TRANSFORMERS AS EFFECTIVE FIELDS: FROM SYMMETRY BREAKING TO ARCHITECTURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Can Transformer-like structures be derived from structural assumptions on a learnable field? We formulate an axiomatic effective-field theory that does not postulate attention in advance. For analytic matrix potentials with token-isometry invariance and feature-basis covariance, we classify the Taylor expansion through fourth order. The quadratic term generates token-wise linear feature mixing, while the quartic Hessian generates a cubic interaction with a finite paired linear-attention decomposition. This pairing reflects the trace-self-adjoint structure imposed by a scalar potential; relaxing the exchange tying opens a trace-anti-self-adjoint, locally non-potential sector that accommodates generic untied attention. A separate low-complexity assumption yields a small number of bottleneck heads. Row-softmax weights remain token-permutation equivariant and admit a row-wise entropy-regularized variational characterization, but softmax normalization does not in general restore potential compatibility for an untied update. We further quantify how a finite Transformer-like model tracks a fixed effective flow. Under uniform regularity, its trajectory error separates into an $O(T^{-1})$ depth-discretization term and width-dependent best field-approximation errors for attention and MLP. Controlled experiments diagnose these terms through frozen-field depth refinement and learned field surrogates. The field perspective therefore provides a low-order route from symmetry to architecture, with explicit limits on potential compatibility and approximation.

1 INTRODUCTION

The Transformer couples long-range, data-dependent token mixing with a feature-wise MLP and has become a general computational template across modalities (Vaswani et al., 2017; Dehghani et al., 2019). Its success has also produced a large space of block-level modifications, many of which are found empirically and do not transfer uniformly across implementations or tasks (Narang et al., 2021). A structural theory should therefore do more than rationalize a finished block: it should explain why attention and MLP arise as the leading operations, state the assumptions under which that conclusion holds, and identify controlled changes to those assumptions that generate new architectures.

Residual networks suggest a natural language for this purpose. With layer index interpreted as time, a residual block is a numerical step of a learned vector field. This viewpoint has been constructive, not merely analogical: dynamical stability and well-posedness lead to constrained deep architectures (E, 2017; Haber & Ruthotto, 2017), numerical integration schemes motivate new residual blocks (Lu et al., 2018), and continuous-depth parameterizations replace a prescribed stack by a learned differential equation (Chen et al., 2018). At the same time, a discrete network represents a single smooth continuous field only under additional regularity across depth (Sander et al., 2022); the field interpretation is a mathematical hypothesis with testable conditions, not an automatic relabeling of every deep network.

Energy and dynamical viewpoints have begun to clarify Transformers as well. Self-attention coincides with an update of a modern Hopfield energy (Ramsauer et al., 2021); Transformer layers can be related to descent steps for an underlying objective (Yang et al., 2022); and continuous-depth attention can be studied as an interacting-particle system (Geshkovski et al., 2025). More constructively, prescribed energies and representation objectives have produced the Energy Transformer and white-box Transformer-like blocks (Hoover et al., 2023; Yu et al., 2023). Together, these results show

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.

why a field or objective can be a design instrument: changing the admissible energy, symmetry, or discretization changes the resulting computation in a traceable way.

These constructions motivate a complementary structural question: which computational building blocks arise from a class of learnable potentials specified by analyticity and symmetry? A low-order classification can reveal the resulting feature-wise and token-mixing operations, identify the parameter relations required for potential compatibility, and distinguish these structural consequences from additional assumptions needed for compact architectural realizations. This provides a systematic framework for constructing and interpreting Transformer-like models through an effective-field perspective.

We develop this theory as a low-order inverse-design problem. We begin with a locally analytic matrix potential and token-space isometries, and classify the Taylor terms needed to produce the first interaction. The quadratic Gram term gives token-wise linear feature mixing, while the quartic Gram Hessian gives a cubic, potential-compatible interaction. Trace self-adjointness of this Hessian yields paired linear attention. Learnable feature tensors transform covariantly under changes of feature basis. A finite head representation follows algebraically, whereas a small number of low-rank heads requires an independent compressibility assumption. This complements geometric deep learning (Bronstein et al., 2021) and the physical principle that asymmetrical effects require asymmetrical causes (Curie, 1894).

Contributions.

