UNIVERSAL INTERPOLATION FOR DEEP RESIDUAL SELF-ATTENTION NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Universal approximation is a necessary qualitative property of learning architectures to benefit from scaling laws. While it is generically verified on a variety of neural architectures and random feature models, it typically involves infinite width limits. In this work, we focus on deep self-attention models and consider instead the 'dual' regime, where approximation power is enabled entirely by depth, and featuring strong parameter sharing across layers, motivated by recent models such as the Looped Transformers. More specifically, we ask whether one can find a predefined finite set of parameters, each defining an attention block, such that the resulting finite set of transformations can map any collection of N sequences of n tokens to any other collection of N sequences of n tokens. Crucially, these transformations are *fixed independently of the input and output* collections: only the order in which the blocks are applied, their signs, and their durations depend on the particular interpolation task. Our main result establishes it for residual softmax attention using only two frozen single-head blocks with Gaussian-initialized projection matrices. The result holds at both continuous and finite depth. We also characterize the restrictions imposed by causal masking and establish corresponding universal interpolation guarantees.

1 INTRODUCTION

Introduced by Vaswani et al. (2017), transformers have driven major advances in natural language processing (Devlin et al., 2019), computer vision (Dosovitskiy et al., 2021) and reasoning (Wei et al., 2022). This widespread success motivates several key questions regarding their architecture, including their expressive power, their 'inductive bias' or feature-learning abilities, and ultimately their ability to generalize in the current pretrain/post-train paradigm.

Towards this long-term goal, the first stepping stone is to characterize a flexible class of transformer architectures with enough expressive power to interpolate any given dataset, as a basis for studying scaling laws. A standard approach to studying expressivity in neural architectures is to consider infinitely wide limits, starting with the classic universal approximation theory of Cybenko (1989), Hornik et al. (1989) and others in the single-hidden layer setting. In this context, the design of simple hypothesis spaces (e.g. random feature models) with universal approximation has enabled the subsequent study of non-linear feature learning. Given the prevalent role of depth in the success of recent deep learning models, it is natural to ask whether other mechanisms for universal approximation exist where depth is the instrumental variable; in other words, to consider the deep analogs of random feature models.

In this work we focus on transformer models, and specifically their self-attention modules. Motivated by recent interest on looped transformers, we make an additional distinction: while traditional memorization asks whether, for any given task, it is possible to adjust the parameters of the architecture to produce prescribed outputs, we are interested in capturing the notion of parameter reuse. Specifically, we ask whether it is possible to predefine a finite family of attention blocks *a priori*, without looking at the data, and then achieve exact interpolation only by choosing which block, and with which scalar gain, to use at each layer. Following the fixed-generator formulations in Cuchiero et al. (2020); Agrachev & Letrouit (2026), we call this property *universal interpolation*.

These questions are particularly relevant to recent Transformer architectures that reuse a small number of blocks across depth. Such recurrent-depth or looped architectures have recently been explored

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.

at scale in language models, including recurrent-depth models (Geiping et al., 2025), Mixture-of-Recursions (Bae et al., 2025), Nanbeige4.2-3B (Nanbeige Lab et al., 2026), and Hyperloop Transformers (Zeitoun et al., 2026). Rather than introducing independent parameters at every layer, these models obtain additional effective depth by repeatedly applying the same transformations. This makes it natural to ask how much expressive power can arise from depth and composition alone when the available transformations are fixed.

Exact interpolation. A large body of work studies expressivity of neural networks by asking whether, for each finite dataset, network parameters can be chosen to match the prescribed outputs exactly. Such results characterize the fitting capacity of an architecture, while allowing all of its parameters to depend on the dataset. Ruiz-Balet & Zuazua (2023) establish exact interpolation of finite datasets of distinct points by ReLU neural ODEs in dimension at least two, using data-dependent controls and biases; one active hidden unit suffices. For transformers, Kajitsuka & Sato (2024) construct contextual representations with one attention head and rank-one projections, followed by a feed-forward decoder whose size grows with the data. Dana et al. (2025) prove next-token memorization for attention-only models with constructed embeddings and a readout, with capacity controlled by the number and dimension of heads. Alcalde et al. (2026) obtain exact interpolation of output token sets, without prescribed token correspondences, using attention and ReLU feed-forward blocks with explicit depth bounds. Geshkovski et al. (2026) instead study interpolation of token distributions on the sphere using attention, perceptrons, and spherical layer normalization. In all of these settings, the transformations used by the network may be adapted to the particular interpolation problem.

Universal approximation. A related notion of expressivity is universal approximation, which asks whether a network class can approximate every mapping in a prescribed function class to arbitrary accuracy, under an appropriate topology. Here too, the network parameters may depend on the target function; universality concerns the network class rather than a fixed collection of transformations. Classical results establish universal approximation for sufficiently wide networks with a single hidden layer and a continuous sigmoidal activation (e.g., Cybenko, 1989; Hornik et al., 1989). Lin & Jegelka (2018) establish L^1 universal approximation for deep ReLU residual networks with one-neuron hidden layers and target-dependent weights and biases. From a control-theoretic perspective, Tabuada & Gharesifard (2021) derive approximation guarantees through ensemble controllability and monotonicity. For transformers, Yun et al. (2020) establish L^p approximation of continuous permutation-equivariant mappings with compact support at fixed sequence length, using two attention heads of width one and feed-forward blocks of width four. Furuya et al. (2025) obtain uniform approximation of continuous in-context mappings on compact token domains, with token dimension and head count bounded independently of context length and accuracy while retaining tokenwise MLPs. Hu et al. (2026) establish approximation using two attention layers supplemented with linear transformations. These results address approximation of function classes rather than exact interpolation under a fixed set of reusable transformations.

Universal interpolation with fixed generators. Beyond classical exact interpolation, one can ask whether the transformations available to the network can be fixed in advance and reused across datasets. Cuchiero et al. (2020) construct five fixed vector fields that interpolate every finite dataset of distinct points in a bounded connected domain of dimension at least two, with only the scalar controls depending on the data. Related fixed-family interpolation results have been obtained for general neural ODEs (Grong & Vega-Molino, 2026) and for smooth-activation residual networks (Tabuada & Gharesifard, 2023). For ReLU dynamics, Cai & Duan (2025) establish exact interpolation using arbitrary affine vector fields together with a fixed ReLU field. Combining their result with an exact factorization of affine flow maps into elementary linear flows and translations yields a finite family of fixed generators, independent of the dataset and its size, with biases allowed. Closest to our work, Agrachev & Letrouit (2026) prove that two generic equivariant vector fields suffice for simultaneous interpolation of arbitrarily many compatible point clouds. Their genericity is formulated over an infinite-dimensional space of vector fields, and they explicitly leave extensions to standard softmax attention as an open direction. Establishing such a result is one focus of our work.

Our contributions. We formalize exact and universal interpolation (Definitions 2.3 and 2.4), two related notions used in the literature, and give simple criteria that follow directly from classical theorems in control theory (Section 2.2.1). We also provide a `PyTorch` implementation to investigate these criteria numerically (Section 2.2.2). Our main contributions are:

- to transfer exact (resp. universal) interpolation from continuous to finite depth under the corresponding Lie-rank conditions (Proposition 2.9, resp. Proposition 2.10);
- to establish universal interpolation with frozen random noncausal attention blocks for any prescribed number of heads, using a number of blocks that grows linearly with the batch size (Theorem 3.3);
- to show that only two frozen random single-head noncausal attention blocks suffice for universal interpolation on a connected domain of full measure (Theorem 3.5);
- to rule out exact and universal interpolation on the full configuration space for causal attention when the batch size exceeds the token dimension, by characterizing the exact dimension of the associated Lie-algebra trace (Proposition 3.8);
- to recover universal interpolation for causal attention on each connected component of a class of batches sharing the same linear relations among their first tokens (Corollary 3.10).

2 Exact interpolation and universal interpolation

In this section, we formulate exact and universal interpolation in continuous depth and show how the resulting criteria transfer to finite-depth residual networks. We begin by recalling the standard continuous-depth formulation of residual architectures.

2.1 From residual layers to continuous-depth networks

Given a partition $0 = s_0 < \cdots < s_L =: S$, a residual layer takes the form

$$x_{k+1} = x_k + h_k f_{\theta_k}(x_k), \qquad h_k = s_{k+1} - s_k.$$

Taking $\theta(s) = \theta_k$ on $[s_k, s_{k+1})$, this update is a forward Euler discretization of the continuous-depth dynamics (Lu et al., 2018; Haber & Ruthotto, 2018; Chen et al., 2018)

$$\frac{dx(s)}{ds} = f_{\theta(s)}(x(s)), \qquad x(0) = x_0, \tag{1}$$

where $s \in [0, S]$ represents continuous depth and the parameters $\theta(s)$ may vary with depth but are shared across inputs.

For self-attention, we represent each sequence by its ordered token coordinates $x = \mathrm{vec}(x_1, \dots, x_n) \in \mathbb{R}^{nd}$ and use the finite-dimensional ODE formulation (1). This representation retains the prescribed tokenwise input–output correspondences, which are not encoded by an empirical-measure description (Sander et al., 2022).

We consider two classes of vector fields, corresponding to residual ReLU networks and residual self-attention networks.

Example 2.1 (Continuous-depth ResNets)**.** A bias-free residual block with one ReLU hidden layer has update

$$f_{U,V} : x \in \mathbb{R}^d \mapsto U\sigma(V^\top x) \in \mathbb{R}^d, \tag{2}$$

where $\sigma(t) \coloneqq \max\{t, 0\}$ is applied componentwise and $U, V \in \mathbb{R}^{d\times r}$, with r the hidden width.

Example 2.2 (Continuous-depth self-attention)**.** For $x = \mathrm{vec}(x_1, \dots, x_n) \in \mathbb{R}^{nd}$, let $\mathcal{J} = (J_i)_{i=1}^n$ specify the visible key positions. An H-head block updates token i by

$$[f^{\mathcal{J}}_{\mathbf{W},\mathbf{V}}(x)]_i \coloneqq \sum_{h=1}^{H} \frac{\sum_{j\in J_i} e^{\langle W_h x_i, x_j\rangle} V_h x_j}{\sum_{q\in J_i} e^{\langle W_h x_i, x_q\rangle}}, \qquad i = 1, \dots, n. \tag{3}$$

Here $\mathbf{W} = (W_h)_{h=1}^H$ and $\mathbf{V} = (V_h)_{h=1}^H$, with the standard parametrization (Vaswani et al., 2017)

$$W_h = \frac{(W_h^K)^\top W_h^Q}{\sqrt{r_h}}, \qquad V_h = W_h^O W_h^V,$$

where $W_h^Q, W_h^K, W_h^V \in \mathbb{R}^{r_h\times d}$ and $W_h^O \in \mathbb{R}^{d\times r_h}$. The head widths r_h are arbitrary positive integers, including the standard choice $r_h = d/H$ when H divides d. We write $f_{\mathbf{W},\mathbf{V}}$ for noncausal attention ($J_i = \{1, \dots, n\}$) and $f^{\mathrm{c}}_{\mathbf{W},\mathbf{V}}$ for causal attention ($J_i = \{1, \dots, i\}$).

2.2 EXACT AND UNIVERSAL INTERPOLATION

Let $\mathcal{F} = \{f_\theta : \mathbb{R}^p \to \mathbb{R}^p : \theta \in \Theta\}$ be the family of available updates, with each f_θ locally Lipschitz. We represent a dataset of N samples by its vectorisation $X = \mathrm{vec}(x^1, \dots, x^N) \in \mathbb{R}^{pN}$, where $(x^1, \dots, x^N)$ is an ordered batch. For residual ReLU networks, each sample is a vector in $\mathbb{R}^d$, so $p = d$. For self-attention, each sample is a sequence of n tokens vectorized in $\mathbb{R}^{nd}$, so $p = nd$ and N counts sequences rather than tokens. Applying the same weights to the whole batch gives the *lifted* update

$$f_\theta^{(N)}(X) \coloneqq \big(f_\theta(x^1), \dots, f_\theta(x^N)\big).$$

A depth-dependent choice of parameters $\theta(\cdot)$ is called a *control*; throughout, controls are piecewise constant with finitely many depth intervals.

Architectural symmetries and preserved relations constrain which input–output pairs can be interpolated. We study interpolation on a chosen configuration domain $\mathcal{M}_N \subseteq \mathbb{R}^{pN}$, specified for each architecture below. Both input and target configurations must lie in this domain, with targets represented in the same space as the inputs.

Definition 2.3 (Exact interpolation)**.** Fix $N \geq 1$ and a nonempty, open, connected set $\mathcal{M}_N \subseteq \mathbb{R}^{pN}$. For $S > 0$, a piecewise-constant control $\theta : [0, S] \to \Theta$ with finitely many depth intervals *exactly interpolates* a pair $X, Y \in \mathcal{M}_N$ if the solution of

$$\dot{Z}(s) = f_{\theta(s)}^{(N)}(Z(s)), \qquad Z(0) = X,$$

remains in $\mathcal{M}_N$ and satisfies $Z(S) = Y$. We say that $\mathcal{F}$ has the *exact interpolation property on* $\mathcal{M}_N$ if every pair $X, Y \in \mathcal{M}_N$ admits such a terminal depth $S > 0$ and a control.

Universal interpolation additionally requires a finite dictionary of updates to be fixed before the input–target dataset is given. For the next definition, the architecture must allow the negative of each selected update. Both architectures above have this property by negating the output matrices.

Definition 2.4 (Universal interpolation)**.** Fix $N \geq 1$ and a nonempty, open, connected set $\mathcal{M}_N \subseteq \mathbb{R}^{pN}$. The architecture described by $\mathcal{F}$ has *universal interpolation on* $\mathcal{M}_N$ if there exist an integer $m \geq 1$ and parameters $\theta_1, \dots, \theta_m \in \Theta$ such that, for every $X, Y \in \mathcal{M}_N$, there exist a terminal depth $S > 0$ and a piecewise-constant control

$$g : [0, S] \to \mathcal{D}_N, \qquad \mathcal{D}_N \coloneqq \{\pm f_{\theta_j}^{(N)} : 1 \leq j \leq m\},$$

with finitely many depth intervals, such that

$$\dot{Z}(s) = g(s)(Z(s)), \qquad Z(0) = X, \qquad Z(S) = Y,$$

with $Z(s) \in \mathcal{M}_N$ for all $s \in [0, S]$.

For fixed N, the parameters are chosen once for all admissible input–target datasets. Fitting a dataset changes only the selected signed updates, their order, and their durations, shared across all samples; in particular, S may depend on X and Y. The selected parameters and their number may depend on N, while p and the architecture remain fixed. We count the m base fields; their lifts and negatives form $\mathcal{D}_N$ without additional parameter choices or random draws.

Our use of *universal interpolation* follows the fixed-vector-field formulations in Cuchiero et al. (2020); Agrachev & Letrouit (2026); Grong & Vega-Molino (2026). Elsewhere, this term also denotes exact interpolation for every N without requiring a fixed finite dictionary (Cheng et al., 2025; Cai & Duan, 2025).

2.2.1 CRITERIA FOR EXACT AND UNIVERSAL INTERPOLATION

The rank criteria below ask whether the lifted updates and their iterated Lie brackets span the entire batch space, of dimension $D = Np$. Composing flows in different orders can produce directions outside the span of the instantaneous updates. For smooth fields, this effect is captured by the Lie bracket

$$[g, h](X) \coloneqq Dh(X)g(X) - Dg(X)h(X),$$

where Dg denotes the Jacobian. For a family $\mathcal{G}$ of smooth fields on an open set $\mathcal{U}$, the Lie algebra $\mathrm{Lie}(\mathcal{G})$ is the real linear span of its fields and their iterated brackets. Its *trace* at $X \in \mathcal{U}$ is

$$\mathrm{Lie}(\mathcal{G})(X) \coloneqq \mathrm{span}\{h(X) : h \in \mathrm{Lie}(\mathcal{G})\}.$$

We use the control formulation of the classical Chow–Rashevskii criterion (Agrachev & Sachkov, 2004, Theorem 5.9 and Corollary 5.11), a consequence of the orbit theorem (Sussmann, 1973). Closure under negation allows flow segments to be traversed in both directions, with no prescribed terminal depth.

Theorem 2.5 (Chow–Rashevskii, control formulation)**.** *Let $D \geq 1$, let $\mathcal{M} \subseteq \mathbb{R}^D$ be nonempty, open, and connected, and let $\mathcal{G}$ be a family of smooth vector fields on $\mathcal{M}$ closed under negation. If*

$$\mathrm{Lie}(\mathcal{G})(X) = \mathbb{R}^D \qquad \textit{for every } X \in \mathcal{M},$$

then, for every $X, Y \in \mathcal{M}$, there exist $S > 0$ and a piecewise-constant control $g : [0, S] \to \mathcal{G}$ with finitely many intervals such that

$$\dot{Z}(s) = g(s)(Z(s)), \qquad Z(0) = X, \qquad Z(S) = Y,$$

with $Z(s) \in \mathcal{M}$ for all $s \in [0, S]$.

Full rank makes every nearby configuration reachable. Connectedness then extends this local property to the whole domain. The two corollaries below (see Sections A and B for the proofs) apply this theorem to all available updates or to a fixed finite selection together with its negatives. In both, we assume that $\mathcal{F}$ is closed under negation. To include ReLU updates, we only require the fields selected at each configuration to be smooth in a neighborhood of that configuration. In the universal criterion, the index set I_X selects these fields from the fixed dictionary and may vary with X.

Corollary 2.6 (Criterion for exact interpolation)**.** *Fix $N \geq 1$ and a nonempty, open, connected configuration space $\mathcal{M}_N \subseteq \mathbb{R}^{pN}$. Suppose that, for every $X \in \mathcal{M}_N$, there are finitely many parameters $\theta_1, \ldots, \theta_m$ whose lifted fields are smooth near X and satisfy*

$$\mathrm{Lie}\big(f^{(N)}_{\theta_1}, \ldots, f^{(N)}_{\theta_m}\big)(X) = \mathbb{R}^{Np}.$$

Then $\mathcal{F}$ has the exact interpolation property on $\mathcal{M}_N$ in the sense of Definition 2.3.

Corollary 2.7 (Criterion for universal interpolation)**.** *Fix $N \geq 1$ and a configuration space $\mathcal{M}_N$ as in Definition 2.4. Suppose there exist finitely many parameters $\theta_1, \ldots, \theta_m$ such that, at each $X \in \mathcal{M}_N$, an index set $I_X \subseteq \{1, \ldots, m\}$ selects lifted fields smooth near X with*

$$\mathrm{Lie}\big(f^{(N)}_{\theta_j} : j \in I_X\big)(X) = \mathbb{R}^{Np}.$$

Then $\mathcal{F}$ has universal interpolation on $\mathcal{M}_N$ in the sense of Definition 2.4.

In the exact criterion, the parameters may depend on the configuration X. In the universal criterion, the same finite parameter set must work for every $X \in \mathcal{M}_N$. For globally smooth updates, one can take $I_X = \{1, \ldots, m\}$.

Remark 2.8 (Analytic converse). For real-analytic fields, the rank conditions above are also necessary, by the analytic orbit theorem (Agrachev & Sachkov, 2004, Corollary 5.17). This applies in particular to linear networks and softmax attention. For merely smooth fields, the conditions remain sufficient but need not be necessary. See Section C.

A stronger sufficient condition is that the update fields themselves span $\mathbb{R}^{Np}$ at every configuration, without using Lie brackets. This direct spanning condition only requires local C^1 regularity of the selected fields; see Corollary D.1 in Section D. For a fixed dictionary, it requires at least Np fields, whereas Lie brackets may yield full rank with fewer fields.

2.2.2 Computing the Lie-algebra rank

Following Marcotte et al. (2023, Section 3.3), we investigate the Lie-algebra rank condition by iteratively generating Lie brackets and computing the rank of their evaluations. For a finite family of smooth lifted fields $g_1, \ldots, g_m$, set $W_0 \coloneqq \mathrm{span}\{g_1, \ldots, g_m\}$ and construct

$$W_k \coloneqq W_{k-1} + [W_0, W_{k-1}], \qquad k \geq 1.$$

Here $[W_0, W_{k-1}]$ is the real span of the corresponding brackets, and $\mathrm{Lie}(W_0) = \bigcup_{k\geq 0} W_k$. At each step, the original generators are bracketed with the fields constructed so far, and their evaluations at a configuration X give the rank $r_k(X) := \dim W_k(X)$. A rank $r_k(X) = D$ certifies full Lie rank at X. More generally, the stopping criterion of Marcotte et al. (2023, Lemma 3.2) applies if $r_{k+1}(Y) = r_k(X)$ for every Y in a neighborhood of X: the traces $W_j(Y)$ then equal $W_k(Y)$ locally for all $j \geq k$. We adapt their `SageMath` implementation to `PyTorch` (code in the supplementary material), replacing symbolic differentiation with automatic differentiation. Lie brackets are evaluated through Jacobian–vector products, which can themselves be differentiated to compute nested brackets without forming full Jacobians or expanding symbolic expressions. This makes it possible to explore larger configuration spaces and higher bracket orders for computationally demanding vector fields, such as the softmax-attention updates in (3), whose repeated symbolic differentiation becomes costly.

2.3 Exact and universal interpolation at finite depth

Euler discretization of an interpolating ODE generally leaves a nonzero output error. For a fixed finite family of smooth fields, full Lie rank nevertheless suffices for finite-depth interpolation using residual layers with signed scalar coefficients; see Proposition 2.9. The stronger direct spanning condition gives an explicit correction bound: under local C^1 regularity of the selected fields, at most $D = Np$ additional residual layers make a sufficiently fine Euler discretization exact. Near the target, D independent update directions make the endpoint map locally invertible in the corresponding step sizes, allowing them to cancel a sufficiently small discretization error exactly.

At finite depth, exact and universal interpolation are defined as in Definitions 2.3 and 2.4, respectively, replacing the controlled ODEs with finite compositions of residual layers on the batch. In the exact case, we write these layers as $\mathrm{Id} + f_\theta^{(N)}$: for both architectures considered here, any scalar step size can be absorbed into the output projections. In the universal case, the layers take the form $\mathrm{Id} + c_k f_{\theta_{j_k}}^{(N)}$, where the parameters $\theta_1, \dots, \theta_m$ are fixed across all admissible datasets, while the indices j_k and scalar coefficients $c_k \in \mathbb{R}$ may depend on the dataset. All intermediate configurations must remain in $\mathcal{M}_N$.

Proposition 2.9 (Exact finite-depth interpolation under Lie rank)**.** *Let $D \geq 1$, let $\mathcal{M} \subseteq \mathbb{R}^D$ be nonempty, open, and connected, and let $\mathcal{G}$ be a family of locally Lipschitz vector fields on $\mathcal{M}$ closed under real scalar multiplication. Suppose that, at each $Z \in \mathcal{M}$, there are finitely many fields $g_1, \dots, g_m \in \mathcal{G}$ that are smooth on a neighborhood of Z and satisfy*

$$\mathrm{Lie}(g_1, \dots, g_m)(Z) = \mathbb{R}^D.$$

Then, for every $X, Y \in \mathcal{M}$, there exist a finite integer $L \geq 0$ and fields $h_0, \dots, h_{L-1} \in \mathcal{G}$ such that

$$Z^0 = X, \qquad Z^{k+1} = Z^k + h_k(Z^k), \qquad Z^L = Y,$$

with $Z^k \in \mathcal{M}$ for every $0 \leq k \leq L$.

For a batch space $\mathcal{M}_N$, the dimension is $D = Np$. The proof is given in Section E.2. The direct-spanning version is stated and proved in Section E.1.

Proposition 2.10 (Universal finite-depth interpolation under Lie rank)**.** *Fix $N \geq 1$ and a configuration space $\mathcal{M}_N$ as in Definition 2.4, and assume that $\mathcal{F}$ allows real scalar multiplication. Suppose there exist finitely many parameters $\theta_1, \dots, \theta_m$ such that, at each $X \in \mathcal{M}_N$, an index set $I_X \subseteq \{1, \dots, m\}$ selects lifted fields smooth near X with*

$$\mathrm{Lie}\big(f_{\theta_j}^{(N)} : j \in I_X\big)(X) = \mathbb{R}^{Np}.$$

Then $\mathcal{F}$ has universal interpolation at finite depth on $\mathcal{M}_N$, using these same parameters.

The proof is given in Section F.2. The direct-spanning version is stated and proved in Section F.1. Under the direct-spanning assumption of Proposition E.1, at most $D = Np$ layers are needed for the final correction. Neither construction gives a bound on total depth or parameter norms in terms of dataset size and architecture dimensions.

3 Applications to continuous- and finite-depth networks

We apply the criteria of Section 2 to noncausal and causal self-attention. The appendix also discusses bias-free ReLU ResNets and the obstruction to interpolation in linear networks (Sections G and G.1).

3.1 NONCAUSAL SELF-ATTENTION NETWORKS

In this section, we show that noncausal self-attention networks achieve universal interpolation with a finite number of frozen Gaussian blocks. For single-head attention of width $r \geq d$, two blocks suffice for every batch size and sequence length (Theorem 3.5).

For $N \geq 1$, write $X = \mathrm{vec}(x^1, \dots, x^N) \in \mathbb{R}^{Nnd}$, where $x^\ell = \mathrm{vec}(x_1^\ell, \dots, x_n^\ell)$ and $x_i^\ell \in \mathbb{R}^d$. Our interpolation domain is

$$\mathcal{B}_{N,n,d} \coloneqq \left\{ X \in \mathbb{R}^{Nnd} : \begin{array}{l} x_i^\ell \neq 0, \quad x_i^\ell \notin \mathbb{R} x_{i'}^\ell \text{ for } i \neq i', \\ x^\ell \neq \pm P x^\ell \text{ for } \ell < \ell',\ P \in S_n \end{array} \right\},$$

where S_n acts by permuting tokens. Tokens are nonzero and pairwise non-collinear within each sequence, may be shared across sequences, and no two sequences are related by a signed permutation. We will also consider the generic subset

$$\mathcal{A}_{N,n,d} \coloneqq \left\{ X \in \mathbb{R}^{Nnd} : \begin{array}{l} x_i^\ell \neq 0, \text{ and} \\ x_i^\ell \notin \mathbb{R} x_{i'}^{\ell'} \text{ whenever } (\ell, i) \neq (\ell', i') \end{array} \right\}.$$

For $d \geq 3$, both domains are open, dense and connected, with $\mathcal{A}_{N,n,d} \subseteq \mathcal{B}_{N,n,d}$; see Sections H and J. Noncausal attention is permutation-equivariant and odd, as are its residual layers and controlled flows. Thus, two input sequences related by $x^{\ell'} = \varepsilon P x^\ell$, where $\ell \neq \ell'$, P permutes tokens, and $\varepsilon \in \{\pm 1\}$, require targets satisfying $y^{\ell'} = \varepsilon P y^\ell$. Both domains exclude such input pairs. These are sufficient restrictions: symmetry-related sequences can also be interpolable when their targets respect those symmetries.