(i) We prove a fourth-order Gram normal form for analytic token-isometry-invariant potentials. Its gradient consists of a linear feed-forward branch and paired linear self-attention. The resulting field has the leading form

$$\frac{\partial V_\Theta}{\partial \boldsymbol{X}^*} = \boldsymbol{X}\boldsymbol{A} + \frac{1}{2}\sum_{h=1}^{r} (\boldsymbol{X}\boldsymbol{B}_h\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{C}_h + \boldsymbol{X}\boldsymbol{C}_h\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{B}_h) + O(\|\boldsymbol{X}\|_{\mathrm{F}}^5),$$

where the two cubic summands form a trace-adjoint pair. Under the bottleneck realization, $\boldsymbol{B}_h = \boldsymbol{Q}_h\boldsymbol{K}_h^*$ and $\boldsymbol{C}_h = \boldsymbol{V}_h\boldsymbol{O}_h^*$. Trace self-adjointness of the quartic Gram operator guarantees the existence of a paired linear-attention realization.

(ii) We prove a finite depth–width refinement theorem that separates numerical resolution from representation error. Depth contributes an explicit $O(T^{-1})$ discretization term, while nested attention and MLP potential classes make the best attainable field errors non-increasing with width. We connect the theorem to controlled evidence. A fixed learned field displays near-first-order convergence when only its Euler depth is refined, and wider potential models reduce a held-out field-error proxy for analytic and projected ViT targets.

(iii) We characterize how relaxing exchange tying extends the low-order construction beyond the potential-compatible class. Writing $G = X^*X$, an untied cubic interaction admits the decomposition

$$\boldsymbol{X}\boldsymbol{B}_h G\boldsymbol{C}_h = \frac{1}{2}\boldsymbol{X}\big(\boldsymbol{B}_h G\boldsymbol{C}_h + \boldsymbol{C}_h G\boldsymbol{B}_h\big) + \frac{1}{2}\boldsymbol{X}\big(\boldsymbol{B}_h G\boldsymbol{C}_h - \boldsymbol{C}_h G\boldsymbol{B}_h\big).$$

The first term corresponds to a trace-self-adjoint Gram operator and is generated by a quartic Gram potential. The second corresponds to a trace-anti-self-adjoint operator, whose nonzero component cannot arise as the Hessian of such a potential. This decomposition identifies the structural constraint imposed by potential generation and the additional interactions admitted when exchange tying is relaxed.

2 Background

Notation Let $[n] = \{0, 1, \ldots, n-1\}$ be the set of n indices. The discrete state is written as $x(t, \omega, \sigma) \in \mathbb{F}$, where $t \in \mathcal{T} := [T]$ is the discrete layer index, $\omega \in \Omega := [N]$ indexes tokens (words/pixels), and $\sigma \in \Sigma := [C]$ indexes neurons or vector field components. $\mathbb{F}$ is a number field either real $\mathbb{R}$ or complex $\mathbb{C}$. The symbols T, N, and C denote the depth, the sequence length (number of tokens), and the feature space dimension (network width), respectively. In a continuous-layer setting we inherit the notation $\mathcal{T} := [0, 1]$, and write $\dot{x} := \frac{\mathrm{d}x}{\mathrm{d}t}$ for the layer-wise derivative and omit the layer as an implicit variable. For brevity, we assume Σ, Ω are finite and denote the discrete