Lemma 3.1 (Pointwise spanning by noncausal self-attention)**.** *Let $d \geq 2$, $n \geq 2$, and $N \geq 1$. At every $X \in \mathcal{A}_{N,n,d}$, the single-head batch fields satisfy*

$$\mathrm{span} \left\{ f_{W,V}^{(N)}(X) : \mathrm{rank}\, W, \mathrm{rank}\, V \leq 1 \right\} = \mathbb{R}^{Nnd}.$$

The proof is given in Section H.

Remark 3.2 (Exact interpolation on $\mathcal{A}_{N,n,d}$). For $d \geq 3$, connectedness and the criteria of Section 2 turn this lemma into exact interpolation on $\mathcal{A}_{N,n,d}$ at both continuous and finite depth, with all intermediate configurations in this domain. One head of width one suffices, so the conclusion holds for any prescribed number of heads and positive head widths. For $n = 1$, attention is linear and exact interpolation fails for $N > d$ by Lemma G.5.

The following theorem (see Section J for a proof) shows that a finite number of random attention blocks suffices to achieve universal interpolation on $\mathcal{B}_{N,n,d}$, and hence exact interpolation, with the required number of blocks growing with N.

Theorem 3.3 (Universal interpolation with Gaussian-initialized attention blocks)**.** *Let $d \geq 3$, $n \geq 2$, and $N \geq 1$, and put $D \coloneqq Nnd$. Fix an H-head architecture with positive widths $r_1, \dots, r_H$. Initialize $m_N = 2D$ noncausal attention blocks by drawing every entry of all their query, key, value, and output projection matrices independently from $\mathcal{N}(0, 1)$. The effective matrices are defined by the factorizations in Example 2.2. With probability one over this single draw, these frozen blocks yield universal interpolation at continuous depth on $\mathcal{B}_{N,n,d}$ in the sense of Definition 2.4, simultaneously for all input and target configurations in this domain. Only the block indices, signs, and durations depend on the prescribed interpolation problem; the terminal depth $S > 0$ may depend on the endpoints.*

The proof has two steps. First, we prove a uniform version of Lemma 3.1: almost surely, $2D$ Gaussian blocks drawn independently of X span the tangent space at every $X \in \mathcal{A}_{N,n,d}$ simultaneously. This follows from the pointwise full-rank witness given by Lemma 3.1 and an analyticity argument (see Lemma I.1). Second, we extend full Lie rank to $\mathcal{B}_{N,n,d}$. By contextual separation, a short flow along one of the sampled blocks moves any configuration into $\mathcal{A}_{N,n,d}$ (see Lemma J.2). The system is analytic, so the Lie rank stays the same along this flow and is therefore full at the starting point. Universal interpolation then follows from the connectedness of $\mathcal{B}_{N,n,d}$ by Corollary 2.7.

Corollary 3.4 (Universal interpolation at finite depth)**.** *Under the assumptions of Theorem 3.3, with probability one, the same $2Nnd$ sampled fields yield universal interpolation at finite depth on $\mathcal{B}_{N,n,d}$, in the sense of Section 2.3. All nonendpoint configurations can be chosen in $\mathcal{A}_{N,n,d}$.*

See Section K for a proof. The $2Nnd$ sampled blocks form a reusable dictionary; the number of layers may depend on the interpolation problem. In Theorem 3.3, the number of frozen blocks grows with N. Our main result removes this dependence: for single-head attention of width $r \geq d$, a dictionary of only two frozen blocks suffices for every batch size and sequence length.

Theorem 3.5 (Universal interpolation with two Gaussian attention blocks)**.** *Fix $d \geq 3$. Consider two noncausal single-head attention blocks with head width $r \geq d$ and query, key, value, and output projection matrices whose entries are all sampled independently from $\mathcal{N}(0,1)$. With probability one, simultaneously for every $N \geq 1$ and $n \geq 2$, there exists an open and connected subset*

$$\mathcal{R}_{N,n} \subseteq \mathcal{B}_{N,n,d},$$

depending on the sampled matrices, whose complement in $\mathcal{B}_{N,n,d}$ has Lebesgue measure zero. On $\mathcal{R}_{N,n}$, these two frozen blocks yield universal interpolation at continuous depth in the sense of Definition 2.4, with $m_N = 2$. The same sampled matrices work for every N and n; only the block selection, signs, and durations depend on the interpolation problem.

For almost every draw of the blocks, input and target batches with i.i.d. standard Gaussian coordinates, sampled independently of the blocks, belong to $\mathcal{R}_{N,n}$ with probability one.

We define $\mathcal{R}_{N,n}$ as the configurations where the two sampled fields and their iterated Lie brackets span the full batch space. The main step is to show that the analytic rank-deficient set has codimension at least two. Removing this set preserves connectedness, allowing the interpolation criterion in Corollary 2.7 to connect any two configurations in $\mathcal{R}_{N,n}$. The full proof and the precise definition of the domain are given in Section L, which also contains the proof of the finite-depth corollary.

Remark 3.6 (Head width and dictionary size). The theorem covers the standard single-head choice $r = d$. For $r < d$, our numerical experiments (Section M) suggest that $m \geq \lceil d/r \rceil$ independent Gaussian-initialized frozen blocks suffice for generic full lifted Lie rank, independently of N and n, with $n \geq 2$. With the standard width $r = d/H$ ($H \geq 2$), this gives the conjectured threshold $m \geq H$, using one active head per block and setting the other heads to zero. A corresponding universal interpolation theorem remains open.

Corollary 3.7 (Universal interpolation at finite depth with two blocks)**.** *Under the assumptions of Theorem 3.5, with probability one, the same two frozen blocks yield universal interpolation at finite depth on the same domains $\mathcal{R}_{N,n}$, in the sense of Section 2.3, simultaneously for every $N \geq 1$ and $n \geq 2$. Every layer uses one of the two sampled fields with a real scalar gain, and every intermediate batch configuration remains in $\mathcal{R}_{N,n}$.*

3.2 Causal self-attention networks

In this section, we show that causal masking makes the first-token update linear, so every linear relation among the first tokens of a batch is preserved at every depth. As a result, exact interpolation fails for $N > d$, whatever the number of heads (Proposition 3.8). Positive results are recovered once input and target batches share the same first-token relations: on each such class, finitely many frozen Gaussian blocks achieve universal interpolation (Corollary 3.10).

We consider the causal attention architecture of Example 2.2 with any prescribed number of heads and positive head widths. Causal masking makes the first-token update linear: the first token attends only to itself. We retain the globally noncollinear domain $\mathcal{A}_{N,n,d}$; the extension allowing shared tokens in Section 3.1 concerns noncausal attention. The interpolation constructions below require only one active head; all other value matrices can be set to zero. We denote the batch lift of a single-head causal update by $F^{\mathrm{c}}_{W,V}$. For a general H-head block, $[f^{\mathrm{c}}_{\mathbf{W},\mathbf{V}}(x)]_1 = V_\Sigma x_1$, where $V_\Sigma := \sum_{h=1}^{H} V_h$. For any batch X, denote its first-token matrix by

$$X_1 := \begin{bmatrix} x_1^1 & \cdots & x_1^N \end{bmatrix} \in \mathbb{R}^{d \times N}.$$

Along continuous and discrete batch trajectories $Z(s)$ and Z^k starting at X, the first-token matrices satisfy $\dot{Z}_1 = V_\Sigma(s) Z_1$ and $Z_1^{k+1} = (I + V_{\Sigma,k}) Z_1^k$, respectively. In both settings, the first-token matrix at any depth has the form MX_1 for some $M \in \mathbb{R}^{d \times d}$. Hence every initial linear relation among the first tokens remains valid at every depth. For example, an input relation $x_1^3 = x_1^1 + x_1^2$ forces every reachable target to satisfy $y_1^3 = y_1^1 + y_1^2$, at both continuous and finite depth. The following proposition shows that later token positions admit independent update directions.

Proposition 3.8 (Exact rank of causal attention). *Let $d \geq 2$, $n \geq 2$, and $N \geq 1$. At a fixed batch $X \in \mathcal{A}_{N,n,d}$, let $\mathcal{S}^{\mathrm{c}}_X$ denote the linear span of the batch updates $F^{\mathrm{c}}_{W,V}(X)$ with $\operatorname{rank} W, \operatorname{rank} V \leq 1$. Then*

$$\mathcal{S}^{\mathrm{c}}_X = \left\{ Z \in \mathbb{R}^{Nnd} : Z_1 = AX_1 \text{ for some } A \in \mathbb{R}^{d\times d} \right\}. \tag{4}$$

Consequently,

$$\dim \mathcal{S}^{\mathrm{c}}_X = N(n-1)d + d \operatorname{rank} X_1. \tag{5}$$

The same span and rank hold for any prescribed number of heads and positive head widths. Lie brackets do not increase this dimension. Full ambient rank Nnd holds exactly when the N first tokens are linearly independent, which requires $N \leq d$.

The proof is given in Section N. The $N(n-1)d$ directions associated with later tokens are unrestricted; the first-token directions are exactly those obtained by applying a common linear map to the first tokens. By the rank criterion and its analytic converse in Corollary 2.6 and Remark 2.8, causal attention has the continuous-depth exact interpolation property on a nonempty connected open subset of $\mathcal{A}_{N,n,d}$ if and only if the first tokens are linearly independent at every configuration in that subset. For $N > d$, this is impossible; the preserved first-token relations also rule out arbitrary interpolation at finite depth. This obstruction concerns interpolation of the full target sequences, including their first tokens. Pairwise noncollinearity alone does not ensure linear independence when $N \geq 3$.

To obtain positive interpolation results for arbitrary N, we restrict input and target batches to have exactly the same linear relations among their first tokens. For a fixed subspace $K \subseteq \mathbb{R}^N$, define

$$\mathcal{A}^K_{N,n,d} \coloneqq \{X \in \mathcal{A}_{N,n,d} : \ker X_1 = K\}, \qquad r \coloneqq N - \dim K.$$

Here $c \in K$ means that $\sum_{\ell=1}^N c_\ell x_1^\ell = 0$. This class is an open subset of the linear space

$$E_K \coloneqq \{Z \in \mathbb{R}^{Nnd} : Z_1 c = 0 \text{ for every } c \in K\}, \qquad \dim E_K = R \coloneqq N(n-1)d + dr.$$

We apply the interpolation criteria of Section 2 within each connected component, using E_K as the ambient space. The rank in (5) is full in this space.

Corollary 3.9 (Exact interpolation by causal attention). *Let $d \geq 2$, $n \geq 2$, $N \geq 1$, and fix $K \subseteq \mathbb{R}^N$ such that $\mathcal{A}^K_{N,n,d}$ is nonempty. The causal attention family has exact interpolation on each connected component of $\mathcal{A}^K_{N,n,d}$, both at continuous depth on $[0, S]$ for a suitable $S > 0$ and at finite depth. One head of width one suffices, with effective parameter matrices W, V of rank at most one shared across all sequences. The continuous trajectory and all intermediate residual configurations remain in the chosen component.*

The proof is given in Section O. Continuous flows keep the first-token kernel fixed, since their linear action is invertible. Their trajectories within $\mathcal{A}^K_{N,n,d}$ remain in a single connected component.

Corollary 3.10 (Universal interpolation by causal attention). *Under the assumptions of Corollary 3.9, fix K before sampling and set $R = N(n-1)d + dr$. Draw independent standard Gaussian vectors $a_j, b_j, u_j, v_j \in \mathbb{R}^d$, $1 \leq j \leq 2R$, and set $W_j = a_j b_j^\top$, $V_j = u_j v_j^\top$. These matrices specify one active head in each block; any additional heads have zero value matrices. With probability one, these $2R$ frozen blocks yield universal interpolation on each connected component of $\mathcal{A}^K_{N,n,d}$, at both continuous and finite depth. The blocks may be used with either sign and, at finite depth, with real scalar gains. A single draw works simultaneously for all components of this fixed relation class, with each interpolation remaining in its chosen component.*

CONCLUSION

In this paper, we establish universal interpolation for residual self-attention networks with frozen attention parameters, at both continuous and finite depth. In the single-head setting, two randomly sampled fixed blocks are sufficient, independently of the dataset size. A natural direction for future work is to better understand the trade-off between the number of fixed generators and the depth required for interpolation.

AI DISCLOSURE SECTION

We used LLMs to aid or polish our writing. We primarily used Astra, and all parts of this paper were polished by an LLM at some point, following the instructions provided by the authors. We used LLMs for retrieval and discovery, particularly to find related work both within the ML community and in other fields that helped us prove certain results, such as Amir et al. (2023) and Ax (1971). We used LLMs for research execution both to assist with code development and to develop proofs (see our final point below). We used LLMs to draft sections of the paper. Many sections were initially drafted by Astra. The authors systematically refined these drafts through extensive iteration, providing precise feedback at each step, and editing the text manually when needed. Finally, we used LLMs to prove mathematical claims at three different levels: i) proving direct consequences of known / previously proven theorems (see e.g. Section A–Section D), which LLMs can do easily and faster than the authors; ii) refining proofs starting from our own drafts (e.g., Section E, Section G, Section H, Section N); and iii) finding ideas for difficult proofs that required tools from other research communities (e.g., parts of Section I, parts of Section L.3–Section L.6). We used Astra for all three purposes. All proofs have been checked by the authors. For iii), we took particular care to explain the tools and arguments drawn from other research communities in a way that is accessible to the ML community. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide code that can reproduce our results as supplementary materials, and describe experimental details and results in Section M.

REFERENCES

Andrei Agrachev and Cyril Letrouit. Generic controllability of equivariant systems and applications to particle systems and neural networks. *Annales de l'Institut Henri Poincaré C, Analyse non linéaire*, 43(3):639–668, 2026. URL https://ems.press/journals/aihpc/articles/14298488.

Andrei A. Agrachev and Yuri L. Sachkov. *Control Theory from the Geometric Viewpoint*, volume 87 of *Encyclopaedia of Mathematical Sciences*. Springer, Berlin, Heidelberg, 2004. doi: 10.1007/978-3-662-06404-7. URL https://doi.org/10.1007/978-3-662-06404-7.

Albert Alcalde, Giovanni Fantuzzi, and Enrique Zuazua. Exact sequence interpolation with transformers. *Mathematical Foundations of Machine Learning*, 2:2, 2026. doi: 10.1007/s44439-026-00005-y.

Tal Amir, Steven Gortler, Ilai Avni, Ravina Ravina, and Nadav Dym. Neural injective functions for multisets, measures and graphs via a finite witness theorem. In *Advances in Neural Information Processing Systems*, volume 36, pp. 42516–42551, 2023. URL https://papers.neurips.cc/paper_files/paper/2023/hash/84b686f7cc7b7751e9aaac0da74f755a-Abstract-Conference.html.

James Ax. On Schanuel's conjectures. *Annals of Mathematics*, 93(2):252–268, 1971. doi: 10.2307/1970774. URL https://annals.math.princeton.edu/1971/93-2/p03.

Sangmin Bae, Yujin Kim, Reza Bayat, Sungnyun Kim, Jiyoun Ha, Tal Schuster, Adam Fisch, Hrayr Harutyunyan, Ziwei Ji, Aaron C. Courville, and Se-Young Yun. Mixture-of-Recursions: Learning dynamic recursive depths for adaptive token-level computation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3229. URL https://papers.nips.cc/paper_files/paper/2025/hash/8b08bbf8b420faa6eeb4020720582ec7-Abstract-Conference.html.

Edward Bierstone and Pierre D. Milman. Semianalytic and subanalytic sets. *Publications Mathématiques de l'IHÉS*, 67:5–42, 1988. doi: 10.1007/BF02699126. URL https://www.numdam.org/item/PMIHES_1988__67__5_0/.

Yongqiang Cai. Vocabulary for universal approximation: A linguistic perspective of mapping compositions. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5189–5208. PMLR, 2024. URL https://proceedings.mlr.press/v235/cai24a.html.

Yongqiang Cai and Yifei Duan. Achieving universal approximation and universal interpolation via nonlinearity of control families. *arXiv preprint arXiv:2510.03676*, 2025. URL https://arxiv.org/abs/2510.03676.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html.

Jingpu Cheng, Qianxiao Li, Ting Lin, and Zuowei Shen. Interpolation, approximation, and controllability of deep neural networks. *SIAM Journal on Control and Optimization*, 63(1):625–649, 2025. doi: 10.1137/23M1599744.

Christa Cuchiero, Martin Larsson, and Josef Teichmann. Deep neural networks, generic universal interpolation, and controlled ODEs. *SIAM Journal on Mathematics of Data Science*, 2(3):901–919, 2020. doi: 10.1137/19M1284117.

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

Léo Dana, Muni Sreenivas Pydi, and Yann Chevaleyre. Memorization in attention-only transformers. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3133–3141. PMLR, 2025. URL https://proceedings.mlr.press/v258/dana25a.html.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Yifei Duan and Yongqiang Cai. A minimal control family of dynamical systems for universal approximation. *IEEE Transactions on Automatic Control*, 70(11):7233–7244, 2025. doi: 10.1109/TAC.2025.3569123. URL https://arxiv.org/abs/2312.12903.

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/file/c9028f7874df04843e7bf435ee4cd3c3-Paper-Conference.pdf.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2502.05171.

Borjan Geshkovski, Philippe Rigollet, and Domènec Ruiz-Balet. Measure-to-measure interpolation using transformers. *Foundations of Computational Mathematics*, 2026. doi: 10.1007/s10208-026-09754-w.

Erlend Grong and Sylvie Vega-Molino. Controllability on landmark manifolds for shapes and neural ODEs. *Mathematics of Control, Signals, and Systems*, 38(2):275–289, 2026. doi: 10.1007/s00498-026-00435-1.

Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34 (1):014004, 2018.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. doi: 10.1016/0893-6080(89)90020-8.

Jerry Yao-Chieh Hu, Hude Liu, Hong-Yu Chen, Weimin Wu, and Han Liu. Universal approximation with softmax attention. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306. PMLR, 2026. URL https://openreview.net/forum?id=GoW78qJ3qM.

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.

Hongzhou Lin and Stefanie Jegelka. Resnet with one-neuron hidden layers is a universal approximator. *Advances in neural information processing systems*, 31, 2018.

Yiping Lu, Aoxiao Zhong, Quanzheng Li, and Bin Dong. Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 3276–3285. PMLR, 2018. URL https://proceedings.mlr.press/v80/lu18d.html.

Sibylle Marcotte, Rémi Gribonval, and Gabriel Peyré. Abide by the law and follow the flow: Conservation laws for gradient flows. In *Advances in Neural Information Processing Systems*, volume 36, pp. 63210–63221, 2023. doi: 10.52202/075280-2760. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/c7bee9b76be21146fd592fc2b46614d5-Abstract.html.

Nanbeige Lab, Chen Yang, Chengrui Huang, Fufeng Lan, Hanhui Chen, Hao Zhou, Huatong Song, Jiaqi Cao, Jiaying Zhu, Jinlin Niu, Kai Wang, Lisheng Huang, Qiliang Liang, Ran Le, Ruixiang Feng, Shuang Sun, Tao Gu, Tao Zhang, Tianyu Luo, Yang Song, Yun Xing, Yuntao Wen, Ziyao Xu, Zongchao Chen, and Zongqiang Li. Nanbeige4.2-3B: Unlocking agentic capabilities in a compact model. *arXiv preprint arXiv:2607.22083*, 2026. URL https://arxiv.org/abs/2607.22083.

Domènec Ruiz-Balet and Enrique Zuazua. Neural ODE control for classification, approximation, and transport. *SIAM Review*, 65(3):735–773, 2023. doi: 10.1137/21M1411433.

Michael E Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pp. 3515–3530. PMLR, 2022. URL https://proceedings.mlr.press/v151/sander22a.html.

I. R. Shafarevich. Basic notions of algebra. In A. I. Kostrikin and I. R. Shafarevich (eds.), *Algebra I*, volume 11 of *Encyclopaedia of Mathematical Sciences*, pp. 1–258. Springer, Berlin, Heidelberg, 1990. doi: 10.1007/978-3-662-39643-8_1. URL https://doi.org/10.1007/978-3-662-39643-8_1.

Héctor J. Sussmann. Orbits of families of vector fields and integrability of distributions. *Transactions of the American Mathematical Society*, 180:171–188, 1973.

Héctor J. Sussmann. Some optimal control applications of real-analytic stratifications and desingularization. *Banach Center Publications*, 44(1):211–232, 1998. doi: 10.4064/-44-1-211-232. URL https://matwbn.icm.edu.pl/ksiazki/bcp/bcp44/bcp44121.pdf.

Paulo Tabuada and Bahman Gharesifard. Universal approximation power of deep residual neural networks via nonlinear control theory. In *International Conference on Learning Representations*, 2021. URL https://iclr.cc/virtual/2021/poster/3010.

Paulo Tabuada and Bahman Gharesifard. Universal approximation power of deep residual neural networks through the lens of control. *IEEE Transactions on Automatic Control*, 68(5):2715–2728, 2023. doi: 10.1109/TAC.2022.3190051. URL https://arxiv.org/abs/2007.06007v4. Corrected version; corrigendum doi: 10.1109/TAC.2024.3390099.

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://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

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*, 2020. URL https://openreview.net/forum?id=ByxRM0Ntvr.

Abbas Zeitoun, Lucas Torroba-Hennigen, and Yoon Kim. Hyperloop Transformers. *arXiv preprint arXiv:2604.21254*, 2026. URL https://arxiv.org/abs/2604.21254.

A Proof of Corollary 2.6

Proof. Write $\mathcal{M} = \mathcal{M}_N$, $D = Np$, and $\mathcal{G} = \{f_\theta^{(N)} : \theta \in \Theta\}$. Fix $X \in \mathcal{M}$ and choose the locally smooth lifted fields $g_1, \dots, g_m$ from the hypothesis. Full Lie rank at X gives D Lie-algebra elements whose evaluations form a basis. Their determinant stays nonzero on a smaller connected neighborhood $\mathcal{U}$ of X. Apply Theorem 2.5 on $\mathcal{U}$ to the family $\{\pm g_j|_{\mathcal{U}} : 1 \le j \le m\}$. This joins every pair in $\mathcal{U}$ by a piecewise-constant control using the selected fields and their negatives. Thus a neighborhood of X belongs to its reachability class.

Reachability by finite concatenations of flows of fields in $\mathcal{G}$, with every segment contained in $\mathcal{M}$, is an equivalence relation: transitivity follows by concatenation, while symmetry follows by reversing the segments and replacing each field by its negative. Applying the preceding local argument at every point shows that every equivalence class is open. Since $\mathcal{M}$ is connected, it has only one reachability class. Thus any distinct $X, Y \in \mathcal{M}$ can be joined by finitely many positive-time flow segments:

$$Y = \Phi_{h_L}^{\tau_L} \circ \cdots \circ \Phi_{h_1}^{\tau_1}(X), \qquad h_j \in \mathcal{G}, \quad \tau_j > 0,$$

with every segment contained in $\mathcal{M}$, where Φ_h^t denotes the local flow of h. Taking $S \coloneqq \sum_j \tau_j$ gives the required piecewise-constant control on $[0, S]$. When $X = Y$, follow any field in $\mathcal{G}$ for a sufficiently short positive time and then its negative for the same duration. Local existence and openness of $\mathcal{M}$ ensure that this trajectory stays in $\mathcal{M}$ and returns to X with $S > 0$.

Conversely, suppose all lifted fields are analytic and exact interpolation holds. Each controlled trajectory is a finite concatenation of their flows, so every orbit is $\mathcal{M}$. By Remark 2.8, $\mathrm{Lie}(\mathcal{G})(X) = \mathbb{R}^D$ at every X. A basis of this trace involves only finitely many Lie brackets and hence finitely many fields of $\mathcal{G}$, giving the required finite parameter selection at X. □

B Proof of Corollary 2.7

Proof. Fix the finite parameter selection from the hypothesis. Apply Corollary 2.6 to

$$\widetilde{\mathcal{F}}_N = \{\pm f_{\theta_j} : 1 \le j \le m\}.$$

This family is closed under negation, and at each configuration the indices I_X supply a locally smooth subfamily with full Lie rank. On the batch, the resulting controls select one signed lifted update on each of finitely many intervals, with a terminal depth $S > 0$, exactly as in Definition 2.4. The parameters were chosen before the endpoints, so $\mathcal{F}$ has universal interpolation on $\mathcal{M}_N$.

Conversely, suppose the lifted fields are analytic and choose the finite parameters supplied by universal interpolation on $\mathcal{M}_N$. These fields and their negatives have one orbit on $\mathcal{M}_N$. By Remark 2.8 their Lie algebra has rank Np everywhere. Adding negatives does not change the generated Lie algebra, so the fixed lifted fields themselves have full Lie rank; take $I_X = \{1, \dots, m\}$. □

C Justification of Remark 2.8

Proof. Let $\mathcal{M} \subseteq \mathbb{R}^D$ be nonempty, open, and connected, and let $\mathcal{G}$ be a family of real-analytic fields on $\mathcal{M}$ closed under negation. Let $\mathcal{O}(X)$ denote the orbit through X generated by their local flows. The analytic orbit theorem gives

$$T_X\mathcal{O}(X) = \mathrm{Lie}(\mathcal{G})(X)$$

for every $X \in \mathcal{M}$ (Agrachev & Sachkov, 2004, Corollary 5.17) (see also Sussmann, 1998, Theorem 5). Closure under negation allows flow segments to be reversed. Thus $\mathcal{O}(X)$ is precisely the set reachable from X by piecewise-constant controls on some finite interval $[0, S]$. The point X itself is reached by a short flow segment followed by its reverse. If the conclusion of Theorem 2.5 holds, then $\mathcal{O}(X) = \mathcal{M}$, so $\mathrm{Lie}(\mathcal{G})(X) = T_X\mathcal{M} = \mathbb{R}^D$. Sufficiency is already given by Theorem 2.5. □

D Pointwise full-rank criterion

Corollary D.1 (Pointwise full-rank criterion)**.** *Let $D \ge 1$, let $\mathcal{M} \subseteq \mathbb{R}^D$ be nonempty, open, and connected, and let $\mathcal{G}$ be a family of locally Lipschitz vector fields on $\mathcal{M}$. Suppose that:*

1. *for every $X \in \mathcal{M}$, there are D vector fields $g_1, \dots, g_D \in \mathcal{G}$ that are C^1 on a neighborhood of X and whose values $g_1(X), \dots, g_D(X)$ form a basis of $\mathbb{R}^D$; and*

2. *$g \in \mathcal{G}$ implies $-g \in \mathcal{G}$.*

Then piecewise-constant controls with values in $\mathcal{G}$ exactly interpolate every pair $X, Y \in \mathcal{M}$ on $[0, S]$ for some $S > 0$ depending on the pair.

Proof. Fix $X \in \mathcal{M}$ and choose $g_1, \dots, g_D$ as in the first assumption. If φ_j^t denotes the local flow of g_j, the endpoint map

$$E(t_1, \dots, t_D) \coloneqq \varphi_D^{t_D} \circ \cdots \circ \varphi_1^{t_1}(X)$$

has derivative at the origin whose columns are $g_1(X), \dots, g_D(X)$. The inverse function theorem therefore shows that the reachability class of X contains a neighborhood of X; negative times are realized using the fields $-g_j$. The symmetry, concatenation, and connectedness arguments in the proof of Corollary 2.6 then give the conclusion. □

E FINITE-DEPTH INTERPOLATION

E.1 DIRECT SPANNING

Proposition E.1 (Exact finite-depth interpolation)**.** *Let $D \geq 1$, let $\mathcal{M} \subseteq \mathbb{R}^D$ be nonempty, open, and connected, and let $\mathcal{G}$ be a family of locally Lipschitz vector fields on $\mathcal{M}$ closed under real scalar multiplication. Suppose that, at each $X \in \mathcal{M}$, $\mathcal{G}$ contains D fields that are C^1 near X and whose values at X form a basis of $\mathbb{R}^D$. Then, for every $X, Y \in \mathcal{M}$, there exist a finite integer $L \geq 0$ and fields $g_0, \dots, g_{L-1} \in \mathcal{G}$ such that*

$$Z^0 = X, \qquad Z^{k+1} = Z^k + g_k(Z^k), \qquad Z^L = Y, \tag{6}$$

with $Z^k \in \mathcal{M}$ for every $0 \leq k \leq L$.

Proof. Fix $X, Y \in \mathcal{M}$.

Step 1: construct an exact correction near the target. We first find a neighborhood of Y such that each point in this neighborhood can be sent exactly to Y by at most D admissible residual layers, with all intermediate states remaining in $\mathcal{M}$. The layer coefficients may depend on the point being corrected. Choose $g_1, \dots, g_D \in \mathcal{G}$ that are C^1 on a neighborhood of Y and whose values at Y form a basis of $\mathbb{R}^D$. For $a = (a_1, \dots, a_D) \in \mathbb{R}^D$, set

$$R_j^{a_j}(Z) \coloneqq Z + a_j g_j(Z), \qquad E(Z, a) \coloneqq R_D^{a_D} \circ \cdots \circ R_1^{a_1}(Z).$$

This map is C^1 near $(Y, 0)$. By shrinking this neighborhood, all intermediate states in the composition remain in $\mathcal{M}$ and in the common C^1 neighborhood of the selected fields. At $(Y, 0)$,

$$E(Y, 0) = Y, \qquad D_a E(Y, 0) = \begin{bmatrix} g_1(Y) & \cdots & g_D(Y) \end{bmatrix}. \tag{7}$$

The latter matrix is invertible. Applying the implicit function theorem to $E(Z, a) - Y = 0$ gives an open neighborhood $\mathcal{U}_Y \subseteq \mathcal{M}$ of Y and a C^1 map $a : \mathcal{U}_Y \to \mathbb{R}^D$, with $a(Y) = 0$, such that

$$E(Z, a(Z)) = Y \qquad \text{for every } Z \in \mathcal{U}_Y. \tag{8}$$

Each nonzero coefficient $a_j(Z)$ can be absorbed into its field, since closure under real scalar multiplication gives $a_j(Z) g_j \in \mathcal{G}$. A zero coefficient gives an identity layer and can be omitted. Thus every $Z \in \mathcal{U}_Y$ can be mapped exactly to Y by at most D admissible residual layers.

Step 2: reach the correction neighborhood by Euler discretization. We next approximate a controlled trajectory from X to Y by admissible residual layers. We show that a sufficiently fine mesh keeps every iterate in $\mathcal{M}$ and places the endpoint in the neighborhood $\mathcal{U}_Y$ constructed in Step 1. By Corollary D.1, a piecewise-constant controlled trajectory $\gamma : [0, S] \to \mathcal{M}$ joins X to Y for some $S > 0$. Fix any such trajectory. Its image Γ is compact, so there is $\rho > 0$ such that the compact tube

$$K_\rho \coloneqq \{Z \in \mathbb{R}^D : \operatorname{dist}(Z, \Gamma) \leq \rho\} \subseteq \mathcal{M}.$$

Let $0 = s_0 < \cdots < s_q = S$ be the switching times of the fixed control, and let $h_i \in \mathcal{G}$ denote the vector field selected on $[s_{i-1}, s_i)$, so that

$$\dot{\gamma}(t) = h_i(\gamma(t)) \qquad \text{for } t \in (s_{i-1}, s_i).$$

The fields $h_1, \ldots, h_q$ need not be distinct. By compactness and local Lipschitz continuity, these finitely many fields have a common norm bound $B \geq 1$ and a common Lipschitz constant $\Lambda \geq 1$ on K_ρ.

Indeed, each h_i is continuous, so compactness of K_ρ gives

$$B := \max\left\{1, \max_{1\leq i\leq q} \max_{Z\in K_\rho} \|h_i(Z)\|\right\} < \infty.$$

To obtain a common Lipschitz constant, local Lipschitz continuity and finiteness of the family imply that, for every $x \in K_\rho$, there exist $r_x > 0$ and $L_x \geq 1$ such that $\mathbb{B}(x, 2r_x) \subseteq \mathcal{M}$ and all the h_i are L_x-Lipschitz on this ball. Here $\mathbb{B}(x, r)$ denotes the open Euclidean ball of radius r centered at x.

By compactness, finitely many balls $\mathbb{B}(x_\ell, r_\ell)$, $\ell = 1, \ldots, J$, cover K_ρ. Here $r_\ell = r_{x_\ell}$. Set

$$\eta := \min_{1\leq \ell\leq J} r_\ell > 0, \qquad L_{\mathrm{loc}} := \max_{1\leq \ell\leq J} L_{x_\ell}.$$

If $Z, W \in K_\rho$ satisfy $\|Z - W\| < \eta$, choose ℓ with $Z \in \mathbb{B}(x_\ell, r_\ell)$. Then both Z and W belong to $\mathbb{B}(x_\ell, 2r_\ell)$, and hence

$$\|h_i(Z) - h_i(W)\| \leq L_{\mathrm{loc}}\|Z - W\| \qquad \text{for every } i.$$

If instead $\|Z - W\| \geq \eta$, the norm bound gives

$$\|h_i(Z) - h_i(W)\| \leq 2B \leq \frac{2B}{\eta}\|Z - W\|.$$

Thus

$$\Lambda := \max\{1, L_{\mathrm{loc}}, 2B/\eta\}$$

is a Lipschitz constant on K_ρ shared by all the fields $h_1, \ldots, h_q$.

Choose a finite mesh $0 = t_0 < \cdots < t_m = S$ containing every switching time. Write $\Delta_k = t_{k+1} - t_k$, $h = \max_k \Delta_k$, and let $H_k \in \{h_1, \ldots, h_q\}$ be the field active on $[t_k, t_{k+1})$. Forward Euler gives $Z^0 = X$ and $Z^{k+1} = Z^k + \Delta_k H_k(Z^k)$. On each mesh interval, the integral equation for γ yields

$$\|\gamma(t_{k+1}) - \gamma(t_k) - \Delta_k H_k(\gamma(t_k))\| \leq \tfrac{1}{2}\Lambda B \Delta_k^2.$$

Consequently, while $Z^k \in K_\rho$, the errors $e_k := \|Z^k - \gamma(t_k)\|$ satisfy

$$e_{k+1} \leq (1 + \Lambda\Delta_k)e_k + \tfrac{1}{2}\Lambda B\Delta_k^2.$$

Since $e_0 = 0$ and $\Delta_k^2 \leq h\Delta_k$, induction gives

$$e_k \leq \frac{Bh}{2}\left(\prod_{j<k}(1 + \Lambda\Delta_j) - 1\right) \leq \frac{Bh}{2}(e^{\Lambda t_k} - 1) \leq Ch, \qquad C := \frac{B}{2}(e^{\Lambda S} - 1). \tag{9}$$

Choose $\varepsilon > 0$ with $\{Z : \|Z - Y\| < \varepsilon\} \subseteq \mathcal{U}_Y$, and refine the mesh until $Ch < \min\{\rho, \varepsilon\}$. This can be done with finitely many steps while retaining every switching time. There can be no first iterate outside K_ρ: the bound for that iterate, computed from its predecessor inside the tube, would give distance less than ρ from Γ. Thus all iterates lie in $\mathcal{M}$, (9) holds throughout, and $\widetilde{Y} := Z^m$ lies in the fixed neighborhood $\mathcal{U}_Y$. Each Euler step is an admissible residual layer, since $\Delta_k H_k \in \mathcal{G}$ by positive scaling.

Step 3: combine the Euler and correction layers to obtain exact interpolation. Step 2 gives an Euler endpoint $\widetilde{Y} \in \mathcal{U}_Y$. We now apply the correction from Step 1 at this point to reach Y exactly with at most D additional layers. With the controlled trajectory and Euler mesh now fixed, set $c_j := a_j(\widetilde{Y})$ for $j = 1, \ldots, D$, and define the correction fields

$$\widehat{g}_j : W \longmapsto c_j g_j(W).$$

Each c_j is a real number, so closure under real scalar multiplication gives $\widehat{g}_j \in \mathcal{G}$. Append the residual layers $\mathrm{Id} + \widehat{g}_j$ to the Euler network. Their intermediate states satisfy

$$W_0 = \widetilde{Y}, \qquad W_j = W_{j-1} + c_j g_j(W_{j-1}), \qquad j = 1, \ldots, D.$$

By (8), their final output is $W_D = E(\widetilde{Y}, c) = Y$, where $c = (c_1, \ldots, c_D)$, and all intermediate states remain in $\mathcal{M}$. Coefficients $c_j = 0$ give identity layers and may be omitted. Together with the Euler layers, this gives (6) with at most $m + D$ layers. □

E.2 Extension under a Lie-rank condition

We now prove Proposition 2.9, which replaces the direct-spanning assumption of Proposition E.1 by a local Lie-rank condition. The proof follows the same three steps, with a new local correction in Step 1.

Proposition 2.9 (Exact finite-depth interpolation under Lie rank)**.** *Let $D \geq 1$, let $\mathcal{M} \subseteq \mathbb{R}^D$ be nonempty, open, and connected, and let $\mathcal{G}$ be a family of locally Lipschitz vector fields on $\mathcal{M}$ closed under real scalar multiplication. Suppose that, at each $Z \in \mathcal{M}$, there are finitely many fields $g_1, \dots, g_m \in \mathcal{G}$ that are smooth on a neighborhood of Z and satisfy*

$$\mathrm{Lie}(g_1, \dots, g_m)(Z) = \mathbb{R}^D.$$

Then, for every $X, Y \in \mathcal{M}$, there exist a finite integer $L \geq 0$ and fields $h_0, \dots, h_{L-1} \in \mathcal{G}$ such that

$$Z^0 = X, \qquad Z^{k+1} = Z^k + h_k(Z^k), \qquad Z^L = Y,$$

with $Z^k \in \mathcal{M}$ for every $0 \leq k \leq L$.

Proof. Fix $X, Y \in \mathcal{M}$. We construct a local correction at Y, reach its domain by Euler discretization, and concatenate the two networks. If $X = Y$, no layers are needed to interpolate.

Step 1: construct an exact correction near the target. We seek a fixed finite sequence of residual layers whose coefficients can be adjusted to send any point near Y to Y. Choose a finite family $g_1, \dots, g_m \in \mathcal{G}$ that is smooth on an open neighborhood $U \subseteq \mathcal{M}$ of Y and has full Lie rank at Y. Write φ_j^t for the local flow of g_j.

A flow correction with independent parameter directions. Let Ψ range over finite compositions of the local flows φ_i^t, with signed times, defined near Y and with all their flow segments contained in U, including the identity composition. For a smooth vector field h, its pullback by Ψ is the vector field

$$(\Psi^* h)(Z) \coloneqq \big(D\Psi(Z)\big)^{-1} h(\Psi(Z)).$$

The Jacobian $D\Psi(Z)$ is invertible since Ψ is a composition of local flow maps. Let $\mathcal{V}_Y$ be the linear span of $(\Psi^* g_j)(Y)$ over all such Ψ and $1 \leq j \leq m$. We write $\mathrm{ad}_g\, h \coloneqq [g, h]$. The standard differentiation formula for pullbacks (Agrachev & Sachkov, 2004, Section 2.5, Eq. (2.21)) gives

$$\frac{d}{dt}(\varphi_i^t)^* h \bigg|_{t=0} = [g_i, h] = \mathrm{ad}_{g_i}\, h.$$

Mixed derivatives of pullbacks by successive flows therefore give the evaluations at Y of all brackets of the form $\mathrm{ad}_{g_{i_1}} \cdots \mathrm{ad}_{g_{i_k}}\, g_j$. Every such derivative is a limit of finite linear combinations of vectors in $\mathcal{V}_Y$. Since this subspace of $\mathbb{R}^D$ is closed, the derivative also belongs to $\mathcal{V}_Y$. These brackets span the generated Lie algebra by the Jacobi identity. The rank assumption consequently implies $\mathcal{V}_Y = \mathbb{R}^D$.

Choose compositions $\Psi_1, \dots, \Psi_D$ and indices $i_1, \dots, i_D$ such that

$$v_\ell = (\Psi_\ell^* g_{i_\ell})(Y), \qquad 1 \leq \ell \leq D,$$

form a basis of $\mathbb{R}^D$. For $a = (a_1, \dots, a_D)$ near zero, define

$$C_\ell^{a_\ell} \coloneqq \Psi_\ell^{-1} \circ \varphi_{i_\ell}^{a_\ell} \circ \Psi_\ell, \qquad E(Z, a) \coloneqq C_D^{a_D} \circ \cdots \circ C_1^{a_1}(Z).$$

Each $C_\ell^{a_\ell}$ follows Ψ_ℓ, flows along g_{i_ℓ} for time a_ℓ, and follows Ψ_ℓ^{-1}. The inverse uses the same flows in reverse order with negated times. Thus E uses only the g_j, and $E(Z, 0) = Z$. Differentiating with respect to a gives

$$E(Y, 0) = Y, \qquad A \coloneqq D_a E(Y, 0) = [v_1 \quad \cdots \quad v_D] \tag{10}$$

with A invertible. On a sufficiently small compact product neighborhood of $(Y, 0)$, all trajectories defining $E(Z, a)$ lie in a common compact subset of U.

Replace each flow segment by residual layers. Number the J successive flow segments in E as

$$E(Z, a) = \varphi_{j_J}^{\tau_J(a)} \circ \cdots \circ \varphi_{j_1}^{\tau_1(a)}(Z).$$

Here j_ℓ is the field index and $\tau_\ell(a)$ is the signed time of segment ℓ. Each τ_ℓ is either a constant, from a fixed composition Ψ_k or its inverse, or a coordinate a_k. For fixed a, segment ℓ starting at an incoming state ζ is the solution, on $s \in [0,1]$, of

$$\frac{dz_\ell}{ds}(s) = \tau_\ell(a) g_{j_\ell}(z_\ell(s)), \qquad z_\ell(0) = \zeta.$$

Its endpoint is $z_\ell(1) = \varphi_{j_\ell}^{\tau_\ell(a)}(\zeta)$. This rescaling also covers negative or zero values of $\tau_\ell(a)$. Choose an integer $q \geq 1$ and approximate each such ODE by q Euler steps on the grid $s_r = r/q$:

$$z_\ell^0 = \zeta, \qquad z_\ell^{r+1} = z_\ell^r + \frac{\tau_\ell(a)}{q} g_{j_\ell}(z_\ell^r), \qquad r = 0, \dots, q-1.$$

Thus the exact flow of segment ℓ is approximated by q residual layers with coefficient $\tau_\ell(a)/q$. Starting from Z and feeding each segment's Euler endpoint into the next defines

$$E_q(Z, a) = \left(\mathrm{Id} + \frac{\tau_J(a)}{q} g_{j_J}\right)^{\circ q} \circ \cdots \circ \left(\mathrm{Id} + \frac{\tau_1(a)}{q} g_{j_1}\right)^{\circ q} (Z).$$

We use the same q for all J segments, giving qJ layers. Increasing q refines every segment's approximation; for each fixed q, the starting state Z and parameters a can vary.

On a sufficiently small compact product neighborhood of $(Y, 0)$,

$$E_q \longrightarrow E \qquad \text{in } C^1 \quad (q \to \infty), \tag{11}$$

where derivatives are taken with respect to (Z, a). All intermediate residual states remain in U for large q. To control derivatives with respect to a as well as the incoming state, adjoin the constant parameter a to each segment's ODE. The augmented state $u = (z, a) \in \mathbb{R}^{2D}$ evolves under

$$H_\ell(z, a) \coloneqq (\tau_\ell(a) g_{j_\ell}(z), 0).$$

Its time-one flow is $(z, a) \mapsto (\varphi_{j_\ell}^{\tau_\ell(a)}(z), a)$, and its Euler scheme is the one above with a unchanged. The Jacobian $P(s) = D_{u(0)} u(s)$ satisfies

$$\dot{u} = H_\ell(u), \qquad \dot{P} = DH_\ell(u) P, \qquad P(0) = I_{2D}.$$

Differentiating the Euler updates gives exactly the Euler scheme for this variational system. Smoothness on the compact neighborhoods above provides uniform bounds, so the Euler estimate in the preceding proof applies to both u and P. It gives uniform convergence in (z, a) of the maps and their Jacobians, while keeping the state iterates in U. The chain rule for the finitely many segments then gives (11).

Recover an exact residual correction. Although $E(Y, 0) = Y$, generally $E_q(Y, 0) \neq Y$. We first adjust a to recover an exact return from Y to Y. Choose $r > 0$ small enough that the closed parameter ball $\overline{B}_r(0)$ lies in the preceding neighborhood and

$$\sup_{a \in \overline{B}_r(0)} \|I - A^{-1} D_a E(Y, a)\| \leq \tfrac{1}{4}.$$

By (11), for all sufficiently large q,

$$\sup_{a \in \overline{B}_r(0)} \|I - A^{-1} D_a E_q(Y, a)\| \leq \tfrac{1}{2}, \qquad \|A^{-1}(E_q(Y, 0) - Y)\| \leq \tfrac{r}{4}.$$

Fix one such q, also large enough to keep every intermediate state in U. The map

$$T_q(a) \coloneqq a - A^{-1}(E_q(Y, a) - Y)$$

is a contraction with constant at most $1/2$ on $\overline{B}_r(0)$ and satisfies

$$\|T_q(a)\| \leq \|T_q(0)\| + \tfrac{1}{2}\|a\| \leq \tfrac{3r}{4}.$$

The contraction mapping theorem supplies a fixed point $a_q \in B_r(0)$. Thus $E_q(Y, a_q) = Y$, and the derivative bound implies that $D_a E_q(Y, a_q)$ is invertible.

Now allow the starting point Z to vary near Y. The implicit function theorem at (Y, a_q) gives a neighborhood $\mathcal{U}_Y \subseteq U$ and a C^1 map $\alpha : \mathcal{U}_Y \to B_r(0)$, with $\alpha(Y) = a_q$, such that

$$E_q(Z, \alpha(Z)) = Y \qquad \text{for every } Z \in \mathcal{U}_Y. \tag{12}$$

Shrinking $\mathcal{U}_Y$ keeps all intermediate states in U. The correction uses $L_Y := qJ$ layers: in segment ℓ, each coefficient is $\tau_\ell(\alpha(Z))/q$. These are fixed real scalars once Z is specified, so scalar closure makes every layer admissible. The integer q, the layer sequence, and $\mathcal{U}_Y$ are fixed before the next step.

Step 2: reach the correction neighborhood by Euler discretization. We now approximate a trajectory from X to Y closely enough to reach the fixed neighborhood $\mathcal{U}_Y$. By the local rank and connectedness argument proving Corollary 2.6, a piecewise-constant controlled trajectory $\gamma : [0, S] \to \mathcal{M}$ joins X to Y for some $S > 0$. Fix any such trajectory. Its image Γ is compact, so there is $\rho > 0$ such that the compact tube

$$K_\rho := \{Z \in \mathbb{R}^D : \operatorname{dist}(Z, \Gamma) \le \rho\} \subseteq \mathcal{M}.$$

Let $0 = s_0 < \cdots < s_\nu = S$ be the switching times of the fixed control, and let $h_i \in \mathcal{G}$ denote the vector field selected on $[s_{i-1}, s_i)$, so that

$$\dot\gamma(t) = h_i(\gamma(t)) \qquad \text{for } t \in (s_{i-1}, s_i).$$

These finitely many locally Lipschitz fields have a common norm bound $B \ge 1$ and Lipschitz constant $\Lambda \ge 1$ on K_ρ. Indeed, continuity and compactness give

$$B := \max\left\{1, \max_{1\le i\le \nu} \max_{Z\in K_\rho} \|h_i(Z)\|\right\} < \infty.$$

The finite-cover argument in the preceding proof gives $\eta > 0$ and a common L_{loc} for pairs of points in K_ρ at distance less than η. For more distant pairs, the norm bound gives $\|h_i(Z) - h_i(W)\| \le 2B \le (2B/\eta)\|Z - W\|$. Hence $\Lambda = \max\{1, L_{\mathrm{loc}}, 2B/\eta\}$ works on all of K_ρ.

Choose a finite mesh $0 = t_0 < \cdots < t_{L_E} = S$ containing every switching time. Write $\Delta_k = t_{k+1} - t_k$, $h = \max_k \Delta_k$, and let $H_k \in \{h_1, \dots, h_\nu\}$ be the field active on $[t_k, t_{k+1})$. Forward Euler gives $Z^0 = X$ and $Z^{k+1} = Z^k + \Delta_k H_k(Z^k)$. On each mesh interval, the integral equation for γ yields

$$\|\gamma(t_{k+1}) - \gamma(t_k) - \Delta_k H_k(\gamma(t_k))\| \le \tfrac{1}{2}\Lambda B \Delta_k^2.$$

Consequently, while $Z^k \in K_\rho$, the errors $e_k := \|Z^k - \gamma(t_k)\|$ satisfy

$$e_{k+1} \le (1 + \Lambda\Delta_k)e_k + \tfrac{1}{2}\Lambda B\Delta_k^2.$$

Since $e_0 = 0$ and $\Delta_k^2 \le h\Delta_k$, induction gives

$$e_k \le \frac{Bh}{2}\left(\prod_{j<k}(1 + \Lambda\Delta_j) - 1\right) \le \frac{Bh}{2}(e^{\Lambda t_k} - 1) \le Ch, \qquad C := \frac{B}{2}(e^{\Lambda S} - 1). \tag{13}$$

Choose $\varepsilon > 0$ with $\{Z : \|Z - Y\| < \varepsilon\} \subseteq \mathcal{U}_Y$, and refine the mesh until $Ch < \min\{\rho, \varepsilon\}$. This can be done with finitely many steps while retaining every switching time. There can be no first iterate outside K_ρ: the bound for that iterate, computed from its predecessor inside the tube, would give distance less than ρ from Γ. Thus all iterates lie in $\mathcal{M}$, (13) holds throughout, and $\widetilde{Y} := Z^{L_E}$ lies in the fixed neighborhood $\mathcal{U}_Y$. Each Euler step is an admissible residual layer, since $\Delta_k H_k \in \mathcal{G}$ by positive scaling.

Step 3: concatenate the Euler and correction layers. Step 2 gives an Euler endpoint $\widetilde{Y} \in \mathcal{U}_Y$. Set $a = \alpha(\widetilde{Y})$ and append the L_Y residual layers defining $E_q(\cdot, a)$ to this Euler network. Their fields all belong to $\mathcal{G}$: each is a fixed real scalar multiple of one of the fields selected in Step 1. By (12), the final output is $E_q(\widetilde{Y}, a) = Y$, and all intermediate configurations remain in $\mathcal{M}$. Together with the L_E Euler layers, this gives a finite admissible network with at most $L_E + L_Y$ layers. All its scalar coefficients are fixed for the prescribed interpolation problem. In a batch application, these scalars are shared across all samples. □

F UNIVERSAL FINITE-DEPTH INTERPOLATION

F.1 DIRECT SPANNING

Proposition F.1 (Universal finite-depth interpolation)**.** *Fix $N \geq 1$ and a configuration space $\mathcal{M}_N$ as in Definition 2.4, and assume that $\mathcal{F}$ allows real scalar multiplication. Suppose there exist finitely many parameters $\theta_1, \dots, \theta_m$ such that, at each $X \in \mathcal{M}_N$, one can select Np of their lifted fields that are C^1 near X and whose values at X form a basis of $\mathbb{R}^{Np}$. Then $\mathcal{F}$ has universal interpolation at finite depth on $\mathcal{M}_N$, using these same parameters.*

Proof. Fix the parameters $\theta_1, \dots, \theta_m$ in the hypothesis and set

$$\mathcal{G}_N \coloneqq \{a f_{\theta_j}^{(N)} : a \in \mathbb{R},\ 1 \leq j \leq m\}.$$

Every field in $\mathcal{G}_N$ is locally Lipschitz on $\mathbb{R}^{Np}$, hence also on $\mathcal{M}_N$. The family is closed under real scalar multiplication and, on $\mathcal{M}_N$, satisfies the pointwise basis condition of Corollary D.1 with $D = Np$. Thus Proposition E.1 joins every pair of configurations by residual layers $\mathrm{Id} + c_k f_{\theta_{j_k}}^{(N)}$, with all intermediate configurations in $\mathcal{M}_N$. The parameters were fixed before the endpoints; only the depth, indices, and scalar coefficients depend on them. This is universal interpolation at finite depth as defined in Section 2.3. □

F.2 EXTENSION UNDER A LIE-RANK CONDITION

Proposition 2.10 (Universal finite-depth interpolation under Lie rank)**.** *Fix $N \geq 1$ and a configuration space $\mathcal{M}_N$ as in Definition 2.4, and assume that $\mathcal{F}$ allows real scalar multiplication. Suppose there exist finitely many parameters $\theta_1, \dots, \theta_m$ such that, at each $X \in \mathcal{M}_N$, an index set $I_X \subseteq \{1, \dots, m\}$ selects lifted fields smooth near X with*

$$\mathrm{Lie}\big(f_{\theta_j}^{(N)} : j \in I_X\big)(X) = \mathbb{R}^{Np}.$$

Then $\mathcal{F}$ has universal interpolation at finite depth on $\mathcal{M}_N$, using these same parameters.

Proof. Fix the parameters $\theta_1, \dots, \theta_m$ in the hypothesis and set

$$\mathcal{G}_N \coloneqq \{c f_{\theta_j}^{(N)} : c \in \mathbb{R},\ 1 \leq j \leq m\}.$$

This family is locally Lipschitz and closed under real scalar multiplication. At each $X \in \mathcal{M}_N$, the fields indexed by I_X form a finite subfamily that is smooth near X and has full Lie rank there. Thus Proposition 2.9, with $D = Np$, joins every pair $X, Y \in \mathcal{M}_N$ by finitely many residual layers

$$\mathrm{Id} + c_k f_{\theta_{j_k}}^{(N)}, \qquad c_k \in \mathbb{R}, \quad j_k \in \{1, \dots, m\},$$

with all intermediate configurations in $\mathcal{M}_N$. The parameters were fixed before the endpoints; only the depth, indices, and scalar coefficients depend on them. This is universal interpolation at finite depth using the same dictionary. □

G RESIDUAL NETWORKS

For $N \geq 1$ and $d \geq 2$, define the generic configuration space

$$\mathcal{C}_{N,d} \coloneqq \left\{(x_1, \dots, x_N) \in (\mathbb{R}^d)^N : x_i \neq 0 \text{ and } x_i \notin \mathbb{R}x_j \text{ for all } i \neq j\right\}.$$

It is nonempty and open: distinct nonzero projective directions give examples, while the excluded zero vectors and collinear pairs form a finite union of closed sets.