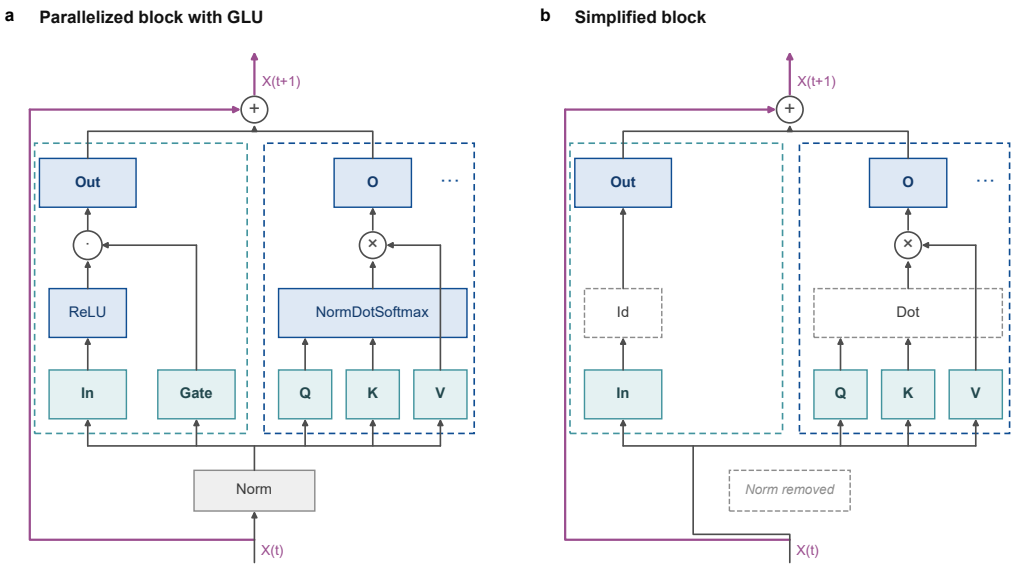

Figure 1: **Left:** a practical parallelized Transformer block with GLU activation. The wide MLP can be equivalently separated into multiple narrow heads. **Right:** a simplified Transformer block eliminating activation functions, normalizations and softmax.

state as a layer-dependent matrix $\boldsymbol{X} \in \mathbb{F}^{N\times C}$ whose row and column entries are $\boldsymbol{X}_{\omega\sigma} \coloneqq x(\omega,\sigma)$, and denote each token at $\omega \in \Omega$ as a vector $x_\omega \in \mathbb{F}^C$ in the feature space whose components are $(x_\omega)_\sigma \coloneqq x(\omega,\sigma)$. The parameters are the layer-dependent matrices $\boldsymbol{W} = \boldsymbol{W}(t) \in \mathbb{F}^{C\times C}$. We use $\odot$ for element-wise (Hadamard) operations. The Lie bracket $[f,g] \coloneqq fg - gf$ denotes the commutator. $\boldsymbol{W}^\top$ and $\boldsymbol{W}^*$ are the real and conjugate / Hermitian transposes of $\boldsymbol{W}$ respectively. $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. $U(C)$ is the C-dimensional unitary group, S_N is the N-dimensional permutation group, and $M_C(\mathbb{C}) \coloneqq \mathbb{C}^{C\times C}$ is the space of complex feature matrices. We use T for an attention superoperator and $\mathcal{L}$ for the quartic Gram Hessian; their trace adjoints are defined below.

2.1 Transformer as ODE

Figure 1 (left) visualizes the architecture of a parallelized Transformer (Dehghani et al., 2023) as a two-layer ResNet with mixed nonlinearities. Suppose the parameters of the i-th attention head are $\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i, \boldsymbol{O}_i \in \mathbb{F}^{C\times C_{\mathrm{A}}}$, the parameters of the MLP are $\boldsymbol{W}_{\text{gate}}, \boldsymbol{W}_{\text{in}}, \boldsymbol{W}_{\text{out}} \in \mathbb{F}^{C\times C_{\text{MLP}}}$, and the number field is real $\mathbb{F} = \mathbb{R}$. Layer dependency is implicit in the notation, and normalization is ommited as it is approximately linear in high dimension.

Definition 1 (Transformers)**.** The ODE form of a Transformer is defined as

$$\dot{\boldsymbol{X}} = ((\boldsymbol{X}\boldsymbol{W}_{\text{gate}}) \odot \mathrm{ReLU}(\boldsymbol{X}\boldsymbol{W}_{\text{in}}))\boldsymbol{W}_{\text{out}} + \sum_{i=1}^{n} \mathrm{Softmax}(C_{\mathrm{A}}^{-1/2}\boldsymbol{X}\boldsymbol{Q}_i\boldsymbol{K}_i^\top\boldsymbol{X}^\top)\boldsymbol{X}\boldsymbol{V}_i\boldsymbol{O}_i^\top. \quad (1)$$