Theorem G.1 (Exact interpolation by bias-free ResNets)**.** *Let $d \geq 3$ and $r \geq 1$. For every $N \geq 1$, the family of vector fields in* (2) *has the exact interpolation property on $\mathcal{C}_{N,d}$, both at continuous depth in the sense of Definition 2.3 and at finite depth in the sense of Section 2.3. One active hidden unit per layer suffices, and every intermediate configuration remains in $\mathcal{C}_{N,d}$.*

The proof is given in Section G.2.

Remark G.2. The assumption $d \geq 3$ in Theorem G.1 ensures that $\mathcal{C}_{N,d}$ is connected. In dimension two, $\mathcal{C}_{N,2}$ is not connected when $N \geq 2$, but exact interpolation at both continuous and finite depth holds on each connected component. These claims are proved in Section G.2.

With identity activation, both continuous- and finite-depth networks apply a common linear map to all samples and preserve their linear relations. For $N > d$, the lifted Lie algebra has rank at most $d^2 < Nd$, so the analytic converse in Corollary 2.6 rules out continuous-depth exact interpolation on any nonempty open subset of $(\mathbb{R}^d)^N$. The preserved linear relations give the same obstruction at finite depth, thereby also excluding universal interpolation at both depths. The proof is given in Section G.1. The following theorem establishes universal interpolation for bias-free ReLU ResNets on $\mathcal{C}_{N,d}$, using a fixed family of updates that works for every N.

Theorem G.3 (Universal interpolation by bias-free ResNets)**.** *Let $d \geq 3$ and $r \geq 1$, and let $e_1, \dots, e_d$ be the standard basis of $\mathbb{R}^d$. Define the fixed family*

$$\mathcal{R}_d \coloneqq \left\{ g_{ij}^{\varepsilon} : x \mapsto e_i \sigma(\varepsilon x_j) : 1 \leq i, j \leq d,\ \varepsilon \in \{-1, 1\} \right\}. \tag{14}$$

For every $N \geq 1$, the ResNet family in (2) *has universal interpolation on $\mathcal{C}_{N,d}$, both at continuous depth in the sense of Definition 2.4 and at finite depth in the sense of Section 2.3, using the same $2d^2$ generators from $\mathcal{R}_d$. At continuous depth, each generator may be used in either sign. These generators are independent of N and require only one active hidden unit per layer.*

The proof in Section G.3 realizes the flows used in Theorem G.1 by composing flows from the fixed family (14). Each such flow is itself exactly a residual layer with a suitable scalar gain.

Fixed families giving exact interpolation are known for general neural ODEs (Cuchiero et al., 2020; Grong & Vega-Molino, 2026) and for certain smooth-activation ResNets (Tabuada & Gharesifard, 2023). For ReLU, Duan & Cai (2025) prove universal approximation using a finite family of linear fields, translations, and ReLU, while Cai & Duan (2025) prove exact interpolation using affine fields and ReLU. Cai (2024) also constructs a finite vocabulary of maps for universal approximation. Theorem G.3 gives an explicit family of bias-free, width-one ReLU updates, reusable for every N, with exact interpolation at both continuous and finite depth.

Remark G.4 (Comparison with interpolation using biases). Ruiz-Balet & Zuazua (2023) prove exact interpolation with biases for $d \geq 2$, using one active neuron at a time. Their admissible configurations require only pairwise distinct points, allowing zero vectors and collinear pairs. Their construction chooses the weights according to the dataset and does not assert interpolation with a fixed finite family. It appears to adapt directly to universal interpolation by adding coordinate translations to (14) and replacing its data-dependent flows by affine conjugations of fixed ReLU flows. These translations are unavailable in the bias-free architecture considered here.

G.1 Linear activation and the obstruction to interpolation

The following lemma also applies to the shared linear dynamics of attention with sequence length $n = 1$.

Lemma G.5 (Linear-relation obstruction)**.** *Let $d \geq 1$, $N > d$, and let $\mathcal{M} \subseteq (\mathbb{R}^d)^N$ be nonempty and open. For every $X \in \mathcal{M}$, there are arbitrarily close targets $Y \in \mathcal{M}$ in the same connected component that cannot be obtained from X by a shared linear map. Thus neither controlled linear ODEs nor finite compositions of linear residual layers can interpolate every pair of configurations in $\mathcal{M}$.*

Proof. For $A \in \mathbb{R}^{d \times d}$, define $f_A(x) = Ax$ and its lift $F_A(X) = (Ax^1, \dots, Ax^N)$. The bracket convention gives

$$[F_A, F_B] = F_{BA-AB}.$$

Therefore, for any family of linear fields,

$$\mathrm{Lie}\{F_{A_\theta} : \theta \in \Theta\}(X) \subseteq \{(Ax^1, \dots, Ax^N) : A \in \mathbb{R}^{d \times d}\}.$$

The space on the right is the image of a linear map from $\mathbb{R}^{d \times d}$, so its dimension is at most $d^2 < Nd$.

To establish the obstruction for both depth conventions, observe that a continuous linear network applies a common fundamental matrix to all samples, and a finite linear residual network applies the

common matrix $(I + A_{L-1}) \cdots (I + A_0)$. In either case, write A_* for the common output matrix. Every input linear relation is retained at the output:

$$\sum_{i=1}^{N} c_i x^i = 0 \quad \Longrightarrow \quad \sum_{i=1}^{N} c_i A_* x^i = A_* \sum_{i=1}^{N} c_i x^i = 0.$$

Fix any $X = (x^1, \ldots, x^N) \in \mathcal{M}$. Since $N > d$, choose a nonzero relation $\sum_i c_i x^i = 0$ and an index j with $c_j \neq 0$. Choose any nonzero $u \in \mathbb{R}^d$ and set $y_j = x_j + \varepsilon u$ and $y_i = x_i$ for $i \neq j$. For sufficiently small $\varepsilon \neq 0$, openness of $\mathcal{M}$ ensures that the segment from X to Y lies in $\mathcal{M}$. Thus Y is in the same connected component as X, but

$$\sum_i c_i y_i = \varepsilon c_j u \neq 0.$$

No shared linear map can send X to Y. Taking arbitrarily small nonzero ε proves the lemma. □

For ResNets with identity activation, the vector fields take the form

$$f_{U,V}(x) = U\sigma(V^\top x) = UV^\top x, \qquad U, V \in \mathbb{R}^{d \times r}.$$

Since Lemma G.5 applies to arbitrary linear fields $x \mapsto Ax$, it applies in particular to $A = UV^\top$, for every hidden width r. Thus, when $N > d$, increasing the hidden width cannot remove the obstruction to interpolation, at either continuous or finite depth.

G.2 Proof of Theorem G.1

Proof. We first prove the stronger statement that exact interpolation holds at both continuous and finite depth on every connected component of $\mathcal{C}_{N,d}$ for all $d \geq 2$. For $X = (x^1, \ldots, x^N)$, write

$$f^{(N)}_{U,V}(X) \coloneqq \big(U\sigma(V^\top x^1), \ldots, U\sigma(V^\top x^N)\big).$$

The width-one subfamily

$$G_{u,v}(X) \coloneqq \big(u\sigma(v^\top x^1), \ldots, u\sigma(v^\top x^N)\big), \qquad u, v \in \mathbb{R}^d,$$

is contained in the width-r family for every $r \geq 1$: take $U = (u, 0, \ldots, 0)$ and $V = (v, \ldots, v)$.

Fix $X \in \mathcal{C}_{N,d}$ and set

$$\mathcal{H}_i \coloneqq \{v \in \mathbb{R}^d : v^\top x^i = 0\}, \qquad \mathcal{O}_X \coloneqq \mathbb{R}^d \setminus \bigcup_{j=1}^{N} \mathcal{H}_j.$$

The hyperplanes $\mathcal{H}_1, \ldots, \mathcal{H}_N$ are pairwise distinct. Hence, for each i, we may choose

$$w \in \mathcal{H}_i \setminus \bigcup_{j \neq i} \mathcal{H}_j.$$

For sufficiently small $\delta > 0$, the four vectors

$$v_k \coloneqq w + k\delta x^i, \qquad k \in \{-2, -1, 1, 2\},$$

all belong to $\mathcal{O}_X$, and, for $j \neq i$, the signs of $v_k^\top xj$ do not depend on k. Therefore, for every $u \in \mathbb{R}^d$,

$$\begin{aligned} G_{u,v_2}(X) - G_{u,v_1}(X) - G_{u,v_{-1}}(X) + G_{u,v_{-2}}(X) \\ = \delta \|x^i\|^2 (0, \ldots, 0, u, 0, \ldots, 0), \end{aligned} \tag{15}$$

where u occupies the i-th block on the right-hand side. Indeed, in every block $j \neq i$, the ReLU is either zero or affine as a function of k, so its finite difference vanishes. In the i-th block, $v_k^\top x^i = k\delta \|x^i\|^2$, and the finite difference is $\delta \|x^i\|^2 u$.

Since i and u are arbitrary, (15) gives

$$\operatorname{span}\left\{ f^{(N)}_{U,V}(X) : U, V \in \mathbb{R}^{d \times r} \right\} = (\mathbb{R}^d)^N. \tag{16}$$

In fact, the width-one subfamily $\{G_{u,v}\}$ already has this full pointwise span. Moreover, every field used in (15) is smooth on a neighborhood of X, because $v_k^\top x^j \neq 0$ for all j. Thus, at each $X \in \mathcal{C}_{N,d}$, we may select Nd locally smooth width-one fields whose values form a basis of $(\mathbb{R}^d)^N$.

The lifted ResNet family is closed under negation:

$$-f_{U,V}^{(N)} = f_{-U,V}^{(N)}.$$

Since $\mathcal{C}_{N,d}$ is open, Corollary D.1 and (16) give exact interpolation at continuous depth on every connected component of $\mathcal{C}_{N,d}$. For finite depth, the family is also closed under real scalar multiplication, since $af_{U,V}^{(N)} = f_{aU,V}^{(N)}$. The locally smooth width-one fields selected above therefore satisfy the hypotheses of Proposition E.1, with $D = Nd$. It gives exact interpolation at finite depth, with every intermediate configuration in the same component. Each scalar gain is absorbed into the output matrix U.

It remains to prove that $\mathcal{C}_{N,d}$ is connected when $d \geq 3$. Write $x^i = r_i u_i$, where $r_i > 0$ and $u_i \in \mathbb{S}^{d-1}$. This identifies $\mathcal{C}_{N,d}$ with

$$(0,\infty)^N \times \mathcal{D}_{N,d}, \qquad \mathcal{D}_{N,d} \coloneqq \left\{(u_1,\dots,u_N) \in (\mathbb{S}^{d-1})^N : u_i \neq \pm u_j \text{ for } i \neq j\right\}.$$

Let $u = (u_1,\dots,u_N)$ and $\widetilde{u} = (\widetilde{u}_1,\dots,\widetilde{u}_N)$ belong to $\mathcal{D}_{N,d}$. Choose points $p_1,\dots,p_N \in \mathbb{S}^{d-1}$ inductively, with each p_i different from the finitely many points

$$\{\pm u_j, \pm\widetilde{u}_j : 1 \leq j \leq N\} \cup \{\pm p_j : 1 \leq j < i\}.$$

Because $d - 1 \geq 2$, a sphere with finitely many points removed is path connected. We may consequently move $u_1,\dots,u_N$, one at a time, to $p_1,\dots,p_N$ while avoiding the directions and antipodal directions of the points kept fixed. Moving the p_i one at a time to $\widetilde{u}_i$ gives a path in $\mathcal{D}_{N,d}$ from u to $\widetilde{u}$. The factor $(0,\infty)^N$ is path connected as well. Hence $\mathcal{C}_{N,d}$ is path connected for $d \geq 3$, and the componentwise result proved above yields the theorem. For $d = 2$ and $N \geq 2$, the continuous function $X \mapsto \det(x^1, x^2)$ never vanishes on $\mathcal{C}_{N,2}$ by definition of $\mathcal{C}_{N,2}$, and takes both signs, so this space is disconnected. □

G.3 Proof of Theorem G.3

Proof. Fix $d \geq 3$, $N \geq 1$, and the family $\mathcal{R}_d$ in the statement. Write ϕ_g^t for the flow of a field g, acting componentwise on configurations. For $g(x) = u\sigma(v^\top x)$ and $\alpha = v^\top u$, this flow is defined for every $t \in \mathbb{R}$ and satisfies

$$\phi_g^t(x) = x + \beta_\alpha(t) g(x), \qquad \beta_\alpha(t) = \begin{cases} (e^{\alpha t} - 1)/\alpha, & \alpha \neq 0, \\ t, & \alpha = 0. \end{cases} \tag{17}$$

Indeed, $q(t) = v^\top \phi_g^t(x)$ solves $\dot{q} = \alpha\sigma(q)$: it remains constant when $q(0) \leq 0$ and equals $e^{\alpha t} q(0)$ otherwise. Integrating the equation for $\phi_g^t(x)$ gives (17).

We first show that the fixed family can realize every linear map $A \in GL^+(d) \coloneqq \{A : \det A > 0\}$ while keeping a prescribed configuration in $\mathcal{C}_{N,d}$. Put $E_{ij} = e_i e_j^\top$. The flows of g_{ij}^{+1} and g_{ij}^{-1} commute: for $i \neq j$ they preserve x_j, and for $i = j$ they act on separate invariant half-lines. Since $g_{ij}^{+1}(x) - g_{ij}^{-1}(x) = E_{ij}x$, we have

$$\phi_{g_{ij}^{+1}}^t \circ \phi_{g_{ij}^{-1}}^{-t}(x) = e^{tE_{ij}} x. \tag{18}$$

Every matrix in $GL^+(d)$ is a finite product of these exponentials. To see this, the derivative at zero of $(t_{ij}) \mapsto \prod_{i,j} \exp(t_{ij}E_{ij})$, with any fixed ordering of the factors, is the isomorphism $(t_{ij}) \mapsto \sum_{i,j} t_{ij}E_{ij}$. The inverse function theorem shows that the subgroup generated by these exponentials contains a neighborhood of the identity. It is therefore open, and connectedness of $GL^+(d)$ implies that it is the whole group.

Although the endpoints in (18) are related by a linear map, its two intermediate flows need not preserve noncollinearity. This is resolved by subdivision. For $X \in \mathcal{C}_{N,d}$, the linear path $\Gamma = \{e^{stE_{ij}}X : 0 \leq s \leq 1\}$ is a compact subset of $\mathcal{C}_{N,d}$. Choose $\delta > 0$ such that its δ-neighborhood is contained

in $\mathcal{C}_{N,d}$, and put $M = \max_{Z\in\Gamma} \|Z\|$. Replace the pair in (18) by K repetitions with $h = t/K$. Their successive endpoints are exactly $e^{khE_{ij}}X$. Every lifted field in $\mathcal{R}_d$ satisfies $\|g^{(N)}(Z)\| \le \|Z\|$. Hence, during either subflow of a pair, Gronwall's inequality bounds the distance from the pair's initial point on Γ by $M(e^{2|h|} - 1)$. For sufficiently large K, this is less than δ. All intermediate configurations then stay in $\mathcal{C}_{N,d}$. Applying this construction to each exponential factor realizes A with finitely many flows from the same fixed family, with the required constraint throughout.

Next, every width-one field $g(x) = u\sigma(v^\top x)$ is conjugate, up to a real scalar, to one of

$$H(x) = e_1\sigma(x_1), \qquad T(x) = e_2\sigma(x_1),$$

which belong to $\mathcal{R}_d$. Zero fields require no flow, so assume $u, v \neq 0$. If $\alpha = v^\top u \neq 0$, choose $A \in GL^+(d)$ whose first column is u/α and whose remaining columns form a basis of $\ker v^\top$. Reversing one of those columns if necessary makes the determinant positive. Then $e_1^\top A^{-1} = v^\top$ and

$$g(x) = \alpha AH(A^{-1}x), \qquad \phi_g^t = A \circ \phi_H^{\alpha t} \circ A^{-1}.$$

If $\alpha = 0$, choose the first column a_1 so that $v^\top a_1 = 1$, take the second column to be u, and complete these columns to a basis using vectors in $\ker v^\top$. Since $d \ge 3$, the sign of a further column can be chosen to ensure $\det A > 0$. This gives

$$g(x) = AT(A^{-1}x), \qquad \phi_g^t = A \circ \phi_T^t \circ A^{-1}.$$

Suppose a segment of the flow of g starting at X remains in $\mathcal{C}_{N,d}$. Realize A^{-1} and A by the admissible compositions just constructed. Between them, the flow of H or T is the image under A^{-1} of the prescribed segment, so it also stays in $\mathcal{C}_{N,d}$. Thus that segment can be replaced exactly by finitely many flows from $\mathcal{R}_d$, without leaving the configuration space.

By Theorem G.1 at width one, any $X, Y \in \mathcal{C}_{N,d}$ are joined inside this set by a finite concatenation of flows of fields $x \mapsto u\sigma(v^\top x)$. Replacing each segment as above gives a finite concatenation of flows from $\mathcal{R}_d$ with signed times $\tau_1, \dots, \tau_L$. Discard zero-time segments and run each remaining segment on the batch with $\operatorname{sgn}(\tau_k) g_k^{(N)}$ for duration $|\tau_k|$, where $g_k \in \mathcal{R}_d$ is its generator. With $S = \sum_k |\tau_k|$, this gives a piecewise-constant control in $\{\pm g^{(N)} : g \in \mathcal{R}_d\}$ on $[0, S]$, whose trajectory satisfies $Z(0) = X$, $Z(S) = Y$ and remains in $\mathcal{C}_{N,d}$. If no segment remains, a sufficiently short flow of any generator followed by its reverse gives the identity with $S > 0$ and stays in this open set. The $2d^2$ generators were fixed before both N and the endpoints, proving the continuous-depth conclusion in Definition 2.4.

Finally, (17) turns each of these flows exactly into a residual layer $\mathrm{Id} + \beta_{\varepsilon\delta_{ij}}(\tau_k) g_{ij}^{\varepsilon}$. The resulting finite network uses the same family and has the same intermediate configurations at the layer endpoints. This proves universal interpolation at finite depth directly, without a discretization error. All blocks embed into every hidden width $r \ge 1$ by setting unused output columns to zero. □

The preceding proof uses changes of activation patterns along the trajectory. The local condition of Corollary 2.7 is not satisfied by this fixed family when $N > d$: when all inputs lie in one open orthant, its fields are locally linear, so their lifted Lie algebra has rank at most $d^2 < Nd$. More generally, the same observation applies to any finite family of bias-free ReLU blocks on an open cone where all their activation patterns are constant. This does not contradict the analytic converse in Remark 2.8, since these fields are not analytic across their activation hyperplanes.

H Proof of Lemma 3.1

Lemma 3.1 (Pointwise spanning by noncausal self-attention)**.** *Let $d \ge 2$, $n \ge 2$, and $N \ge 1$. At every $X \in \mathcal{A}_{N,n,d}$, the single-head batch fields satisfy*

$$\operatorname{span}\left\{ f_{W,V}^{(N)}(X) : \operatorname{rank} W, \operatorname{rank} V \le 1 \right\} = \mathbb{R}^{Nnd}.$$

Proof. The pointwise rank argument below only requires $d \ge 2$. We will also prove exact interpolation at both continuous and finite depth on every connected component of $\mathcal{A}_{N,n,2}$. Fix $N \ge 1$ and $d \ge 2$. The set $\mathcal{A}_{N,n,d}$ is open: its complement is a finite union of the closed sets where one token is

zero or where all 2×2 minors of a pair of token columns vanish. Hence its connected components are open. Take $X = (x_i^\ell)_{\ell,i} \in \mathcal{A}_{N,n,d}$. It is enough to use one head: set $V_h = 0$ for $h \geq 2$, and denote the effective matrices of the remaining head by $W, V \in \mathbb{R}^{d\times d}$. We will further restrict both matrices to have rank at most one. Let $f_{W,V}^{(N)}$ be the corresponding lifted vector field and define its pointwise span

$$\mathcal{S}_X := \operatorname{span}\left\{ f_{W,V}^{(N)}(X) : \operatorname{rank} W, \operatorname{rank} V \leq 1 \right\} \subseteq (\mathbb{R}^d)^{Nn}.$$

This is a finite-dimensional linear subspace and is therefore closed.

For $b \in \mathbb{R}^d$ and $1 \leq \ell \leq N$, introduce the softmax barycenter

$$z_\ell(b) := \frac{\sum_{j=1}^n \exp\big(\langle b, x_j^\ell\rangle\big) x_j^\ell}{\sum_{j=1}^n \exp\big(\langle b, x_j^\ell\rangle\big)}. \tag{19}$$

The (i,ℓ) block of the lifted field is then

$$f_{W,V}^{(N)}(X)_{i,\ell} = V z_\ell(W x_i^\ell).$$

We show that an arbitrary block can be controlled independently. Fix a target (i_0, ℓ_0) and put $q := x_{i_0}^{\ell_0}$. First choose $a \in \mathbb{R}^d$ such that

$$\langle a, x_j^\ell\rangle \neq \langle a, x_k^\ell\rangle \qquad (1 \leq \ell \leq N,\ j \neq k). \tag{20}$$

Such an a exists: each forbidden equality defines the proper hyperplane $(x_j^\ell - x_k^\ell)^\perp$, and only finitely many hyperplanes are excluded.

Next choose $\alpha \in q^\perp$ satisfying

$$\langle \alpha, x_i^\ell\rangle \neq 0 \qquad \text{whenever } (i,\ell) \neq (i_0,\ell_0). \tag{21}$$

Indeed, every other token $r = x_i^\ell$ is not collinear with q, so $q^\perp \cap r^\perp$ is a proper linear subspace of $q^\perp$. A finite union of proper linear subspaces cannot cover $q^\perp$.

Let $V \in \mathbb{R}^{d\times d}$ have rank at most one; its value will be chosen below. For $t > 0$, define the rank-one score matrices

$$W_t^\pm := a\left(t\alpha \pm \frac{q}{\|q\|^2}\right)^\top. \tag{22}$$

Since $\alpha \perp q$, they satisfy $W_t^\pm q = \pm a$. Consequently, at the target block,

$$f_{W_t^+,V}^{(N)}(X)_{i_0,\ell_0} - f_{W_t^-,V}^{(N)}(X)_{i_0,\ell_0} = Vc, \qquad c := z_{\ell_0}(a) - z_{\ell_0}(-a).$$

The vector c is nonzero. To see this, let

$$\psi(s) := \log\left(\sum_{j=1}^n \exp\big(s\langle a, x_j^{\ell_0}\rangle\big)\right).$$

Then $\psi'(s) = \langle a, z_{\ell_0}(sa)\rangle$, and

$$\begin{aligned}\psi''(s) &= \sum_{j=1}^n p_j(s)\,\langle a, x_j^{\ell_0}\rangle^2 - \left(\sum_{j=1}^n p_j(s)\,\langle a, x_j^{\ell_0}\rangle\right)^2 \\ &= \sum_{j=1}^n p_j(s)\left(\langle a, x_j^{\ell_0}\rangle - \sum_{k=1}^n p_k(s)\,\langle a, x_k^{\ell_0}\rangle\right)^2 > 0,\end{aligned}$$

where $p_j(s) := \frac{e^{s\langle a, x_j^{\ell_0}\rangle}}{\sum_{k=1}^n e^{s\langle a, x_k^{\ell_0}\rangle}}$. Strict positivity follows from $n \geq 2$ and (20). Hence

$$\langle a, c\rangle = \psi'(1) - \psi'(-1) > 0.$$

Consider now a non-target query $r = x_i^\ell$. From (22),

$$W_t^\pm r = \lambda_t^\pm(r)a, \qquad \lambda_t^\pm(r) \coloneqq t\langle \alpha, r\rangle \pm \frac{\langle q, r\rangle}{\|q\|^2}.$$

By (21), both $\lambda_t^+(r)$ and $\lambda_t^-(r)$ tend to the same signed infinity. If the sign is positive, both softmax barycenters converge to the unique key in sequence ℓ maximizing $\langle a, x_j^\ell\rangle$; if it is negative, both converge to the unique minimizing key. The uniqueness is guaranteed by (20). Therefore

$$z_\ell(W_t^+ r) - z_\ell(W_t^- r) \longrightarrow 0 \qquad (t \to \infty)$$

at every non-target block.

Let $E_{i_0,\ell_0}(u)$ denote the vector in $(\mathbb{R}^d)^{Nn}$ whose (i_0, ℓ_0) block equals u and whose other blocks vanish. The preceding limits give

$$f^{(N)}_{W_t^+,V}(X) - f^{(N)}_{W_t^-,V}(X) \longrightarrow E_{i_0,\ell_0}(Vc).$$

Every vector on the left belongs to $\mathcal{S}_X$, and this subspace is closed. Given any $u \in \mathbb{R}^d$, choose the value matrix of rank at most one

$$V \coloneqq \frac{uc^\top}{\|c\|^2},$$

for which $Vc = u$. Thus $E_{i_0,\ell_0}(u)$ belongs to $\mathcal{S}_X$. Since the target block and u were arbitrary,

$$\mathcal{S}_X = (\mathbb{R}^d)^{Nn}. \tag{23}$$

□

Theorem H.1 (Exact interpolation by noncausal self-attention)**.** *Let $d \geq 3$, $n \geq 2$, and $H \geq 1$. For every $N \geq 1$, the noncausal self-attention family in* (3) *has the exact interpolation property on $\mathcal{A}_{N,n,d}$, both at continuous depth in the sense of Definition 2.3 and at finite depth in the sense of Section 2.3. The conclusion holds for any prescribed positive head widths. The token dimension remains d, and every intermediate configuration remains in $\mathcal{A}_{N,n,d}$.*

Proof. Fix $N \geq 1$ and $X \in \mathcal{A}_{N,n,d}$. The attention vector fields are analytic, because every softmax denominator is strictly positive. Negating all value matrices negates the field, so the family is closed under negation. Equation (23) supplies a finite collection of actual rank-one attention fields whose values form a basis at X. In particular, the limiting argument only establishes a full finite-dimensional span: the selected basis fields have finite parameter values, and no infinite-temperature layer is used. Since X was arbitrary, Corollary D.1 gives exact interpolation at continuous depth on every connected component of $\mathcal{A}_{N,n,d}$. Multiplying the value matrix by any real scalar scales the field by that scalar, so the same rank-one family satisfies Proposition E.1, with $D = Nnd$, and therefore gives exact interpolation at finite depth, with every intermediate configuration in the same component. Absorbing the scalar gains into V preserves its rank. Rank-one factorizations of W, V realize each layer with a head of width one; setting unused value matrices to zero embeds it into every prescribed H-head architecture.

It remains to determine these components. Put $m \coloneqq Nn$. Writing each token as $x_k = r_k u_k$, with $r_k > 0$ and $u_k \in \mathbb{S}^{d-1}$, gives a homeomorphism

$$\mathcal{A}_{N,n,d} \simeq (0,\infty)^m \times \mathcal{D}_{m,d}, \qquad \mathcal{D}_{m,d} \coloneqq \left\{(u_1, \ldots, u_m) \in (\mathbb{S}^{d-1})^m : u_k \neq \pm u_l \text{ for } k \neq l\right\}.$$

For $d \geq 3$, the path-connectedness argument in Section G.2 applies with N replaced by m. Hence $\mathcal{A}_{N,n,d} = \mathcal{C}_{m,d}$ is path connected, proving the theorem.

□

I Uniform direct spanning on the generic attention domain

Fix an H-head architecture with positive integer widths $r_1, \ldots, r_H$. Its full projection parameter space is

$$\Theta = \prod_{h=1}^{H} \left((\mathbb{R}^{r_h \times d})^3 \times \mathbb{R}^{d \times r_h}\right) \simeq \mathbb{R}^p, \qquad p = 4d\sum_{h=1}^{H} r_h. \tag{24}$$

Write $\theta = (Q_h, K_h, U_h, O_h)_{h=1}^H$, set $W_h = K_h^\top Q_h/\sqrt{r_h}$ and $V_h = O_h U_h$, and denote the corresponding batch field by $F_\theta = f^{(N)}_{\mathbf{W},\mathbf{V}}$. Thus Θ parametrizes the query, key, value, and output projections, not independent effective matrices W_h, V_h.

Lemma I.1 (Uniform direct spanning on the generic domain)**.** *Let $d \geq 3$, $n \geq 2$, and $N \geq 1$, and put $D \coloneqq Nnd$. Draw $(\theta_1, \ldots, \theta_{2D})$ from any jointly absolutely continuous law on Θ^{2D}. Then, with probability one over this single draw,*

$$\operatorname{span}\{F_{\theta_k}(X) : 1 \leq k \leq 2D\} = \mathbb{R}^D \qquad \text{simultaneously for every } X \in \mathcal{A}_{N,n,d}. \tag{25}$$

Thus, the same frozen finite dictionary is full rank at every generic configuration. In particular, it satisfies the uniform Lie-rank condition without using any Lie brackets, with $m_N \leq 2D = 2Nnd$.

We first isolate the general principle behind this lemma. Amir et al. (2023, Theorem A.2) establish a more general finite-witness theorem using tools from o-minimal geometry. For accessibility, we state only its consequence for jointly real analytic maps on open Euclidean domains, which covers all applications below. We apply their argument to covectors annihilating the sampled fields and include the resulting dimension-counting proof.

Theorem I.2 (Uniform random spanning principle)**.** *Let $p, q, r \geq 1$. Let $\Theta \subseteq \mathbb{R}^p$ be a nonempty, connected open set, and let $\mathcal{X} \subseteq \mathbb{R}^q$ be a nonempty open set. Suppose that $G : \Theta \times \mathcal{X} \to \mathbb{R}^r$ is jointly real analytic and that*

$$\operatorname{span}\{G(\theta, X) : \theta \in \Theta\} = \mathbb{R}^r \qquad \text{for every } X \in \mathcal{X}. \tag{26}$$

For every $m \geq q + r$, the exceptional set

$$\mathcal{B}_m \coloneqq \{(\theta_1, \ldots, \theta_m) \in \Theta^m : \operatorname{span}\{G(\theta_k, X) : 1 \leq k \leq m\} \neq \mathbb{R}^r \text{ for some } X \in \mathcal{X}\}$$

satisfies

$$\dim_{\mathrm{H}} \mathcal{B}_m \leq mp - m + q + r - 1 \leq mp - 1. \tag{27}$$

Here $\dim_{\mathrm{H}}$ denotes Hausdorff dimension. In particular, $\mathcal{B}_m$ has Lebesgue measure zero, so every jointly absolutely continuous random bank in Θ^m spans $\mathbb{R}^r$ simultaneously at every $X \in \mathcal{X}$ with probability one.

Proof. For $(X, \lambda) \in \mathcal{X} \times \mathbb{S}^{r-1}$, define

$$h_{X,\lambda}(\theta) \coloneqq \langle \lambda, G(\theta, X) \rangle.$$

By (26), this analytic function is not identically zero on the connected set Θ. The zero set $Z_{X,\lambda}$ of this nontrivial analytic function therefore satisfies

$$\dim_{\mathrm{H}} Z_{X,\lambda} \leq p - 1.$$

Consider the incidence set

$$\mathcal{I}_m \coloneqq \left\{(\theta_1, \ldots, \theta_m, X, \lambda) \in \Theta^m \times \mathcal{X} \times \mathbb{S}^{r-1} : \theta_k \in Z_{X,\lambda} \text{ for every } 1 \leq k \leq m\right\}.$$

Joint analyticity ensures that $\mathcal{I}_m$, like each $Z_{X,\lambda}$, is an analytic set and hence a countable union of smooth submanifolds; see Amir et al. (2023, Appendix A). For each fixed (X, λ), the parameter tuples satisfying the incidence conditions form exactly the Cartesian product $Z_{X,\lambda}^m$, whose Hausdorff dimension is at most $m(p-1)$ (Amir et al., 2023, Lemma A.17). Since (X, λ) ranges over $\mathcal{X} \times \mathbb{S}^{r-1}$, of dimension $q + r - 1$, the dimension bound for countable unions of smooth submanifolds (Amir et al., 2023, Lemma A.18) gives

$$\dim_{\mathrm{H}} \mathcal{I}_m \leq mp - m + q + r - 1.$$

A bank fails to span at some X if and only if a unit covector λ annihilates all its values there. Hence $\mathcal{B}_m$ is exactly the projection of $\mathcal{I}_m$ onto Θ^m. Coordinate projections are Lipschitz and do not increase Hausdorff dimension, which proves (27). For $m \geq q + r$, this dimension is strictly smaller than mp, so $\mathcal{B}_m$ has Lebesgue measure zero. Moreover, $\mathcal{I}_m$ is a countable union of compact sets, as is its projection, so $\mathcal{B}_m$ is Borel. Any jointly absolutely continuous law assigns probability zero to this exceptional set. □

Proof of Lemma I.1. Every rank-one single-head field belongs to the architecture family. Indeed, for $a, b, u, v \in \mathbb{R}^d$, choose a head of width r, let $e_1 \in \mathbb{R}^r$ be its first coordinate vector, and set

$$Q = \sqrt{r}\, e_1 b^\top, \qquad K = e_1 a^\top, \qquad U = e_1 v^\top, \qquad O = u e_1^\top. \tag{28}$$

Set the output projections of all other heads to zero. The effective matrices of the active head are $W = ab^\top$ and $V = uv^\top$. These parameter choices are witnesses inside Θ; no such restriction is imposed on the random draw. Therefore the rank-one pointwise spanning result (23) implies

$$\operatorname{span}\{F_\theta(X) : \theta \in \Theta\} = \mathbb{R}^D \qquad \text{for every } X \in \mathcal{A}_{N,n,d}.$$

The map $(\theta, X) \mapsto F_\theta(X)$ is jointly real analytic: the projection factorizations are polynomial, and the attention formula involves only finite sums, products, exponentials, and division by strictly positive softmax denominators. Thus the analytic regularity assumption of Theorem I.2 is satisfied.

The configuration space $\mathcal{A}_{N,n,d}$ is nonempty and open in $\mathbb{R}^D$, so it has dimension D. Apply Theorem I.2 with Θ as in (24), $\mathcal{X} = \mathcal{A}_{N,n,d}$, $G(\theta, X) = F_\theta(X)$, and

$$p = 4d \sum_{h=1}^{H} r_h, \qquad q = D, \qquad r = D, \qquad m = 2D.$$

The exceptional banks form a subset of Θ^{2D} of Hausdorff dimension at most $2Dp - 1$, hence of Lebesgue measure zero. Every jointly absolutely continuous draw, including independent Gaussian projection entries with prescribed positive variances, avoids this set almost surely and satisfies (25). The Lie algebra generated by the sampled fields contains their real linear span. □

J Proof of Theorem 3.3

Theorem 3.3 (Universal interpolation with Gaussian-initialized attention blocks)**.** *Let $d \geq 3$, $n \geq 2$, and $N \geq 1$, and put $D := Nnd$. Fix an H-head architecture with positive widths $r_1, \ldots, r_H$. Initialize $m_N = 2D$ noncausal attention blocks by drawing every entry of all their query, key, value, and output projection matrices independently from $\mathcal{N}(0, 1)$. The effective matrices are defined by the factorizations in Example 2.2. With probability one over this single draw, these frozen blocks yield universal interpolation at continuous depth on $\mathcal{B}_{N,n,d}$ in the sense of Definition 2.4, simultaneously for all input and target configurations in this domain. Only the block indices, signs, and durations depend on the prescribed interpolation problem; the terminal depth $S > 0$ may depend on the endpoints.*

Throughout this section, put $D = Nnd$, $\mathcal{A} = \mathcal{A}_{N,n,d}$, $\mathcal{B} = \mathcal{B}_{N,n,d}$, and use the full projection space Θ from (24). For $x, y \in \mathbb{R}^d$, we identify their exterior product $x \wedge y$ with the vector $(x_i y_j - x_j y_i)_{i<j} \in \mathbb{R}^{\binom{d}{2}}$. It vanishes exactly when x and y are linearly dependent. The set $\mathcal{B}$ is nonempty and open: its defining conditions are open, and $\mathcal{A} \subseteq \mathcal{B}$. For $d \geq 3$, $\mathcal{A}$ is connected by Section H and dense in $\mathcal{B}$, so $\mathcal{B}$ is connected as well.

Lemma J.1 (Contextual separation)**.** *Let $d \geq 2$, $n \geq 2$, and $N \geq 1$. For every $X \in \mathcal{B}$, there exists $(a, b, u, v) \in (\mathbb{R}^d)^4$ such that, setting $W = ab^\top$ and $V = uv^\top$, one has*

$$f_{W,V}(x^\ell)_i \wedge x_j^m + x_i^\ell \wedge f_{W,V}(x^m)_j \neq 0 \tag{29}$$

for every pair $(\ell, i) \neq (m, j)$ with $x_i^\ell \wedge x_j^m = 0$. Consequently, one rank-one residual attention layer with sufficiently small nonzero scalar gain sends X into $\mathcal{A}$.

Proof. Fix $X \in \mathcal{B}$. For each sequence, define its normalized exponential sum and softmax barycenter by

$$M_\ell(\eta) := \frac{1}{n} \sum_{j=1}^{n} e^{\langle \eta, x_j^\ell \rangle}, \qquad z_\ell(\eta) := \nabla \log M_\ell(\eta) = \frac{\sum_j e^{\langle \eta, x_j^\ell \rangle} x_j^\ell}{\sum_j e^{\langle \eta, x_j^\ell \rangle}}.$$

The token update is $f_{W,V}(x^\ell)_i = V z_\ell(W x_i^\ell)$.

Step 1: separate one initially collinear pair. Consider a collinear pair $q = x_i^\ell$ and $x_j^m = cq$. Nonzero tokens imply $c \neq 0$, and within-sequence noncollinearity implies $\ell \neq m$. We claim that

$$\delta(\eta) := z_m(c\eta) - c z_\ell(\eta) \quad \text{is not identically zero.} \tag{30}$$

Otherwise, $z_m(c\eta) = cz_\ell(\eta)$ for all $\eta \in \mathbb{R}^d$. Since $z_k = \nabla \log M_k$, the chain rule gives

$$\nabla_\eta \big[\log M_m(c\eta) - c^2 \log M_\ell(\eta)\big] = cz_m(c\eta) - c^2 z_\ell(\eta) = 0.$$

Thus the difference is constant on $\mathbb{R}^d$. Evaluating at $\eta = 0$ and using $M_m(0) = M_\ell(0) = 1$ shows that this constant is zero, hence

$$\log M_m(c\eta) = c^2 \log M_\ell(\eta) \qquad \text{for all } \eta \in \mathbb{R}^d.$$

Choose $h \in \mathbb{R}^d$ such that

$$\langle h, x_i^k \rangle \neq \langle h, x_j^k \rangle \qquad (k \in \{\ell, m\},\ i \neq j).$$

Such an h exists: each forbidden equality defines the proper hyperplane $(x_i^k - x_j^k)^\perp$, and only finitely many hyperplanes are excluded. Since $c \neq 0$, each exponential sum has a unique dominant term along $\eta = th$ as $t \to +\infty$, which gives

$$\log M_m(cth) = tA - \log n + o(1), \qquad \log M_\ell(th) = tB - \log n + o(1),$$

where $A = \max_j c\langle h, x_j^m \rangle$ and $B = \max_i \langle h, x_i^\ell \rangle$. This also holds when $c < 0$. Dividing the identity by t first yields $A = c^2 B$; comparing the remaining constants yields $(c^2 - 1) \log n = 0$. Since $n \geq 2$, $c = \pm 1$, and the identity reduces to $M_m(c\eta) = M_\ell(\eta)$. Let $y_1, \ldots, y_R$ be the distinct vectors in $\{cx_j^m\}_j \cup \{x_i^\ell\}_i$, and let a_r be the multiplicity of y_r in the first multiset minus its multiplicity in the second. Multiplying the identity by n and grouping equal vectors gives

$$\sum_{r=1}^{R} a_r e^{\langle \eta, y_r \rangle} = 0 \qquad \text{for all } \eta \in \mathbb{R}^d.$$

Choose $g \in \mathbb{R}^d$ outside the finitely many proper hyperplanes $(y_r - y_s)^\perp$, $r \neq s$, and relabel so that $\lambda_1 < \cdots < \lambda_R$, where $\lambda_r = \langle g, y_r \rangle$. Taking $\eta = tg$ and dividing by $e^{t\lambda_R}$ yields

$$a_R + \sum_{r<R} a_r e^{-t(\lambda_R - \lambda_r)} = 0.$$

Letting $t \to +\infty$ gives $a_R = 0$; repeating with the remaining exponents gives $a_r = 0$ for every r. Thus the two multisets agree, so there is a permutation π such that $cx_j^m = x_{\pi(j)}^\ell$ for all j. Since $c = \pm 1$, the two sequences are related by a signed permutation, contradicting $X \in \mathcal{B}$. This proves the claim.

Choose η with $\delta(\eta) \neq 0$ and $u \notin \mathbb{R}q$, and set

$$W = \frac{\eta q^\top}{\|q\|^2}, \qquad V = \frac{u\delta(\eta)^\top}{\|\delta(\eta)\|^2}.$$

These matrices have rank at most one, $Wq = \eta$, and

$$f_{W,V}(x^m)_j - c f_{W,V}(x^\ell)_i = V\delta(\eta) = u.$$

By bilinearity and antisymmetry of the exterior product, the left-hand side of (29) for this pair is

$$\begin{aligned} f_{W,V}(x^\ell)_i \wedge x_j^m + x_i^\ell \wedge f_{W,V}(x^m)_j &= f_{W,V}(x^\ell)_i \wedge (cq) + q \wedge f_{W,V}(x^m)_j \\ &= -c\, q \wedge f_{W,V}(x^\ell)_i + q \wedge f_{W,V}(x^m)_j \\ &= q \wedge \big(f_{W,V}(x^m)_j - c f_{W,V}(x^\ell)_i\big) \\ &= q \wedge u \neq 0, \end{aligned}$$

where the last inequality follows from $u \notin \mathbb{R}q$.

Step 2: choose common parameters for all initially collinear pairs. For each initially collinear pair, the squared norm of the left-hand side of (29) is a real analytic function of $(a, b, u, v) \in (\mathbb{R}^d)^4$, with $W = ab^\top$ and $V = uv^\top$. The preceding construction shows that it is not identically zero. Its zero set is therefore closed with empty interior, so its complement is open and dense. There are only finitely many initially collinear pairs; hence the intersection of these open dense complements is itself open and dense, and in particular nonempty. Any (a, b, u, v) in this intersection satisfies all the required inequalities.

Step 3: apply one residual layer to reach $\mathcal{A}$. Fix such a common choice of a, b, u, v. All token updates lie in the image of $V = uv^\top$, which has dimension at most one. Hence, for every pair,

$$\begin{aligned}&(x_i^\ell + t f_{W,V}(x^\ell)_i) \wedge (x_j^m + t f_{W,V}(x^m)_j)\\ &\quad = x_i^\ell \wedge x_j^m + t\big(f_{W,V}(x^\ell)_i \wedge x_j^m + x_i^\ell \wedge f_{W,V}(x^m)_j\big).\end{aligned}$$

Every initially collinear pair separates for all $t \neq 0$. Previously noncollinear pairs remain noncollinear, and nonzero tokens remain nonzero, for all sufficiently small $|t|$. There are finitely many tokens and pairs, giving one common neighborhood of $t = 0$. The resulting batch therefore belongs to $\mathcal{A}$ for every sufficiently small nonzero t. □

For $\theta \in \Theta$, write $f_\theta = f_{\mathbf{W},\mathbf{V}}$ for the field on one sequence and $f_\theta^{(N)}$ for its corresponding batch field.

Lemma J.2 (A uniform separating dictionary)**.** *For the fixed attention architecture above, let $m \geq D + 1$ and draw $(\theta_1, \dots, \theta_m)$ from a jointly absolutely continuous law on Θ^m. With probability one, simultaneously for every $X \in \mathcal{B}$, there is $k \in \{1, \dots, m\}$ such that*

$$f_{\theta_k}(x^\ell)_i \wedge x_j^{\ell'} + x_i^\ell \wedge f_{\theta_k}(x^{\ell'})_j \neq 0$$

for every pair $(\ell, i) \neq (\ell', j)$ with $x_i^\ell \wedge x_j^{\ell'} = 0$. For this same k and all sufficiently small nonzero t, both

$$X + t f_{\theta_k}^{(N)}(X) \in \mathcal{A}, \qquad \Phi^t_{f_{\theta_k}^{(N)}}(X) \in \mathcal{A}.$$

Both curves remain in $\mathcal{B}$ for sufficiently small $|t|$.

Proof. Step 1: establish uniform separation by the sampled bank. Define the scalar selector

$$\mathcal{S}(\theta, X) = \prod_{(\ell,i)<(\ell',j)} \begin{pmatrix} \|x_i^\ell \wedge x_j^{\ell'}\|^2 \\ + \|f_\theta(x^\ell)_i \wedge x_j^{\ell'} + x_i^\ell \wedge f_\theta(x^{\ell'})_j\|^2 \end{pmatrix}. \tag{31}$$

Here $<$ denotes lexicographic order on the token indices, so each pair of distinct token positions occurs once. The function $\mathcal{S}$ is nonnegative and jointly real analytic on $\Theta \times \mathcal{B}$. By Lemma J.1 and the embedding (28), for every $X \in \mathcal{B}$ there is $\theta \in \Theta$ with $\mathcal{S}(\theta, X) > 0$. For a scalar-valued map, this is exactly the spanning assumption of Theorem I.2.

Since $\Theta \simeq \mathbb{R}^p$ and $\mathcal{B} \subseteq \mathbb{R}^D$ is nonempty and open, we can apply that theorem with $\mathcal{X} = \mathcal{B}$, $G = \mathcal{S}$, $q = D$, and $r = 1$. Thus, for $m \geq D + 1$, a jointly absolutely continuous draw $(\theta_1, \dots, \theta_m) \in \Theta^m$ almost surely satisfies

$$\forall X \in \mathcal{B}, \quad \exists k \in \{1, \dots, m\} : \quad \mathcal{S}(\theta_k, X) > 0.$$

By definition of $\mathcal{S}$, this is precisely the required transversality for all initially collinear pairs at X.

Step 2: reach $\mathcal{A}$ with a residual layer. Fix such a bank and configuration, and choose k with $\mathcal{S}(\theta_k, X) > 0$. For an initially collinear pair $x_i^\ell, x_j^{\ell'}$, the wedge after a residual step is

$$\begin{aligned}&(x_i^\ell + t f_{\theta_k}(x^\ell)_i) \wedge (x_j^{\ell'} + t f_{\theta_k}(x^{\ell'})_j)\\ &\quad = t\big(f_{\theta_k}(x^\ell)_i \wedge x_j^{\ell'} + x_i^\ell \wedge f_{\theta_k}(x^{\ell'})_j\big) + t^2 f_{\theta_k}(x^\ell)_i \wedge f_{\theta_k}(x^{\ell'})_j.\end{aligned}$$

The linear coefficient is nonzero, so the pair separates for all sufficiently small nonzero t. Previously noncollinear pairs and nonzero tokens persist by continuity.

Step 3: reach $\mathcal{A}$ along the flow. Keep the same k and write $V_{k,h} = O_{k,h} U_{k,h}$. For $Z = \mathrm{vec}(z^1, \dots, z^N)$, each softmax barycenter is a convex combination of the tokens in its sequence. Hence, for every ℓ, i,

$$\|f_{\theta_k}(z^\ell)_i\| \leq \left(\sum_{h=1}^H \|V_{k,h}\|_{\mathrm{op}}\right) \|z^\ell\|.$$

Summing the squared norms over all tokens and sequences gives

$$\|f_{\theta_k}^{(N)}(Z)\|^2 = \sum_{\ell=1}^N \sum_{i=1}^n \|f_{\theta_k}(z^\ell)_i\|^2 \leq n \left(\sum_{h=1}^H \|V_{k,h}\|_{\mathrm{op}}\right)^2 \sum_{\ell=1}^N \|z^\ell\|^2.$$

Taking square roots yields

$$\|f_{\theta_k}^{(N)}(Z)\| \le \sqrt{n}\left(\sum_{h=1}^{H}\|V_{k,h}\|_{\mathrm{op}}\right)\|Z\| \qquad \text{on } \mathbb{R}^D.$$

Analyticity gives local existence and uniqueness. By Gronwall's inequality, this linear growth bound rules out finite-time blow-up for both the field and its negative, so the flow is defined for all $t \in \mathbb{R}$.

Write $Z(t) = \Phi^t_{f^{(N)}_{\theta_k}}(X)$ and denote its tokens by $z_i^\ell(t)$. Then $z_i^\ell(0) = x_i^\ell$ and $\dot z_i^\ell(0) = f_{\theta_k}(x^\ell)_i$. For every initially collinear pair, bilinearity gives

$$\frac{d}{dt}\big(z_i^\ell(t) \wedge z_j^{\ell'}(t)\big)\Big|_{t=0} = f_{\theta_k}(x^\ell)_i \wedge x_j^{\ell'} + x_i^\ell \wedge f_{\theta_k}(x^{\ell'})_j \neq 0.$$

Since $x_i^\ell \wedge x_j^{\ell'} = 0$, Taylor's formula yields

$$\begin{aligned} z_i^\ell(t) \wedge z_j^{\ell'}(t) = t\, f_{\theta_k}(x^\ell)_i \wedge x_j^{\ell'} \\ + t\, x_i^\ell \wedge f_{\theta_k}(x^{\ell'})_j + O(t^2). \end{aligned}$$

The coefficient of t is nonzero, so this product is nonzero for all sufficiently small nonzero t. Initially noncollinear pairs and nonzero tokens persist by continuity. There are finitely many tokens and pairs, so one common interval works for all of them. Finally, both the residual curve and the flow start at $X \in \mathcal{B}$; by continuity and openness of $\mathcal{B}$, they remain there for sufficiently small $|t|$. □

Proof of Theorem 3.3. Let $(\theta_1, \dots, \theta_{2D}) \in \Theta^{2D}$ be the sampled parameters, with batch fields $f_{\theta_j}^{(N)}$, $1 \le j \le 2D$. Independent Gaussian projection entries have a joint density on Θ^{2D}. By Lemma I.1, with probability one these fields span $\mathbb{R}^D$ simultaneously at every $X \in \mathcal{A}$. For separation, recall that the proof of Lemma J.2 applies Theorem I.2 to the scalar selector $\mathcal{S} : \Theta \times \mathcal{B} \to \mathbb{R}$ from (31), with $q = D$ and $r = 1$. Since our bank has $2D \ge q + r = D + 1$ fields, it almost surely satisfies

$$\forall X \in \mathcal{B}, \quad \exists k \in \{1, \dots, 2D\} : \quad \mathcal{S}(\theta_k, X) > 0.$$

Fix a bank in the intersection of these two probability-one events. Both properties then hold uniformly in the configuration for this same bank.

We now prove that the sampled fields have full Lie rank on $\mathcal{B}$. Fix $X \in \mathcal{B}$ and choose $k \in \{1, \dots, 2D\}$ with $\mathcal{S}(\theta_k, X) > 0$, as in Lemma J.2. Write $\Phi_t = \Phi^t_{f^{(N)}_{\theta_k}}$ for its flow and put

$$L_X = \mathrm{Lie}\big(f_{\theta_1}^{(N)}, \dots, f_{\theta_{2D}}^{(N)}\big)(X) \subseteq \mathbb{R}^D.$$

For an analytic vector field G, its pullback by Φ_t is the vector field defined by

$$(\Phi_t^* G)(Z) \coloneqq \big(D\Phi_t(Z)\big)^{-1} G(\Phi_t(Z)).$$

The Jacobian $D\Phi_t(Z)$ is invertible since Φ_{-t} is the inverse of Φ_t. For each $j \in \{1, \dots, 2D\}$, the map

$$t \longmapsto U_j(t) = \big(\Phi_t^* f_{\theta_j}^{(N)}\big)(X) \in \mathbb{R}^D$$

is real analytic near $t = 0$. The standard differentiation formula for pullbacks (Agrachev & Sachkov, 2004, Section 2.5, Eq. (2.21)) and its iterates give

$$\begin{aligned} \frac{d}{dt}\Phi_t^* G &= \Phi_t^*\big[f_{\theta_k}^{(N)}, G\big], \\ \frac{d^r}{dt^r}\big(\Phi_t^* G\big) &= \Phi_t^*\big(\mathrm{ad}^r_{f^{(N)}_{\theta_k}} G\big), \qquad r \ge 0, \end{aligned}$$

where $\mathrm{ad}_{f^{(N)}_{\theta_k}} G = [f_{\theta_k}^{(N)}, G]$. Taking $G = f_{\theta_j}^{(N)}$ and evaluating at $t = 0$, where Φ_0^* is the identity, gives

$$U_j^{(r)}(0) = \big(\mathrm{ad}^r_{f^{(N)}_{\theta_k}} f_{\theta_j}^{(N)}\big)(X) \in L_X, \qquad r \ge 0.$$

Every partial sum of the Taylor series of U_j belongs to L_X. Since this series converges to $U_j(t)$ for sufficiently small $|t|$ and L_X is closed, we obtain $U_j(t) \in L_X$ for every j and all sufficiently small $|t|$.

Choose such a nonzero t with $X' = \Phi_t(X) \in \mathcal{A}$, as guaranteed by Lemma J.2. Uniform direct spanning gives

$$\operatorname{span}\{f_{\theta_j}^{(N)}(X') : 1 \le j \le 2D\} = \mathbb{R}^D.$$

Since $D\Phi_t(X)$ is invertible, their pullbacks $U_j(t)$ also span $\mathbb{R}^D$. Since they belong to L_X,

$$\mathbb{R}^D = \operatorname{span}\{U_j(t) : 1 \le j \le 2D\} \subseteq L_X \subseteq \mathbb{R}^D.$$

Thus $L_X = \mathbb{R}^D$. This holds at every $X \in \mathcal{B}$ for the same bank of full multi-head blocks. The same proof applies to independent Gaussian entries with any prescribed positive variances, or more generally to a jointly absolutely continuous law on Θ^{2D}.

The sampled fields are analytic, and $\mathcal{B}$ is nonempty, open, and connected for $d \ge 3$ by the domain argument above. Moreover, $-f_{\theta_j}$ belongs to the same architecture family: it is realized by negating all output-projection blocks of θ_j. The draw is fixed before the endpoints, so Corollary 2.7 proves universal interpolation on $\mathcal{B}_{N,n,d}$ with the same $2D$ sampled generators, used in either sign over a horizon $[0, S]$ that may depend on the endpoints. Since N was arbitrary, this holds for every N. □

Remark J.3 (Rank-one attention). Theorem 3.3 includes the case of one head of width one. Taking $H = 1$ and $r_1 = 1$, write the Gaussian projections as $Q_k = b_k^\top$, $K_k = a_k^\top$, $U_k = v_k^\top$, and $O_k = u_k$. Then

$$W_k = a_k b_k^\top, \qquad V_k = u_k v_k^\top, \qquad 1 \le k \le 2D,$$

where all vectors a_k, b_k, u_k, v_k are independent standard Gaussian vectors in $\mathbb{R}^d$. Both effective matrices have rank one almost surely. Thus the same theorem gives universal interpolation at continuous depth on $\mathcal{B}_{N,n,d}$ with a frozen bank of $2Nnd$ such fields.

Dependence of the terminal depth on the endpoints. Allowing the terminal time to vary is essential for this formulation. Each sampled field, or its negative, satisfies $\|G(Z)\| \le C\|Z\|$ with a common $C < \infty$, since softmax takes convex combinations of token values. Gronwall's inequality therefore bounds every trajectory using these unscaled fields by

$$\|Z(S)\| \le e^{CS}\|Z(0)\|.$$

Yet $\lambda X \in \mathcal{B}_{N,n,d}$ for every $X \in \mathcal{B}_{N,n,d}$ and $\lambda > 0$. For any fixed horizon S_0, choosing $\lambda > e^{CS_0}$ gives an admissible target $Y = \lambda X$ that cannot be reached by time S_0. A horizon depending on the endpoints removes this obstruction without introducing adjustable scalar gains.

K Proof of Corollary 3.4

Proof. Fix $N \ge 1$, put $D = Nnd$, and choose a draw $(\theta_1, \dots, \theta_{2D}) \in \Theta^{2D}$ in the intersection of the probability-one events in Lemmas I.1 and J.2. These events hold simultaneously for all configurations in their respective domains, so the bank is fixed before the endpoints are chosen. The batch fields $f_{\theta_1}^{(N)}, \dots, f_{\theta_{2D}}^{(N)}$ are analytic and span $\mathbb{R}^D$ at every point of $\mathcal{A}_{N,n,d}$ by Lemma I.1. This set is open and connected for $d \ge 3$, and the attention architecture allows real scalar multiplication since $cf_{\mathbf{W},\mathbf{V}} = f_{\mathbf{W},c\mathbf{V}}$. Hence Proposition F.1 gives universal interpolation at finite depth on $\mathcal{A}_{N,n,d}$ using this same bank, with all intermediate configurations in $\mathcal{A}_{N,n,d}$.

Now fix arbitrary endpoints $X, Y \in \mathcal{B}_{N,n,d}$. By Lemma J.2, choose an index $k_X \in \{1, \dots, 2D\}$ and set $F \coloneqq f_{\theta_{k_X}}^{(N)}$. For sufficiently small $s > 0$,

$$X^+ \coloneqq X + sF(X) \in \mathcal{A}_{N,n,d}.$$

This is one admissible residual layer.

At the target, choose an index $k_Y \in \{1, \dots, 2D\}$ supplied by the same lemma and set $G \coloneqq f_{\theta_{k_Y}}^{(N)}$. For every pair $(\ell, i) \ne (\ell', j)$ with $y_i^\ell \wedge y_j^{\ell'} = 0$, this choice gives

$$f_{\theta_{k_Y}}(y^\ell)_i \wedge y_j^{\ell'} + y_i^\ell \wedge f_{\theta_{k_Y}}(y^{\ell'})_j \ne 0.$$

Consider the equation

$$Z + tG(Z) - Y = 0.$$

Its derivative with respect to Z at $(t, Z) = (0, Y)$ is the identity. The implicit function theorem therefore gives an analytic curve $Z(t)$ for t near zero satisfying

$$Z(0) = Y, \qquad Z(t) + tG(Z(t)) = Y,$$
$$Z'(0) = -G(Y).$$

Taylor's formula therefore gives

$$Z(t) = Z(0) + tZ'(0) + O(t^2) = Y - tG(Y) + O(t^2).$$

Write $z_i^\ell(t)$ for the tokens of $Z(t)$. For each initially collinear pair, its wedge along this curve is

$$z_i^\ell(t) \wedge z_j^{\ell'}(t) = -t\, f_{\theta_{k_Y}}(y^\ell)_i \wedge y_j^{\ell'} - t\, y_i^\ell \wedge f_{\theta_{k_Y}}(y^{\ell'})_j + O(t^2).$$

It is consequently nonzero for all sufficiently small nonzero t. Initially noncollinear pairs remain noncollinear by continuity, and all tokens remain nonzero. There are finitely many tokens and pairs, so we may choose one sufficiently small $t > 0$ such that

$$Y^- \coloneqq Z(t) \in \mathcal{A}_{N,n,d}, \qquad Y^- + tG(Y^-) = Y.$$

Apply the finite-depth interpolation result on $\mathcal{A}_{N,n,d}$ to connect X^+ to Y^-. Prepend the layer $\mathrm{Id} + sF$ and append the layer $\mathrm{Id} + tG$. This gives the required finite residual network. All configurations other than the prescribed endpoints belong to $\mathcal{A}_{N,n,d}$, and hence every configuration belongs to $\mathcal{B}_{N,n,d}$. The construction uses at most two additional layers around an interpolation inside $\mathcal{A}_{N,n,d}$. The implicit function theorem only defines the intermediate target Y^-; the last network layer is the admissible forward residual map $\mathrm{Id} + tG$.

Every selected index and scalar gain is shared across the entire batch. Finally, each multiple cf_{θ_j} is realized by a scalar gain on the frozen block output; the head widths and all stored projections remain unchanged. The same bank works for every endpoint pair. If $X = Y$, the depth-zero network also suffices. Since the sampled fields are fixed before the endpoints, this is universal interpolation at finite depth on $\mathcal{B}_{N,n,d}$ in the sense of Section 2.3. The argument applies to every $N \geq 1$. □

L Proof of Theorem 3.5

This appendix proves Theorem 3.5 for noncausal attention with unrestricted effective score and value matrices. The admissible interpolation domain here is allowed to depend on the sampled pair, whereas the Gaussian multihead bank in Theorem 3.3 works on the fixed domain $\mathcal{B}_{N,n,d}$. The same pair works for every finite batch size and sequence length. For a single sequence $x = \mathrm{vec}(x_1, \ldots, x_n)$, we use the single-head field

$$f_{W,V}(x)_i = V \frac{\sum_{j=1}^n x_j \exp(x_j^\top W x_i)}{\sum_{j=1}^n \exp(x_j^\top W x_i)}, \qquad W, V \in \mathbb{R}^{d\times d}, \tag{32}$$

and its batch lift $f_{W,V}^{(N)}$. One head of width at least d realizes every such field; in a multihead block, the remaining heads can be set to zero. Under the standard equal-width convention $r_h = d/H$, this single-active-head realization corresponds to $H = 1$.

Write $\theta = (W_1, V_1, W_2, V_2) \in \mathbb{R}^{4d^2}$ and, for $N \geq 1$ and $n \geq 2$, put $D = Nnd$ and $f_r^{(N)} = f_{W_r,V_r}^{(N)}$. We choose the interpolation domain as

$$\mathcal{R}_{\theta,N,n} \coloneqq \left\{ X \in \mathcal{B}_{N,n,d} : \mathrm{Lie}(f_1^{(N)}, f_2^{(N)})(X) = \mathbb{R}^D \right\}, \tag{33}$$
$$\Sigma_{\theta,N,n} \coloneqq \mathcal{B}_{N,n,d} \setminus \mathcal{R}_{\theta,N,n}.$$

Thus $\mathcal{R}_{N,n} = \mathcal{R}_{\theta,N,n}$ in the main statement. We prove the stronger bound

$$\dim_{\mathbb{R}} \Sigma_{\theta,N,n} \leq Nnd - 2, \tag{34}$$

almost surely, simultaneously for every $N \geq 1$ and $n \geq 2$. In particular, the domain also has full Lebesgue measure in $\mathcal{B}_{N,n,d}$ and is path connected. More generally, the same conclusions hold when θ is drawn from any joint law absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^{4d^2}$. Both value matrices V_1, V_2 are then invertible almost surely.

The Gaussian projection initialization in Theorem 3.5 induces such a law. Indeed, condition on the key and value projections, which almost surely have full column rank when $r \geq d$. The maps from the query and output projections to the effective matrices are then surjective linear maps. Since all projection entries are independent standard Gaussians, the conditional joint law of θ is a nondegenerate Gaussian, and its unconditional joint law is therefore absolutely continuous. The arguments below only exclude Lebesgue-null parameter sets.

The two frozen base fields, used with either sign on a freely chosen interval $[0, S]$, give universal interpolation in the sense of Definition 2.4 on $\mathcal{R}_{\theta,N,n}$, with $m_N = 2$. Their negatives are fixed by the same draw and require no additional sampled matrices. The matrices are independent of N and n, while the domain may depend on θ, N, n. The theorem does not assert full Lie rank on all of $\mathcal{B}_{N,n,d}$, nor a width-one realization.

L.1 A finite Lie-minor witness for every batch size

Lemma L.1 (A two-field witness for arbitrary batch size)**.** *For every $d \geq 3$, $n \geq 2$, and $N \geq 1$, there exist invertible $W_1, V_1, W_2, V_2 \in \mathbb{R}^{d\times d}$ and $X \in \mathcal{A}_{N,n,d}$ such that, with $D = Nnd$,*

$$\mathrm{Lie}\big(f^{(N)}_{W_1,V_1}, f^{(N)}_{W_2,V_2}\big)(X) = \mathbb{R}^D.$$

Proof. Put $M = Nn$, $\eta = (4M)^{-1}$, and choose

$$s_{\ell i} = \frac{1}{4} + ((\ell-1)n + i)\eta, \qquad x^\ell_i = s_{\ell i} e_1 + \sqrt{1 - s^2_{\ell i}}\, e_2. \tag{35}$$

All M tokens are distinct points in the open first quadrant of the unit circle, so this batch belongs to $\mathcal{A}_{N,n,d}$. For each token define the real-analytic function

$$h_{\ell i}(\beta) = \frac{\sum_{j=1}^n s_{\ell j} e^{\beta s_{\ell i} s_{\ell j}}}{\sum_{j=1}^n e^{\beta s_{\ell i} s_{\ell j}}}.$$

The largest and second-largest scalars in sequence ℓ differ by η (indeed $s_{\ell n} - s_{\ell,n-1} = \big[n - (n-1)\big]\eta = \eta$). Dividing the numerator and denominator by $e^{\beta s_{\ell i} s_{\ell n}}$ gives

$$h_{\ell i}(\beta) = s_{\ell n} - \eta e^{-\eta s_{\ell i}\beta} + o\big(e^{-\eta s_{\ell i}\beta}\big), \qquad \beta \longrightarrow +\infty. \tag{36}$$

The functions $1, h_{\ell i}$ are linearly independent: indeed suppose that

$$a_0 + \sum_{\ell,i} a_{\ell i} h_{\ell i}(\beta) = 0.$$

Subtracting its limit as $\beta \to +\infty$ gives

$$\sum_{\ell,i} a_{\ell i}\big(h_{\ell i}(\beta) - s_{\ell n}\big) = 0.$$

By the asymptotic expansion,

$$h_{\ell i}(\beta) - s_{\ell n} = (-\eta + o(1))e^{-\eta s_{\ell i}\beta}.$$

If any coefficient $a_{\ell i}$ is nonzero, choose the unique index (as the rates $\eta s_{\ell i}$ are pairwise distinct) (ℓ_*, i_*) with the smallest decay rate among those with nonzero coefficients. Dividing by $e^{-\eta s_{\ell_* i_*}\beta}$ and letting $\beta \to +\infty$, all other terms vanish, yielding

$$-\eta a_{\ell_* i_*} = 0,$$

a contradiction. Thus all $a_{\ell i}$ vanish, and the original relation then gives $a_0 = 0$, proving linear independence.

Consider the auxiliary pair consisting of the tokenwise linear field L_Λ and an attention field H, where

$$\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_d), \qquad \lambda_r = 1 + \frac{r-1}{d}, \qquad v = (1, \dots, 1)^\top, \qquad H = f^{(N)}_{e_1 e_1^\top,\, v e_1^\top}.$$

Here $(L_\Lambda(X))_i^\ell = \Lambda x_i^\ell$. The flow Φ_t of L_Λ multiplies every token by $e^{t\Lambda}$. Its pushforward of H, evaluated at the batch (35), has spatial component r

$$\big((\Phi_t)_* H\big)_{\ell i, r}(X) = e^{(\lambda_r - 1)t} h_{\ell i}(e^{-2t}). \tag{37}$$

We claim that the evaluations of $\mathrm{ad}^k_{L_\Lambda} H$, $k \geq 0$, span $\mathbb{R}^D$. If a covector $(c_{\ell i, r})$ annihilated them all, the real-analytic function obtained by pairing it with (37) would have every derivative zero at $t = 0$, and hence vanish for all real t. Indeed, these derivatives are the iterated brackets, up to signs. Set

$$B_r(\beta) = \sum_{\ell, i} c_{\ell i, r} h_{\ell i}(\beta), \qquad z = e^{-t/d}.$$

The resulting identity is

$$\sum_{r=1}^{d} z^{-(r-1)} B_r(z^{2d}) = 0, \qquad z > 0. \tag{38}$$

Each B_r is analytic near zero. After multiplication by z^{d-1}, its contribution to (38) contains only powers $d - r + 2dk$, $k \geq 0$. These sets of exponents are disjoint for different r. Every Taylor coefficient of every B_r therefore vanishes. Analyticity and the independence of the $h_{\ell i}$ imply $c_{\ell i, r} = 0$ for all indices, proving the claim.

Finite dimensionality now provides nonnegative integers $k_1, \dots, k_D$ for which

$$\det\big((\mathrm{ad}^{k_j}_{L_\Lambda} H)(X)\big)_{j=1}^{D} \neq 0. \tag{39}$$

To replace the linear field by attention, put $G_T = f^{(N)}_{TI, \Lambda}$. At the fixed batch all tokens have norm one and, within each sequence, $\langle x_i^\ell, x_j^\ell \rangle < 1$ when $i \neq j$. Thus, for every fixed derivative order a,

$$D^a G_T(X) \longrightarrow D^a L_\Lambda(X) \qquad (T \longrightarrow +\infty). \tag{40}$$

For completeness, divide each softmax row by its diagonal exponential. Every remaining exponential is $e^{-T g_{ij}^\ell}$ at the witness, with $g_{ij}^\ell = 1 - \langle x_i^\ell, x_j^\ell \rangle > 0$. Every fixed-order derivative is bounded by a polynomial in T times e^{-Tg}, where g is the smallest of these finitely many gaps. The denominators after division are bounded below by one. This proves (40), including derivatives transverse to the plane containing the witness tokens.

A fixed Lie word evaluated at a point depends polynomially on finitely many derivatives of its generators. Hence the determinant in (39) remains nonzero after replacing L_Λ by G_T, for some finite sufficiently large T. Finally replace the second field's parameters by

$$W_2 = e_1 e_1^\top + \varepsilon I, \qquad V_2 = v e_1^\top + \varepsilon I.$$

Both determinants equal $\varepsilon^{d-1}(1+\varepsilon)$, and the selected Lie determinant remains nonzero for sufficiently small $\varepsilon > 0$ by continuity. Together with $W_1 = TI$ and $V_1 = \Lambda$, this gives four invertible matrices and proves the lemma. □

Lemma L.2 (Generic rank simultaneously for all finite batches)**.** *Fix $d \geq 3$, and draw (W_1, V_1, W_2, V_2) from a joint law absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^{4d^2}$. Almost surely, simultaneously for every $n \geq 2$ and $N \geq 1$, the full Lie-rank locus in $\mathcal{B}_{N,n,d}$ is open, dense, and has full Lebesgue measure.*

Proof. For fixed n, N, Lemma L.1 provides a fixed batch X_0 and finitely many Lie words whose determinant $\Delta_{n,N}(\theta, X)$ is nonzero at one parameter choice, where $\theta = (W_1, V_1, W_2, V_2)$. The determinant is jointly real analytic in all parameters and state coordinates: the softmax denominators are strictly positive on the entire real domain. Therefore $\theta \mapsto \Delta_{n,N}(\theta, X_0)$ is a nonzero analytic function, whose zero set has Lebesgue measure zero. For almost every draw, $\Delta_{n,N}(\theta, \cdot)$ is consequently nonzero analytic. Its nonvanishing locus is open, dense, and of full measure, also after restriction to $\mathcal{B}_{N,n,d}$. The full Lie-rank locus contains it and is itself open, so it has the same three properties. Finally intersect these probability-one parameter events over the countably many pairs (n, N). The resulting event still has probability one and uses the same two generators for every pair. □

L.2 GENERIC MATRIX CONDITIONS

Write

$$S_r = \frac{W_r + W_r^\top}{2}, \qquad K_r = \frac{W_r - W_r^\top}{2}, \qquad r = 1, 2.$$

In addition to the event supplied by Lemma L.2, we impose the following conditions:

1. $\mathrm{Lie}_{\mathrm{mat}}(V_1, V_2) = \mathrm{gl}_d(\mathbb{R})$;
2. S_1, S_2 are invertible and nonproportional, and K_1, K_2 are nonzero and nonproportional;
3. for $r \neq s$,
$$V_s^\top S_r + S_r V_s \notin \mathbb{R} S_r, \qquad V_s^\top K_r + K_r V_s \notin \mathbb{R} K_r; \tag{41}$$
4. each V_r has simple complex spectrum and no two of its eigenvalues, allowing repetition, sum to zero.

These conditions hold almost surely and do not depend on N, n. For the first, the matrices

$$D_0 = \mathrm{diag}(1, 2, 4, \ldots, 2^{d-1}), \qquad R_0 = I + \mathbf{1}\mathbf{1}^\top$$

generate gl_d: their distinct nonzero adjoint eigenvalue gaps isolate every off-diagonal matrix unit from R_0; brackets give the traceless diagonals, and D_0 has nonzero trace. A finite matrix Lie determinant is therefore a nonzero polynomial in the two values. The second and fourth conditions also exclude proper algebraic sets. Here the skew-symmetric matrix space has dimension at least two because $d \geq 3$.

For fixed invertible symmetric S, the map $V \mapsto V^\top S + SV$ is onto the symmetric matrices, so the inverse image of $\mathbb{R}S$ is a proper linear subspace. For fixed nonzero skew-symmetric K, put K in real block-diagonal canonical form with $K_{12} \neq 0$. The choice $V = E_{13}$ gives $V^\top K + KV$ a nonzero $(2, 3)$ entry, whereas $K_{23} = 0$. Thus the skew condition also excludes a proper linear subspace. Fubini's theorem proves (41) almost surely. Independence of the parameter blocks is unnecessary: the joint law has a density.

Fix a parameter pair satisfying these conditions. Enumerate the $p = 2Nn^2$ score quadratics as

$$Q_{r,a,i,j}(X) = (x_j^a)^\top W_r x_i^a. \tag{42}$$

They are linearly independent over $\mathbb{C}$. Indeed, terms from different sequences or different unordered token pairs have distinct multidegrees. For a diagonal token pair, independence follows from that of S_1, S_2. For an off-diagonal pair, a relation among $W_1, W_1^\top, W_2, W_2^\top$ separates into a relation among S_1, S_2 and one among K_1, K_2, so all coefficients vanish.

Let L_{V_r} denote the shared tokenwise linear field $(L_{V_r} X)_i^a = V_r x_i^a$. The fourth condition implies that L_{V_r}, acting by differentiation, is invertible on the space of homogeneous quadratic polynomials in all D state coordinates. In an eigenbasis, its eigenvalues on that space are precisely sums of two spatial eigenvalues, with repetitions from the token indices.

L.3 EXCLUSION OF INVARIANT ALGEBRAIC HYPERSURFACES

We first record the independence property needed for polynomial exponents on a regular real hypersurface patch.

Lemma L.3 (Exponential independence on a hypersurface patch)**.** *Let $h \in \mathbb{R}[x_1, \ldots, x_q]$, with $q \geq 2$, and let $\mathcal{M}$ be a nonempty connected relatively open subset of*

$$\{x \in \mathbb{R}^q : h(x) = 0,\ \nabla h(x) \neq 0\}.$$

Let $P_1, \ldots, P_s$ and $A_1, \ldots, A_s$ be real polynomials in these coordinates. Suppose that $P_i - P_j$ is nonconstant on $\mathcal{M}$ whenever $i \neq j$. If

$$\sum_{j=1}^{s} A_j(x) e^{P_j(x)} = 0 \qquad \textit{for every } x \in \mathcal{M},$$

then $A_j(x) = 0$ for every $x \in \mathcal{M}$ and every j. The conclusion concerns the restrictions of the coefficients to $\mathcal{M}$, not their values outside this patch.

Proof. We first explain a fact about functions of one real variable, and then use the implicit function theorem to restrict the proposed identity to such functions.

Algebraic functions of one variable. A real analytic function $b(t)$ on an interval is called algebraic if it satisfies

$$Q(t, b(t)) = 0$$

for a nonzero polynomial Q in two variables. For example, $b(t) = \sqrt{1+t^2}$ satisfies $b(t)^2 - 1 - t^2 = 0$. We use the elementary elimination properties of polynomial equations: sums, products, and quotients of algebraic functions are algebraic wherever defined; moreover, any two algebraic functions $b(t), c(t)$ satisfy a nonzero polynomial relation $R(b(t), c(t)) = 0$. In the latter property the parameter t is eliminated from their polynomial equations. Derivatives are algebraic as well. Indeed, choose an equation $Q(t, b(t)) = 0$ of smallest possible degree in its second variable. Its partial derivative $Q_z(t, b(t))$ cannot vanish identically, since that would give an equation of smaller degree. Implicit differentiation therefore gives, after restricting to a smaller interval if needed,

$$b'(t) = -\frac{Q_t(t, b(t))}{Q_z(t, b(t))}.$$

An exponential of a nonconstant algebraic function is not algebraic. Suppose that both b and e^b were algebraic, with b nonconstant. The preceding elimination property would give $R(b(t), e^{b(t)}) = 0$ for some nonzero polynomial R. The range of b contains a nonempty open interval, so $R(u, e^u) = 0$ on that interval. Since this expression is real analytic on all of $\mathbb{R}$, it vanishes for every real u. Writing $R(u, v) = \sum_{k=0}^{m} p_k(u) v^k$ with $p_m \neq 0$, and dividing by e^{mu}, gives

$$p_m(u) + \sum_{k<m} p_k(u) e^{-(m-k)u} = 0.$$

As $u \to +\infty$, every term in the sum tends to zero. This would force the nonzero polynomial $p_m(u)$ to tend to zero, which is impossible.

Independence in one variable. Let $f_1(t), \ldots, f_s(t)$ be algebraic real analytic functions on an open interval I, with pairwise nonconstant differences. Suppose that their exponentials satisfy a nontrivial relation with algebraic analytic coefficients. A coefficient is counted as nonzero if it is not identically zero as a function, even though it may vanish at some individual points.

Keep the functions $f_1, \ldots, f_s$ fixed, but consider all possible choices of algebraic real analytic coefficients giving a nontrivial relation, on I or on a nonempty open subinterval. Among these relations, choose one with the smallest number m of nonzero coefficients. Such a minimum exists because this number is an integer between 1 and s. Discard the identically zero terms and relabel the remaining functions and coefficients, so that the chosen relation reads

$$a_1(t) e^{f_1(t)} + \cdots + a_m(t) e^{f_m(t)} = 0,$$

with every a_j not identically zero. Necessarily $m \geq 2$: a relation with only one term would force its coefficient to vanish identically, since an exponential never vanishes.

Since a_m is not identically zero, choose a point t_0 where $a_m(t_0) \neq 0$. By continuity, there is a smaller open interval J containing t_0 on which $a_m(t)$ never vanishes. The product $a_m(t) e^{f_m(t)}$ is therefore nonzero throughout J, and division gives

$$1 + \sum_{j=1}^{m-1} \frac{a_j(t)}{a_m(t)} e^{f_j(t) - f_m(t)} = 0 \qquad (t \in J).$$

Define

$$r_j(t) = \frac{a_j(t)}{a_m(t)}, \qquad b_j(t) = f_j(t) - f_m(t), \qquad j < m.$$

Each r_j is algebraic and real analytic on J, and remains not identically zero. Each b_j is algebraic and nonconstant; analyticity ensures that a nonconstant difference cannot become constant on a smaller open interval. We thus obtain

$$1 + \sum_{j<m} r_j(t) e^{b_j(t)} = 0.$$

The purpose of this division is to turn the last term into the constant 1, which disappears when we differentiate:

$$\sum_{j<m} \big(r_j' + r_j b_j'\big) e^{b_j} = 0.$$

All the new coefficients are algebraic and real analytic. If any were nonzero, multiplying by e^{f_m} would give a relation among the original exponentials with fewer than m terms. Minimality thus gives $r_j' + r_j b_j' = 0$ for every j. Consequently,

$$(r_j e^{b_j})' = 0, \qquad r_j = c_j e^{-b_j}, \quad c_j \neq 0.$$

This contradicts the preceding paragraph: $-b_j$ is a nonconstant algebraic function, whereas $e^{-b_j} = r_j/c_j$ would be algebraic.

Restriction of the hypersurface identity to one variable. Suppose, toward a contradiction, that there exists $p \in \mathcal{M}$ such that $A_\ell(p) \neq 0$. Since $\nabla h(p) \neq 0$, we may permute coordinates so that $\partial_{x_q} h(p) \neq 0$. The real analytic implicit function theorem then solves $h(x) = 0$ locally for the last coordinate as a real analytic function $x_q = \varphi(u)$ of $u = (x_1, \dots, x_{q-1})$. Thus a neighborhood of p in $\mathcal{M}$ is parametrized by

$$x = \Phi(u) = (u, \varphi(u)), \qquad h(u, \varphi(u)) = 0, \qquad u \in U \subseteq \mathbb{R}^{q-1},$$

where U is open. In particular, the coefficient on this graph is $u \mapsto A_\ell(\Phi(u))$. By shrinking U, continuity ensures that both $\partial_{x_q} h(\Phi(u))$ and $A_\ell(\Phi(u))$ remain nonzero throughout U. Each difference $(P_i - P_j) \circ \Phi$ is nonconstant on every open subset of U: otherwise analytic continuation on the connected patch $\mathcal{M}$ would make $P_i - P_j$ constant throughout $\mathcal{M}$. Its gradient is therefore nonzero on an open dense subset of U. There are only finitely many differences, so we can choose $u_0 \in U$ where all these gradients are nonzero and $A_\ell(\Phi(u_0)) \neq 0$. Choose a direction v such that

$$\nabla\big((P_i - P_j) \circ \Phi\big)(u_0) \cdot v \neq 0 \qquad (i \neq j).$$

The forbidden directions lie in finitely many proper linear hyperplanes, which cannot cover $\mathbb{R}^{q-1}$.

For sufficiently small t, set $x(t) = \Phi(u_0 + tv)$. Its first $q-1$ coordinates are affine functions of t. Its last coordinate satisfies the polynomial equation

$$h(u_0 + tv, x_q(t)) = 0.$$

This equation is nonzero as a polynomial in (t, x_q), because $\partial_{x_q} h$ is nonzero at the starting point. Thus all coordinates of $x(t)$ are algebraic functions of t, and so are $f_j(t) = P_j(x(t))$ and $a_j(t) = A_j(x(t))$. Our choice of v ensures $(f_i - f_j)'(0) \neq 0$, while $a_\ell(0) \neq 0$. Restricting the assumed identity gives

$$\sum_j a_j(t) e^{f_j(t)} = 0,$$

contradicting the one-variable independence just proved. □

Lemma L.4 (No algebraic hypersurface barrier)**.** *Under conditions 1, 2, and 3 of Section L.2, put $D = Nnd$ and write $X = (x_i^a)_{a,i} \in \mathbb{R}^D$ for the batch. There do not exist a nonconstant real polynomial h in these D scalar coordinates and a nonempty connected relatively open subset $\mathcal{M}$ of*

$$\{X \in \mathcal{B}_{N,n,d} : h(X) = 0,\ \nabla h(X) \neq 0\}$$

such that both batch attention fields are tangent to $\mathcal{M}$ at every point.

Proof. Suppose that such a polynomial h and set $\mathcal{M}$ exist. Thus $h = 0$ is a polynomial equation for the hypothetical hypersurface, and $\mathcal{M}$ is a local portion of this hypersurface, called a *patch*.

Step 1: Polynomial divisibility on $\mathcal{M}$. After shrinking $\mathcal{M}$, we may assume that h is irreducible over $\mathbb{R}$. Indeed, factor h and fix a point of $\mathcal{M}$. Since $\nabla h \neq 0$ there, exactly one irreducible factor vanishes at that point, with multiplicity one. The other factors remain nonzero in a sufficiently small neighborhood. Replacing h by the vanishing factor and restricting $\mathcal{M}$ to a connected portion of this neighborhood preserves the hypersurface and its tangent spaces.

We will use the following polynomial divisibility fact: if a real polynomial F vanishes on $\mathcal{M}$, then h divides F. Fix a point $p \in \mathcal{M}$. Since $\nabla h(p) \neq 0$, we can relabel the coordinates as

$(u, y) \in \mathbb{R}^{D-1} \times \mathbb{R}$ so that $\partial_y h(p) \neq 0$. The implicit function theorem gives a local parametrization $X = (u, \varphi(u))$ of $\mathcal{M}$, where u ranges over a nonempty open set $U \subseteq \mathbb{R}^{D-1}$. On this graph,

$$h(u, \varphi(u)) = F(u, \varphi(u)) = 0 \qquad \text{for every } u \in U.$$

Suppose that h does not divide F. Regard both polynomials as polynomials in the single variable y, and allow their coefficients to be rational functions of $u = (u_1, \dots, u_{D-1})$. These coefficients form the field $K = \mathbb{R}(u_1, \dots, u_{D-1})$. The condition $\partial_y h(p) \neq 0$ ensures that h has positive degree in y. Moreover, its coefficients in $\mathbb{R}[u]$ have no common nonconstant factor, since such a factor would contradict the irreducibility of h in $\mathbb{R}[u, y]$. Gauss's lemma therefore implies that h remains irreducible in $K[y]$, and that divisibility by h of a polynomial $F \in \mathbb{R}[u, y]$ is the same in these two polynomial rings. Thus h does not divide F in $K[y]$ either, so their greatest common divisor in $K[y]$ is 1.

The Euclidean algorithm for polynomials in one variable over a field now gives a Bézout identity

$$\widetilde{B}(u, y)h(u, y) + \widetilde{C}(u, y)F(u, y) = 1, \qquad \widetilde{B}, \widetilde{C} \in K[y].$$

Choose a nonzero polynomial $c(u)$ that is a common denominator of the finitely many rational coefficients of $\widetilde{B}$ and $\widetilde{C}$. Multiplying by $c(u)$ gives $B = c\widetilde{B}$ and $C = c\widetilde{C}$ in $\mathbb{R}[u, y]$, and hence the polynomial identity

$$B(u, y)h(u, y) + C(u, y)F(u, y) = c(u), \qquad c \neq 0.$$

This identity is valid for every (u, y), including points where $c(u) = 0$. Substituting $y = \varphi(u)$ makes both terms on the left zero, so $c(u) = 0$ for every $u \in U$. A real polynomial vanishing on a nonempty open set is the zero polynomial: one can apply the corresponding one-variable fact successively to each coordinate on an open box contained in U. This contradicts the choice $c \neq 0$ and proves that h divides F.

Step 2: At least one generator is resonant. Since $T_X\mathcal{M} = \ker dh_X$, the assumed tangency of the two batch fields means precisely

$$dh_X\big[f_r^{(N)}(X)\big] = \sum_{a=1}^{N}\sum_{i=1}^{n} \partial_{x_i^a} h(X) \cdot f_r^{(N)}(X)_i^a = 0, \qquad X \in \mathcal{M}, \quad r = 1, 2,$$

where $\partial_{x_i^a} h \in \mathbb{R}^d$ is the gradient with respect to the coordinates of token x_i^a. For one generator (W, V), fix $X \in \mathcal{M}$ and write

$$s_{a,i,j} = (x_j^a)^\top W x_i^a, \qquad Z_{a,i} = \sum_{j=1}^{n} e^{s_{a,i,j}}, \qquad b_{a,i,j} = (V x_j^a) \cdot \partial_{x_i^a} h(X).$$

The tangency equation is

$$0 = \sum_{a,i} \frac{\sum_j b_{a,i,j} e^{s_{a,i,j}}}{Z_{a,i}}.$$

Each $Z_{a,i}$ is positive. Multiplying by $\prod_{a,i} Z_{a,i}$ cancels the denominator in each summand, giving

$$0 = \sum_{a,i} \left(\sum_j b_{a,i,j} e^{s_{a,i,j}} \right) \prod_{(b,k) \neq (a,i)} \left(\sum_\ell e^{s_{b,k,\ell}} \right).$$

Expanding the products amounts to choosing one key index for each token position. Such a choice is a map $\sigma : \{1, \dots, N\} \times \{1, \dots, n\} \to \{1, \dots, n\}$, where $\sigma(a, i)$ selects a key in sequence a; different token positions may select the same key. For a fixed σ, every summand has the common exponential factor $\prod_{a,i} e^{s_{a,i,\sigma(a,i)}} = e^{\sum_{a,i} s_{a,i,\sigma(a,i)}}$. Summing over the distinguished token position (a, i) adds the coefficients $b_{a,i,\sigma(a,i)}$. Thus we obtain

$$\sum_\sigma e^{P_\sigma(X)} A_\sigma(X) = 0 \quad \text{on } \mathcal{M}, \tag{43}$$

where

$$P_\sigma = \sum_{a,i} (x^a_{\sigma(a,i)})^\top W x^a_i, \qquad A_\sigma = \sum_{a,i} (V x^a_{\sigma(a,i)}) \cdot \partial_{x^a_i} h.$$

Call the generator nonresonant on $\mathcal{M}$ if every difference $P_\sigma - P_\tau$, $\sigma \neq \tau$, is nonconstant on $\mathcal{M}$. We first suppose that both generators are nonresonant on $\mathcal{M}$. We will show that $T_X\mathcal{M} = \mathbb{R}^D$ for every $X \in \mathcal{M}$, contradicting the fact that $\mathcal{M}$ is a hypersurface of dimension $D-1$.

For either generator (W, V), Lemma L.3 applies to (43) and gives $A_\sigma = 0$ on $\mathcal{M}$ for every σ. Let L_V denote the linear batch field defined by $(L_V(X))^a_i = V x^a_i$, which applies the same matrix V to every token. Taking $\sigma(a,i) = i$ for every token position gives

$$A_\sigma(X) = \sum_{a,i} (V x^a_i) \cdot \partial_{x^a_i} h(X) = dh_X[L_V(X)] = 0 \qquad \text{on } \mathcal{M}.$$

Hence L_V is tangent to $\mathcal{M}$.

Now take two of the identities $A_\sigma = A_\tau = 0$ whose index maps agree at every token position except (a,i), where $\sigma(a,i) = j$ and $\tau(a,i) = k$. Keeping the same X, W, and V, subtracting these identities gives

$$0 = A_\sigma(X) - A_\tau(X) = V(x^a_j - x^a_k) \cdot \partial_{x^a_i} h(X).$$

Thus the vector in $\mathbb{R}^D$ with component $V(x^a_j - x^a_k)$ at position (a,i) and zero at every other position belongs to $T_X\mathcal{M}$. Since this holds for every $X \in \mathcal{M}$, the corresponding linear field is tangent to $\mathcal{M}$.

Fix a, i, j, k. We have obtained tangent linear fields whose only nonzero component is $B(x^a_j - x^a_k)$ at position (a,i), for $B = V_1$ and $B = V_2$. The fields L_{V_1} and L_{V_2} are also tangent to $\mathcal{M}$. Taking the bracket of L_{V_r} with such a field, for $r = 1, 2$, gives a field of the same form, with B replaced by $BV_r - V_r B$. Since linear combinations and Lie brackets preserve tangency, repeating these operations gives tangent fields for all matrices in the linear span of nested commutators of V_1, V_2, including the generators themselves. This span is $\mathrm{Lie}_{\mathrm{mat}}(V_1, V_2)$, which equals $\mathrm{gl}_d(\mathbb{R})$ by condition 1 of Section L.2. Choose $j \neq k$. Since the tokens within each sequence are distinct, $v = x^a_j - x^a_k \neq 0$, and $\{Bv : B \in \mathrm{gl}_d(\mathbb{R})\} = \mathbb{R}^d$. Consequently, $T_X\mathcal{M}$ contains every vector supported at a single token position. These vectors span $\mathbb{R}^D$, contradicting $\dim T_X\mathcal{M} = D - 1$.

Consequently, for at least one of the two generators (W, V), there exist distinct index maps σ, τ and a real constant c such that $P_\sigma - P_\tau = c$ on $\mathcal{M}$.

Step 3: At most one generator is resonant. Fix such a generator, such maps, and this constant. Order all $M = Nn$ tokens and write the chosen exponent difference as

$$\begin{aligned} Q_{C,W}(X) &\coloneqq P_\sigma(X) - P_\tau(X) \\ &= \sum_{r,s=1}^{M} C_{rs} x_r^\top W x_s = X^\top (C \otimes W) X, \end{aligned}$$

where $C \neq 0$ is block diagonal by sequences, with key rows and query columns, and $X = (x_1^\top, \ldots, x_M^\top)^\top$. By construction,

$$Q_{C,W}(X) = P_\sigma(X) - P_\tau(X) = c \qquad \text{for every } X \in \mathcal{M}.$$

Let $H = H^\top$ be the symmetric matrix representing this quadratic form, so that

$$Q_{C,W}(X) = X^\top H X.$$

A matrix and its symmetric part define the same quadratic form. Thus, writing $S = (W + W^\top)/2$ and $K = (W - W^\top)/2$, we obtain

$$H = \frac{C \otimes W + C^\top \otimes W^\top}{2} = \mathrm{sym}(C) \otimes S + \mathrm{skew}(C) \otimes K. \tag{44}$$

If $\mathrm{sym}(C) \neq 0$, choose a fixed $t = (t_1, \ldots, t_M) \in \mathbb{R}^M$ with $t^\top C t \neq 0$. Restrict $Q_{C,W}$ to the linear subspace $X = (t_1 z, \ldots, t_M z)$, with $z \in \mathbb{R}^d$ variable: each token x_r is replaced by the scalar multiple $t_r z$ of the same vector z. Substitution gives

$$Q_{C,W}(t_1 z, \ldots, t_M z) = \sum_{r,s} C_{rs} t_r t_s \, z^\top W z = (t^\top C t) z^\top S z.$$

This restricted quadratic form has matrix $(t^\top Ct)S$, of rank d because $t^\top Ct \neq 0$ and S is invertible. Restriction to a linear subspace cannot increase the rank of a quadratic form, so $\mathrm{rank}(H) \geq d > 2$. This is an algebraic restriction of the polynomial; the substituted configurations need not belong to $\mathcal{M}$ or to $\mathcal{B}_{N,n,d}$. If $\mathrm{sym}(C) = 0$, then C is skew and $\mathrm{rank}(H) = \mathrm{rank}(C)\,\mathrm{rank}(K) \geq 4 > 2$. That means that in any case: $\mathrm{rank}(H) > 2$. Thus $Q_{C,W} - c$ is irreducible over $\mathbb{C}$ for every constant c: a product of two affine linear factors would have quadratic part of rank at most two. Since $Q_{C,W} - c$ vanishes on $\mathcal{M}$ by construction, the divisibility fact above gives $h \mid Q_{C,W} - c$, so irreducibility forces h to be proportional to $Q_{C,W} - c$.

Suppose now that both generators are resonant on $\mathcal{M}$. Their index pairs may differ, giving coefficient matrices $C_1, C_2 \neq 0$ and constants c_1, c_2. By the preceding argument, both $Q_{C_1,W_1} - c_1$ and $Q_{C_2,W_2} - c_2$ are proportional to the same polynomial h. Hence, for some $\lambda \in \mathbb{R} \setminus \{0\}$,

$$Q_{C_1,W_1} - c_1 = \lambda\big(Q_{C_2,W_2} - c_2\big).$$

Comparing the terms of degree two gives

$$Q_{C_1,W_1} = \lambda Q_{C_2,W_2}, \qquad H_1 = \lambda H_2,$$

where H_r is the symmetric matrix representing Q_{C_r,W_r}. The decomposition in (44) then forces either S_1, S_2 or K_1, K_2 to be proportional, contradicting condition 2 of Section L.2.

Step 4: Excluding the remaining case. Thus exactly one generator is resonant; denote it by r and the nonresonant generator by s. Applying the nonresonance argument to generator s and choosing $\sigma(a, i) = i$ shows that L_{V_s} is tangent to $\mathcal{M}$. After multiplying h by a nonzero constant, we may write $h = Q_{C,W_r} - c$. Tangency means that the polynomial $F(X) := dh_X[L_{V_s}(X)]$ vanishes on $\mathcal{M}$. Since L_{V_s} applies V_s to every token, differentiating each quadratic term in both token variables gives

$$\begin{aligned} F(X) &= \sum_{p,q} C_{pq}\big[(V_s x_p)^\top W_r x_q + x_p^\top W_r (V_s x_q)\big] \\ &= Q_{C,V_s^\top W_r + W_r V_s}(X). \end{aligned}$$

The divisibility fact gives $h \mid F$. As h has degree two and F has degree at most two, the quotient is a scalar:

$$F = \gamma h = \gamma(Q_{C,W_r} - c), \qquad \gamma \in \mathbb{R}.$$

Comparing the terms of degree two therefore yields

$$Q_{C,V_s^\top W_r + W_r V_s} = \gamma Q_{C,W_r}$$

(the constant terms also give $\gamma c = 0$).

The symmetric and skew-symmetric parts of $V_s^\top W_r + W_r V_s$ are respectively $V_s^\top S_r + S_r V_s$ and $V_s^\top K_r + K_r V_s$. Using (44) and separating the symmetric and skew-symmetric factors in the token indices therefore gives

$$\begin{aligned} \mathrm{sym}(C) \otimes (V_s^\top S_r + S_r V_s - \gamma S_r) = 0, \\ \mathrm{skew}(C) \otimes (V_s^\top K_r + K_r V_s - \gamma K_r) = 0. \end{aligned}$$

If $\mathrm{sym}(C) \neq 0$, the first identity implies $V_s^\top S_r + S_r V_s = \gamma S_r$. Otherwise $\mathrm{skew}(C) = C \neq 0$, and the second implies $V_s^\top K_r + K_r V_s = \gamma K_r$. Either relation contradicts condition 3 of Section L.2, namely (41). □

L.4 Polynomial lifts and common invariant divisors

Some notations and background. Write $X = (x_1, \ldots, x_D) \in \mathbb{C}^D$ for the complexified batch. Enumerate the $p = 2Nn^2$ score quadratics in (42) as $Q_1, \ldots, Q_p$. For each Q_ν, introduce a formal variable z_ν representing the exponential $e^{Q_\nu(X)}$. In the algebraic calculations, the z_ν are algebraically independent of one another and of the state coordinates: the relation $z_\nu = e^{Q_\nu(X)}$ is imposed only when evaluating the resulting expressions at the original exponential functions.

Since exponentials never vanish, we also allow inverses of these variables and work in the Laurent polynomial ring

$$A = \mathbb{C}[x_1, \ldots, x_D, z_1^{\pm 1}, \ldots, z_p^{\pm 1}],$$

where $z_\nu^{\pm 1}$ means that both z_ν and its inverse $z_\nu^{-1} = 1/z_\nu$ are allowed. Thus an element of A has the form

$$P(X, z) = \sum_{\beta \in \mathbb{Z}^p} p_\beta(X) z^\beta, \qquad z^\beta = \prod_{\nu=1}^{p} z_\nu^{\beta_\nu},$$

with only finitely many nonzero coefficients $p_\beta \in \mathbb{C}[X] = \mathbb{C}[x_1, \dots, x_D]$. The exponents of the z_ν may therefore be negative, while the coefficients remain polynomials in the state coordinates.

We also write $z_{r,a,i,j}$ for the variable corresponding to $Q_{r,a,i,j}$; the variables with a fixed $r \in \{1, 2\}$ form that generator's score block. A single subscript $\nu \in \{1, \dots, p\}$ enumerates all these variables, and $z = (z_\nu)_{\nu=1}^p$.

For a linear state field $L(X) = (L_1(X), \dots, L_D(X))$, we also use L to denote the associated differential operator: for a polynomial f in the state coordinates,

$$(Lf)(X) \coloneqq df_X[L(X)] = \sum_{k=1}^{D} L_k(X) \frac{\partial f}{\partial x_k}(X).$$

In particular, LQ_ν is the derivative of Q_ν in the direction of the field L. Define the horizontal lift of L by

$$\mathcal{D}_L = L + \sum_{\nu=1}^{p} z_\nu (LQ_\nu) \partial_{z_\nu}.$$

For a key choice σ, put $z_r^\sigma = \prod_{a,i} z_{r,a,i,\sigma(a,i)}$ and $(L_{r,\sigma}(X))_i^a = V_r x^a_{\sigma(a,i)}$. Multiplying the lifted attention field by the product of its row denominators yields the exact polynomial derivation

$$\delta_r = \sum_\sigma z_r^\sigma \mathcal{D}_{L_{r,\sigma}}. \tag{45}$$

Indeed, let $F_r(X, z)$ denote the attention field with each $e^{Q_{r,a,i,j}(X)}$ replaced by $z_{r,a,i,j}$, and set $H_r(z) = \prod_{a,i} (\sum_j z_{r,a,i,j})$. Clearing denominators and expanding the product of row sums gives

$$H_r F_r = \sum_\sigma z_r^\sigma L_{r,\sigma}.$$

The lift $F \mapsto F + \sum_\nu z_\nu dQ_\nu[F]\, \partial_{z_\nu}$ is linear in the field components. Applying it to this identity therefore gives (45). The identity key choice $\sigma(a, i) = i$ gives $L_{r,\sigma} = L_{V_r}$.

Lemma L.5 (Reduction to a single monomial in z)**.** *Suppose the score quadratics are linearly independent and L_{V_1}, L_{V_2} are invertible on homogeneous quadratics. If $0 \neq P \in A$ satisfies $\delta_r P = B_r P$ with $B_r \in A$, $r = 1, 2$, then*

$$P = z^\beta p(X)$$

for a Laurent monomial z^β and a polynomial $p \in \mathbb{C}[X]$.

Proof. Write $P = \sum_\beta p_\beta(X) z^\beta$ and $Q_\beta = \sum_\nu \beta_\nu Q_\nu$. The horizontal lift preserves each exponent:

$$\mathcal{D}_L(p_\beta z^\beta) = z^\beta \big(L p_\beta + (LQ_\beta) p_\beta \big). \tag{46}$$

Step 1: Determine the possible monomials in B_r. Fix $r \in \{1, 2\}$ and first suppose that $B_r \neq 0$. We use an elementary property of Laurent polynomials: the smallest and largest exponents of any one variable add under multiplication of nonzero polynomials. The same holds for the smallest and largest sums of exponents over any fixed group of variables. Indeed, group the terms by the exponent or sum in question. The product of the two groups with largest values is nonzero and gives the largest value in the product; the argument for the smallest values is identical.

By (46), every monomial of $\delta_r P$ comes from multiplying a monomial of P by some z_r^σ. Such multiplication never decreases any individual exponent. Since $\delta_r P = B_r P$, additivity of the smallest exponents implies that every exponent in B_r is nonnegative. The variables of the other score block do not occur in z_r^σ: their smallest exponents cannot decrease and their largest exponents cannot increase in $\delta_r P$. Thus their smallest exponents in B_r are at least zero and their largest at most zero, so they do not occur in B_r.

For each position (a, i), multiplication by z_r^σ increases the sum of the exponents of $z_{r,a,i,1}, \ldots, z_{r,a,i,n}$ by exactly one. The same additivity argument shows that, among the monomials of B_r, the smallest such sum is at least one and the largest is at most one. Every such sum therefore equals one. As the individual exponents are nonnegative integers, each monomial of B_r contains exactly one of these variables, to the first power, for every position (a, i). Moreover, (46) increases degree in X by at most two. Since this degree adds in nonzero products, $\deg_X B_r \le 2$. We have proved

$$B_r = \sum_\sigma b_{r,\sigma}(X) z_r^\sigma, \qquad \deg_X b_{r,\sigma} \le 2. \tag{47}$$

If $B_r = 0$, the same formula holds with all coefficients zero.

Step 2: Compare a monomial of largest weight. Assign a real weight w_ν to each variable z_ν, so that the weight of z^β is $w \cdot \beta = \sum_\nu w_\nu \beta_\nu$. Choose weights satisfying

$$w_{r,a,i,i} > w_{r,a,i,j} \qquad \text{for every } a, i \text{ and } j \ne i.$$

Then z_r^{id}, where $\mathrm{id}(a, i) = i$, has strictly larger weight than every other z_r^σ. The weights in the other score block are unrestricted. We may also choose the weights so that distinct monomials of P have different weights: ties require one of finitely many equalities $w \cdot (\alpha - \beta) = 0$, which an arbitrarily small perturbation avoids while preserving the strict inequalities.

Let β be the exponent of the monomial of P with largest weight. In $\delta_r P = B_r P$, the coefficient of $z^\beta z_r^{\mathrm{id}}$ can only come from this monomial of P and the identity choice: every other pair has smaller total weight. Comparing this coefficient gives

$$L_{V_r} p_\beta + (L_{V_r} Q_\beta - b_{r,\mathrm{id}}) p_\beta = 0.$$

Since $L_{V_r} p_\beta$ has degree at most $\deg_X p_\beta$, comparison at degree $\deg_X p_\beta + 2$, followed by cancelling the nonzero homogeneous part of highest degree of p_β, gives

$$L_{V_r} Q_\beta = (b_{r,\mathrm{id}})_{[2]}, \tag{48}$$

where $[2]$ denotes the homogeneous quadratic part. The right side does not depend on the weights. Invertibility of L_{V_r} on homogeneous quadratics therefore determines Q_β uniquely, and linear independence of the Q_ν determines β uniquely. For this fixed r, the same exponent, denoted by β_*, thus has largest weight for every such choice without ties.

Step 3: Show that all monomials have the same exponent. Every weight vector satisfying the strict inequalities above can be approximated by choices without ties. Passing to the limit gives $w \cdot (\alpha - \beta_*) \le 0$ for every exponent α appearing in P and every such weight vector. Split the weights and exponents into their two score blocks. Fix any weights w_r satisfying the strict inequalities and let w_{3-r} vary freely. We obtain

$$w_r \cdot (\alpha_r - \beta_{*,r}) + w_{3-r} \cdot (\alpha_{3-r} - \beta_{*,3-r}) \le 0 \quad \text{for every } w_{3-r}.$$

If $\alpha_{3-r} - \beta_{*,3-r} \ne 0$, taking $w_{3-r} = t(\alpha_{3-r} - \beta_{*,3-r})$ and letting $t \to +\infty$ contradicts this inequality. Hence every exponent appearing in P has the same coordinates in block $3 - r$. Applying the argument for $r = 1$ and $r = 2$ shows that all exponents coincide. Thus $P = z^\beta p(X)$, as claimed. □

L.5 Analytic hypersurfaces and the Ax–Schanuel theorem

For functions $f_1, \ldots, f_\ell$, the notation $\mathbb{C}(f_1, \ldots, f_\ell)$ denotes the field of rational expressions in these functions with complex coefficients. Its *transcendence degree*, denoted by $\mathrm{trdeg}_{\mathbb{C}}\, \mathbb{C}(f_1, \ldots, f_\ell)$, is the largest number of these functions that are *algebraically independent*: no nonzero polynomial with complex coefficients vanishes identically when evaluated on them. For example, $\mathrm{trdeg}_{\mathbb{C}}\, \mathbb{C}(t, t^2) = 1$, since t^2 is already a polynomial in t.

We use the following theorem for complex-analytic functions.

Theorem L.6 (Ax–Schanuel; Ax, 1971)**.** *Let $g_1, \ldots, g_s$ be complex-analytic functions on a connected open neighborhood $U \subseteq \mathbb{C}^m$ of the origin. Suppose they are linearly independent over $\mathbb{Q}$ modulo constants: if $\sum_i q_i g_i$ is constant on U, with $q_i \in \mathbb{Q}$, then every q_i is zero. Then*

$$\mathrm{trdeg}_{\mathbb{C}}\, \mathbb{C}(g_1, \ldots, g_s, e^{g_1}, \ldots, e^{g_s}) \ge s + \max_{t \in U} \mathrm{rank} \left(\frac{\partial g_i}{\partial t_j}(t) \right).$$

The rank on the right is the largest rank attained by the Jacobian on U.

Only the functions on this neighborhood are needed; no extension to all of $\mathbb{C}^m$ is required.

Lemma L.7 (An analytic singular hypersurface gives a lifted divisor)**.** *Assume the conditions of Section L.2, and suppose a Lie determinant is nonzero at some state. A smooth real-analytic hypersurface patch $H \subseteq \Sigma_{\theta,N,n}$ tangent to both fields would give a nonunit irreducible $P \in A$ satisfying $\delta_r P \in (P)$ for $r = 1, 2$. Moreover, no nonzero polynomial with complex coefficients vanishes on any nonempty relatively open subset of H.*

Proof. Suppose a nonzero real polynomial vanished on a nonempty relatively open subset of H. Restrict to a connected open portion. Factoring the polynomial and using the identity theorem for real-analytic functions shows that one real irreducible factor h vanishes throughout this portion. The set $\{h = 0, \nabla h = 0\}$ has dimension at most $D-2$, so it cannot contain this smooth $(D-1)$-dimensional portion. Near a point where $\nabla h \neq 0$, the portion is therefore relatively open in the smooth algebraic hypersurface $h = 0$ and remains tangent to both fields, contradicting Lemma L.4. Separating real and imaginary parts gives the same conclusion for polynomials with complex coefficients.

Extend a real-analytic local parametrization of H to complex parameters using its convergent power series. After restricting the parameter domain, this gives a complex-analytic map $X(t) = (x_1(t), \dots, x_D(t))$ on a connected open neighborhood $U \subseteq \mathbb{C}^{D-1}$ of 0, with $\operatorname{rank} DX(t) = D - 1$; its restriction to real parameters parametrizes an open portion of H. The image $X(U)$ is *Zariski dense* in $\mathbb{C}^D$: this means that the only polynomial in $\mathbb{C}[X]$ vanishing throughout $X(U)$ is the zero polynomial. Indeed, any such identity would also hold on the real portion of H, which was just ruled out. Row denominators are nonzero at $X(0)$ and remain nonzero after shrinking U. The tangency identities likewise extend from real to complex parameters by their convergent power series.

Apply Theorem L.6 to the $p + D$ functions

$$Q_1(X(t)), \dots, Q_p(X(t)), x_1(t), \dots, x_D(t).$$

They are linearly independent modulo constants even over $\mathbb{C}$: a relation would give a polynomial vanishing on $X(U)$, hence the zero polynomial. Separating its homogeneous degrees and using independence of the score quadratics forces every coefficient to vanish. Their Jacobian has rank $D - 1$ because it includes DX. Consequently,

$$\operatorname{trdeg}_{\mathbb{C}} \mathbb{C}\big(X(t), e^{Q(X(t))}, e^{X(t)}\big) \geq p + 2D - 1,$$

where exponentials of tuples are taken componentwise. Removing the D auxiliary coordinate exponentials lowers transcendence degree by at most D, giving

$$\operatorname{trdeg}_{\mathbb{C}} \mathbb{C}\big(X(t), e^{Q(X(t))}\big) \geq p + D - 1. \tag{49}$$

These auxiliary exponentials are used only in the proof.

Let $\mathcal{O}(U)$ denote the ring of complex-analytic functions on U. The substitution homomorphism

$$A \longrightarrow \mathcal{O}(U), \qquad x_i \mapsto x_i(t), \quad z_\nu \mapsto e^{Q_\nu(X(t))}$$

has a prime kernel I, since on connected U a product of two nonzero complex-analytic functions cannot vanish identically. The nonzero Lie determinant, after replacing exponentials by formal variables and clearing row denominators, gives a nonzero element of I: it vanishes on H, but is not identically zero under ambient analytic substitution. This relation shows that the $D + p$ functions $X(t), e^{Q(X(t))}$ are not algebraically independent. Together with (49), it follows that their generated field has transcendence degree exactly $D + p - 1$.

Choose $D + p - 1$ algebraically independent functions among these generators, and let K be their field of rational expressions. The remaining function satisfies a nonzero polynomial equation with coefficients in K. A nonzero polynomial of least degree that vanishes at this function divides every other such polynomial, by the Euclidean division algorithm. Clearing its denominators and removing the common factor of its coefficients gives a polynomial P in the original variables. Gauss's lemma makes P irreducible and shows that every polynomial relation is divisible by P. This remains true for Laurent polynomial relations after clearing negative powers of the z_ν. Hence $I = (P)$ in A. Inverting the z_ν preserves irreducibility unless P becomes a unit. This is impossible: P vanishes under substitution, whereas all inverted variables become nonzero exponentials.

The chain rule intertwines the horizontal lifts with differentiation on the exponential graph. Tangency to H therefore makes I invariant under the cleared derivations (45). Hence $\delta_r P = B_r P$ for some $B_r \in A$, as required. □

L.6 CONNECTEDNESS OF THE INTERPOLATION DOMAIN

Theorem 3.5 (Universal interpolation with two Gaussian attention blocks)**.** *Fix $d \geq 3$. Consider two noncausal single-head attention blocks with head width $r \geq d$ and query, key, value, and output projection matrices whose entries are all sampled independently from $\mathcal{N}(0,1)$. With probability one, simultaneously for every $N \geq 1$ and $n \geq 2$, there exists an open and connected subset*

$$\mathcal{R}_{N,n} \subseteq \mathcal{B}_{N,n,d},$$

depending on the sampled matrices, whose complement in $\mathcal{B}_{N,n,d}$ has Lebesgue measure zero. On $\mathcal{R}_{N,n}$, these two frozen blocks yield universal interpolation at continuous depth in the sense of Definition 2.4, with $m_N = 2$. The same sampled matrices work for every N and n; only the block selection, signs, and durations depend on the interpolation problem.

Proof of Theorem 3.5. Fix a pair in the probability-one event of Lemma L.2 and Section L.2, and fix N, n. Put $D = Nnd$ and $\Sigma = \Sigma_{\theta,N,n}$.

Step 1: A finite rank test and its preservation along flows. Write the reciprocal softmax denominators as

$$u_{r,a,i}(X) \coloneqq \left(\sum_j e^{Q_{r,a,i,j}(X)} \right)^{-1}.$$

Define

$$\mathcal{A} = \mathbb{R}\big[x_1, \dots, x_D, e^{Q_1(X)}, \dots, e^{Q_p(X)}, (u_{r,a,i}(X))_{r,a,i}\big].$$

This algebra is fixed for the chosen θ, N, n; X is the variable of its functions. More precisely, its generators depend only on W_1, W_2 and N, n: the entries of V_1, V_2 appear only as real coefficients in the attention fields. Here square brackets denote all finite sums and products of the listed functions, with real coefficients. No algebraic independence of these functions is assumed. Every function in $\mathcal{A}$ is real analytic because the softmax denominators are positive. The attention formula (32) shows that every component of the two batch fields belongs to $\mathcal{A}$.

Partial derivatives also remain in $\mathcal{A}$. Indeed, derivatives of the state coordinates are constants, while

$$\begin{aligned} \partial_{x_k} e^{Q_\nu} &= (\partial_{x_k} Q_\nu) e^{Q_\nu}, \\ \partial_{x_k} u_{r,a,i} &= -u_{r,a,i}^2 \sum_j (\partial_{x_k} Q_{r,a,i,j}) e^{Q_{r,a,i,j}}. \end{aligned}$$

Since the Q_ν are quadratic, their derivatives are linear polynomials in the state coordinates. Both right sides therefore belong to $\mathcal{A}$, and the sum and product rules give the same conclusion for every function in $\mathcal{A}$. Finally, each component of a Lie bracket is a sum of products of field components and their first derivatives. Hence all iterated Lie brackets of the two fields have components in $\mathcal{A}$ as well.

Since $\mathcal{A}$ is generated by finitely many functions over $\mathbb{R}$, Hilbert's basis theorem has the following consequence: any family of vectors with D components in $\mathcal{A}$ is generated by finitely many of its members, using coefficients in $\mathcal{A}$ (Shafarevich, 1990, §6, pp. 44–45). Applied to the two fields and all their iterated Lie brackets, this gives a finite list $Y_1, \dots, Y_b$ among them such that every original field or iterated bracket Z satisfies

$$Z(X) = \sum_{\ell=1}^{b} a_{Z,\ell}(X) Y_\ell(X), \qquad a_{Z,\ell} \in \mathcal{A}.$$

The same finite list $Y_1, \dots, Y_b$ works for every X. Consequently, the columns of $Y(X) = (Y_1(X), \dots, Y_b(X))$ span the Lie algebra at X, and

$$\Sigma = \{X \in \mathcal{B}_{N,n,d} : \operatorname{rank} Y(X) < D\}.$$

Equivalently, all determinants formed by choosing D columns of $Y(X)$ vanish. This describes Σ by finitely many real-analytic equations. It has no interior by Lemma L.2.

Each bracket $[f_r^{(N)}, Y_\ell]$ is also such a linear combination, so $[f_r^{(N)}, Y] = YC_r$ for an analytic $b \times b$ matrix C_r, with the bracket taken column by column. Let ϕ_t be the local flow of $f_r^{(N)}$ starting at X. Differentiating $D\phi_t(X)^{-1} Y(\phi_t(X))$ gives

$$Y(\phi_t(X)) = D\phi_t(X) Y(X) R(t), \qquad R'(t) = R(t) C_r(\phi_t(X)), \quad R(0) = I_b.$$

Both $D\phi_t(X)$ and $R(t)$ are invertible, the latter by uniqueness for this linear matrix differential equation. Hence the rank of Y is unchanged along either flow. In particular, a trajectory starting in Σ remains in Σ for as long as it stays in $\mathcal{B}_{N,n,d}$.

Step 2: The exceptional set has dimension at most $D-2$. Suppose instead that $\dim_{\mathbb{R}} \Sigma = D-1$. A standard property of real-analytic zero sets then gives a neighborhood in which Σ is a smooth real-analytic hypersurface H; see Bierstone & Milman (1988, Section 2). The preceding flow property makes both fields tangent to H. The construction in Lemma L.7 gives a nonzero $P \in A$ such that

$$P(X, e^{Q(X)}) = 0 \quad \text{on } H, \qquad \delta_r P = B_r P \quad (r = 1, 2), \quad B_r \in A.$$

By Lemma L.5, $P = z^\beta p(X)$ for a nonzero polynomial $p \in \mathbb{C}[X]$. Substitution yields

$$0 = e^{\sum_\nu \beta_\nu Q_\nu(X)} p(X) \qquad \text{on } H.$$

Since the exponential factor never vanishes, p vanishes on H. This contradicts the last assertion of Lemma L.7. Thus $\dim_{\mathbb{R}} \Sigma \le D-2$.

Step 3: The remaining domain is connected. The sets excluded from $\mathcal{B}_{N,n,d}$ consist of proportional token pairs within a sequence, zero tokens, and pairs of sequences related by a permutation and an optional overall sign change. Their dimensions are, respectively, $D-d+1$, $D-d$, and $D-nd$, all at most $D-2$ because $d \ge 3$. Together with Σ, they form the excluded set

$$E = (\mathbb{R}^D \setminus \mathcal{B}_{N,n,d}) \cup \Sigma.$$

Real-analytic zero sets can be covered by countably many smooth pieces, each of dimension at most that of the set itself (Bierstone & Milman, 1988, Corollary 2.11). Each smooth piece is a smooth submanifold of $\mathbb{R}^D$. If its dimension is k, it admits local parametrizations by k real variables. Here Σ is defined inside $\mathcal{B}_{N,n,d}$ by the analytic determinant equations of Step 1, and has dimension at most $D-2$ by Step 2. The sets excluded from $\mathcal{B}_{N,n,d}$ are defined by polynomial equations and satisfy the same dimension bound. Applying the cited result separately to these sets therefore covers E by smooth pieces of dimension at most $D-2$. We can then cover these pieces by countably many local parametrizations of the form

$$Y = \psi(u), \qquad u \in U \subseteq \mathbb{R}^k, \qquad k \le D-2,$$

where U is open and ψ is smooth.

Take $X_0, X_1 \notin E$. We seek an intermediate point Z such that both segments $[X_0, Z]$ and $[Z, X_1]$ avoid E. Fix $i \in \{0, 1\}$. If $[X_i, Z]$ meets E at Y, then $Y = X_i + \lambda(Z - X_i)$ for some $\lambda \in [0, 1]$. Since $X_i \notin E$, we have $Y \ne X_i$ and thus $\lambda > 0$. Solving for Z, setting $t = 1/\lambda \ge 1$, and writing $Y = \psi(u)$ gives

$$Z = X_i + t(Y - X_i) = X_i + t(\psi(u) - X_i), \qquad u \in U, \quad t \ge 1.$$

Thus these forbidden choices of Z lie on rays from X_i through points $Y \in E$, at or beyond Y. For each parametrization, this expression depends smoothly on $k+1 \le D-1$ parameters, namely u and t, so its image has D-dimensional Lebesgue measure zero. Taking the countable union over all parametrizations still gives a measure-zero set. We can therefore choose Z outside the two measure-zero sets of forbidden intermediate points. Both segments $[X_0, Z]$ and $[Z, X_1]$ then avoid E. This proves that $\mathcal{R}_{\theta,N,n} = \mathbb{R}^D \setminus E$ is path connected. It is open by the rank test in Step 1, and dense with full Lebesgue measure in $\mathcal{B}_{N,n,d}$ by Lemma L.2.

Step 4: Interpolation with the same two sampled blocks. At every point of this connected open domain, the two fields and their iterated Lie brackets span $\mathbb{R}^D$. The criterion in Corollary 2.7 therefore joins any two configurations in the domain by finitely many successive flows of fields from $\{\pm f_1^{(N)}, \pm f_2^{(N)}\}$, with the entire trajectory remaining in the domain. The flow durations, and hence the terminal time $S > 0$, may depend on the prescribed endpoints.

Finally, the matrix conditions do not depend on N, n, and the probability-one event of Lemma L.2 already holds simultaneously for every finite N, n. Hence all conclusions hold for the same sampled pair. □

L.7 Exact interpolation at finite depth

The passage to residual layers follows from Proposition 2.9, applied to the scalar multiples of the two sampled fields.

Proof of Corollary 3.7. Fix a draw in the probability-one event established in the proof of Theorem 3.5. For every $N \geq 1$ and $n \geq 2$, the two batch fields $f_1^{(N)}, f_2^{(N)}$ are analytic on the open connected domain $\mathcal{R}_{\theta,N,n}$, and their Lie algebra has full rank there by (33). Apply Proposition 2.9 with $D = Nnd$, $\mathcal{M} = \mathcal{R}_{\theta,N,n}$, and

$$\mathcal{G} = \{cf_r^{(N)} : c \in \mathbb{R},\ r \in \{1, 2\}\}.$$

This family is closed under real scalar multiplication, and its two generators are smooth and have full Lie rank on $\mathcal{M}$. It joins every pair of configurations in this domain by finitely many residual layers

$$\mathrm{Id} + c_k f^{(N)}_{W_{r_k}, V_{r_k}}, \qquad r_k \in \{1, 2\}, \quad c_k \in \mathbb{R},$$

with every intermediate configuration still in the domain. Each scalar gain and block index is shared across the batch. The matrices remain fixed; only the depth, indices, and gains depend on the prescribed endpoints. This is universal interpolation at finite depth with $m_N = 2$ in the sense of Section 2.3. The argument is deterministic for every N, n once the draw is fixed, so no additional probability-one event is needed. □

Remark L.8 (Scope of the conclusion). The exceptional set $\Sigma_{\theta,N,n}$ may be nonempty. The theorem gives two-field interpolation on the connected generic domain (33), whereas Theorem 3.3 gives a bank of $2Nnd$ multihead blocks with Gaussian-initialized projections, valid at every point of $\mathcal{B}_{N,n,d}$ for any fixed number of heads and positive head widths. Its rank-one special case is described in Remark J.3. The two-block result requires a single head of width at least d. Corollary 3.7 gives exact finite-depth interpolation on the same domain, with adjustable real scalar gains and the same two frozen fields. No bound on the required depth is asserted.

M Numerical tests of the head-width threshold

We test whether $m \geq t := \lceil d/r \rceil$ frozen noncausal single-head blocks of width $r < d$ suffice for generic full lifted Lie rank in dimension $D = Nnd$. The condition $mr \geq d$ is necessary: all bracket values lie tokenwise in the sum of the output projection images, of dimension at most mr, so the Lie rank is at most $Nn \min(d, mr)$. A single generator has Lie rank at most one.

Protocol. We use $(d, r) \in \{(3,1),(3,2),(4,1),(4,2),(4,3),(6,1),(6,2),(6,3)\}$, $(N, n) \in \{(1,2),(2,2),(1,4),(2,3)\}$, and $m \in \{t-1, t, t+1\}$. All query, key, value, and output projection entries are independent centered Gaussians of variance $v \in \{1, 1/d\}$. Using seeds $0, 1, 2$ for each setting gives 576 cases with $6 \leq D \leq 36$. The state X is sampled independently with standard Gaussian coordinates. Comparisons are paired: increasing m takes prefixes of the same bank, banks are shared across (N, n), and the two variances rescale the same base draws. Rescaling changes the fields.

We compute the spaces $W_k(X)$ of Section 2.2.2 using nested PyTorch Jacobian-vector products in float64, retaining all generated bracket fields. Step k includes brackets of length at most $k+1$. Evaluated columns of norm above 10^{-14} are normalized, and singular values exceeding $\max\{10^{-12}, \varepsilon\sigma_1\}$ are counted. We use $\varepsilon = 10^{-8}$ for stopping and record $10^{-6}, 10^{-10}$ for sensitivity. Runs stop at full rank or the first pointwise rank plateau, subject to $k \leq 6$ and 4096 words. Analyticity justifies plateau stopping at generic points for exact ranks; numerical plateaus can be premature.

Results. The table reports full-rank counts at the final step of the original runs, before any continuation. Each entry contains 96 cases.

Projection variance v	Blocks m	Full rank at 10^{-8}	At all three tolerances
1	$t-1$	0/96	0/96
1	t	94/96	89/96
1	$t+1$	96/96	96/96
$1/d$	$t-1$	0/96	0/96
$1/d$	t	96/96	90/96
$1/d$	$t+1$	96/96	96/96

All 192 below-threshold cases are structurally excluded from full rank. At $m = t$, each of the 64 combinations (v, d, r, N, n) has at least one seed reaching full rank at all three tolerances. The two primary deficits occur at $v = 1$. For $(d, r, N, n, m) = (4, 1, 2, 2, 4)$, seed 0 stops at rank $15/16$; continuing the same fields and state through $k = 5$ gives rank 16 only at $\varepsilon = 10^{-10}$. For $(6, 3, 2, 3, 2)$, seed 1 reaches rank $35/36$ at $k = 6$ and full primary rank at $k = 7$. These separate checks do not replace the original counts.

Larger dimension. A separate pilot fixes $D = 96$, $d = 3$, $r = 1$, $m = 3 = t$, $v = 1/d$, and seed 0, using the same projections for $(N, n) = (16, 2), (8, 4), (4, 8)$. At $k = 5$, the respective rank triplets, in tolerance order $(10^{-6}, 10^{-8}, 10^{-10})$, are $(93, 95, 96)$, $(94, 96, 96)$, and $(96, 96, 96)$. Continuing the first case through $k = 6$ gives $(92, 95, 96)$; relative numerical rank need not increase when columns are added. Thus full rank is observed for all three shapes at 10^{-10}, but only two at the primary tolerance.

These results support the proposed threshold across the tested batch sizes and sequence lengths, with sensitivity to conditioning at the boundary. They do not establish a universal interpolation theorem, which would also require control of the connectedness of the full-rank domain. The experiment is reproduced by `run_threshold.py`, using `lie_rank.py` and the helpers in `experiment_utils.py`; saved results include rank histories, singular values, and stopping reasons.

N Proof of Proposition 3.8

Proof. Fix $X \in \mathcal{A}_{N,n,d}$. Every single-head causal field has first-token block VX_1; for a multi-head field, this block is $(\sum_{h=1}^{H} V_h)X_1$. More generally, if two fields have first-token blocks AX_1 and CX_1 for constant matrices A, C, their Lie bracket has first-token block $(CA - AC)X_1$. Thus both the direct span and all iterated brackets lie in the space on the right of (4).

It remains to show that all directions in this space are attained. We adapt the argument from the proof of Lemma 3.1 in Section H to causal prefixes. Define the prefix barycenters

$$z_{\ell,i}(b) := \frac{\sum_{j=1}^{i} e^{\langle b, x_j^\ell \rangle} x_j^\ell}{\sum_{j=1}^{i} e^{\langle b, x_j^\ell \rangle}}, \qquad F^{\mathrm{c}}_{W,V}(X)_{i,\ell} = V z_{\ell,i}(W x_i^\ell).$$

Fix a target block (i_0, ℓ_0) with $i_0 \geq 2$, and put $q = x_{i_0}^{\ell_0}$. Choose a separating the keys in each sequence as in (20), and choose $\alpha \in q^\perp$ nonorthogonal to every other token as in (21). These choices are possible because the tokens are nonzero and pairwise noncollinear. Set

$$W_t^{\pm} = a \left(t\alpha \pm \frac{q}{\|q\|^2} \right)^\top.$$

At the target block their difference gives Vc, where $c = z_{\ell_0,i_0}(a) - z_{\ell_0,i_0}(-a)$. This vector is nonzero: we apply the argument in Section H to the function

$$\psi(s) := \log \sum_{j=1}^{i_0} e^{s\langle a, x_j^{\ell_0} \rangle}$$

to obtain $\psi''(s) > 0$, since $i_0 \geq 2$ and the numbers $\langle a, x_j^{\ell_0} \rangle$ are distinct. Hence $\langle a, c \rangle = \psi'(1) - \psi'(-1) > 0$.

At any other query, the two score vectors tend along a to the same signed infinity. Both prefix barycenters therefore converge to the same unique maximizing or minimizing key. At a first-token block they are already identical, since $z_{\ell,1}(b) = x_1^\ell$ for every b. Hence

$$F^{\mathrm{c}}_{W_t^+,V}(X) - F^{\mathrm{c}}_{W_t^-,V}(X) \longrightarrow E_{i_0,\ell_0}(Vc),$$

where $E_{i_0,\ell_0}(u)$ has block u at the target and zeros elsewhere. The pointwise span $\mathcal{S}^{\mathrm{c}}_X$ is finite-dimensional and closed, so it contains this limit. Thus the limiting direction is a finite linear combination of field values with finite parameters. Taking $V = uc^\top/\|c\|^2$ shows that it contains $E_{i_0,\ell_0}(u)$ for every $u \in \mathbb{R}^d$. Thus every later-token direction can be varied independently.

Finally, write any $A \in \mathbb{R}^{d\times d}$ as $A = \sum_r V_r$ with $\operatorname{rank} V_r \le 1$. Then $\sum_r F^{\mathrm{c}}_{0,V_r}(X)$ has first-token block AX_1. The directions constructed above allow all later-token blocks to be adjusted arbitrarily without changing AX_1. This proves (4). The image of $A \mapsto AX_1$ has dimension $d \operatorname{rank} X_1$, giving (5); the bracket upper bound is therefore sharp as well. Any head of positive width realizes these rank-at-most-one fields, with the remaining heads set to zero, so the same span and rank hold for the multi-head family.

For completeness, the first-token constraint also obstructs global interpolation on $\mathcal{A}_{N,n,d}$ when $N > d$. Choose a nonzero relation $X_1 c = 0$ and an index ℓ with $c_\ell \neq 0$. Scaling only x_1^ℓ by two breaks this relation while staying in the same component of $\mathcal{A}_{N,n,d}$ along the positive scaling path. Every causal flow and finite causal residual network preserves the original relation, so neither reaches this target. □

O Proof of Corollary 3.9

Proof. Let $r = N - \dim K$. Choose a full-row-rank matrix $Q \in \mathbb{R}^{r\times N}$ with $\ker Q = K$. A first-token matrix X_1 annihilates K if and only if it has the unique form $X_1 = PQ$ for $P \in \mathbb{R}^{d\times r}$. Thus E_K is a linear space of dimension $R = N(n-1)d + dr$. Within E_K, the condition $\ker X_1 = K$ is equivalent to P being injective, since Q is surjective onto $\mathbb{R}^r$. Equivalently, $\operatorname{rank} P = r$, an open condition because a nonzero $r \times r$ minor remains nonzero under small perturbations. Since $\mathcal{A}_{N,n,d}$ is open, it follows that $\mathcal{A}^K_{N,n,d}$ is relatively open in E_K. Every point therefore has a small ball in E_K contained in this set. This ball is connected and hence lies in the same connected component as its center, so each component is relatively open.

Every causal field maps E_K into E_K: for $X \in E_K$ and $c \in K$, its first-token block satisfies $(V_\Sigma X_1)c = V_\Sigma(X_1 c) = 0$. Now fix $X \in \mathcal{A}^K_{N,n,d}$ and $Z \in E_K$. The rule $A(X_1 u) = Z_1 u$ defines a linear map on $\operatorname{im} X_1$: if $X_1 u = X_1 v$, then $u - v \in \ker X_1 = K$, so $Z_1 u = Z_1 v$. Extending A linearly to $\mathbb{R}^d$ gives $Z_1 = AX_1$. Thus every $Z \in E_K$ satisfies the condition in (4), and Proposition 3.8 yields

$$\operatorname{span}\{F^{\mathrm{c}}_{W,V}(X) : \operatorname{rank} W, \operatorname{rank} V \le 1\} = E_K.$$

Choose linear coordinates identifying E_K with $\mathbb{R}^R$. The restricted fields are analytic and closed under negation by replacing V with $-V$. On each connected component, Corollary D.1 gives the continuous-depth conclusion on a horizon $[0, S]$ depending on the endpoints. Closure under real scalar multiplication by scaling V also permits applying Proposition E.1 to obtain the finite residual network. Both constructions stay in that component. The rank-at-most-one factorizations give one head of width one; additional prescribed heads can be set to zero. □

P Proof of Corollary 3.10

Proof. Fix K as in the statement and a linear isomorphism $T : E_K \to \mathbb{R}^R$. In these coordinates, define

$$G(\theta, z) := T(F^{\mathrm{c}}_{ab^\top,uv^\top}(T^{-1}z)), \qquad \theta = (a, b, u, v) \in \mathbb{R}^{4d}.$$

The domain $T(\mathcal{A}^K_{N,n,d})$ is a nonempty open subset of $\mathbb{R}^R$. The function G is jointly real analytic in (θ, z), since T is linear and all prefix softmax denominators are strictly positive. Its values span $\mathbb{R}^R$ at every configuration by Proposition 3.8. Apply Theorem I.2 with parameter dimension $4d$ and with both configuration and output dimensions equal to R. It follows that $2R$ independent Gaussian-factor

samples span simultaneously throughout this domain with probability one. The same event covers all its connected components, for this fixed K.

Fix a draw on this event and consider the signed family

$$\{\pm F^{\mathrm{c}}_{W_j,V_j} : 1 \le j \le 2R\}.$$

This family is closed under negation and has pointwise full rank in E_K at every point of the class. Applying Corollary D.1 in the coordinates T on each connected component gives exact interpolation on a horizon $[0, S]$ depending on the endpoints, within that component. Each control segment selects one of the same $2R$ sampled generators and its sign, so this is the relative universal interpolation property of Definition 2.4. The signed dictionary contains at most $4R$ fields, all realized by the same architecture since $-F^{\mathrm{c}}_{W_j,V_j} = F^{\mathrm{c}}_{W_j,-V_j}$.

For finite depth, apply Proposition E.1 to the scalar-closed family $\{cF^{\mathrm{c}}_{W_j,V_j} : c \in \mathbb{R},\ 1 \le j \le 2R\}$. A sufficiently fine Euler discretization of an interpolating trajectory first gives a network of L_0 residual layers, where L_0 is finite and may depend on the interpolation problem. Appending at most R correction layers then yields exact interpolation, with total depth at most $L_0 + R$ and scalar gains shared across the batch. Linearity in V preserves the rank-at-most-one realization. The sampled matrices are fixed before the endpoints are chosen, proving universal interpolation at finite depth with the same $2R$ fields.

□