It can be more concisely written as $\dot{\boldsymbol{X}} = \lambda(\boldsymbol{X}\boldsymbol{W}_{\text{all}})\boldsymbol{W}'_{\text{all}}$ where $\lambda \coloneqq \lambda_{\text{MLP}} \oplus \lambda_{\mathrm{A}}$, or in expanded form, $\dot{\boldsymbol{X}} = \lambda_{\text{MLP}}(\boldsymbol{X}\boldsymbol{W}_{\text{MLP}})\boldsymbol{W}'_{\text{MLP}} + \lambda_{\mathrm{A}}(\boldsymbol{X}\boldsymbol{W}_{\mathrm{A}})\boldsymbol{W}'_{\mathrm{A}}$, where $\boldsymbol{W}_{\text{all}} \coloneqq [\boldsymbol{W}_{\text{MLP}} \mid \boldsymbol{W}_{\mathrm{A}}] = [\boldsymbol{W}_{\text{gate}} \mid \boldsymbol{W}_{\text{in}} \mid \boldsymbol{Q}_1 \mid \boldsymbol{K}_1 \mid \boldsymbol{V}_1 \mid \ldots \mid \boldsymbol{Q}_n \mid \boldsymbol{K}_n \mid \boldsymbol{V}_n]$ and $\boldsymbol{W}'_{\text{all}} \coloneqq [\boldsymbol{W}'_{\text{MLP}} \mid \boldsymbol{W}'_{\mathrm{A}}] = [\boldsymbol{W}_{\text{out}} \mid \boldsymbol{O}_1 \mid \ldots \mid \boldsymbol{O}_n]$.

The more common Transformer architecture can be interpreted as a discrete operator-splitting scheme for this ODE. For a fixed continuous field, the split and parallel Euler schemes approach the same trajectory as the step size tends to zero, even when the two branches do not commute, provided the usual boundedness and Lipschitz conditions hold.

2.2 Comparison between Parallelized and Split Transformers

For a unit-time flow resolved by T layers, set $h = 1/T$. A discrete layer is then a forward Euler step $\boldsymbol{X}_{k+1} = \boldsymbol{X}_k + h\lambda(\boldsymbol{X}_k \boldsymbol{W}(t_k))\boldsymbol{W}'(t_k)$, where $t_k = kh$. Formally, the following first-order approximations link parallel and split updates:

$$(1 + h\lambda_{\text{MLP}} + h\lambda_{\text{A}}) \approx e^{h\lambda_{\text{MLP}} + h\lambda_{\text{A}}} \approx e^{h\lambda_{\text{MLP}}} e^{h\lambda_{\text{A}}} \approx (1 + h\lambda_{\text{MLP}})(1 + h\lambda_{\text{A}}) \tag{2}$$

In Equation 2, the left- and right-hand sides denote one parallel and one split step, respectively. For autonomous linear generators the middle exponentials agree exactly when the generators commute. For nonlinear or time-dependent fields, the statement needed here is instead the consistency estimate in Lemma 2; no product of nonlinear fields is interpreted as ordinary operator multiplication.

Lemma 1 (Lie–Trotter for autonomous linear generators)**.** *For bounded linear generators A and M, $e^{M+A} = \lim_{T\to\infty}(e^{M/T} e^{A/T})^T$.*

Lemma 2 (Consistency of Split and Parallel Depth Refinement)**.** *For $j \in \{\text{A}, \text{MLP}\}$, write $F_j(t, \boldsymbol{Z}) = \lambda_j^{\boldsymbol{W}(t)}(\boldsymbol{Z})$. On a control region containing the two discrete trajectories, suppose that*

$$\|F_j(t, \boldsymbol{Z})\|_{\text{F}} \le B_j,$$
$$\|F_j(t, \boldsymbol{Z}) - F_j(t, \boldsymbol{Z}')\|_{\text{F}} \le K_j \|\boldsymbol{Z} - \boldsymbol{Z}'\|_{\text{F}}, \qquad j \in \{\text{A}, \text{MLP}\}.$$

Set $B = B_{\text{A}} + B_{\text{MLP}}$, $K = K_{\text{A}} + K_{\text{MLP}}$, $h = 1/T$, and $t_k = kh$, and define

$$P_{k,h}(\boldsymbol{Z}) = \boldsymbol{Z} + hF_{\text{A}}(t_k, \boldsymbol{Z}) + hF_{\text{MLP}}(t_k, \boldsymbol{Z}), \tag{3}$$
$$S_{k,h}(\boldsymbol{Z}) = \boldsymbol{Z} + hF_{\text{A}}(t_k, \boldsymbol{Z}) + hF_{\text{MLP}}(t_k, \boldsymbol{Z} + hF_{\text{A}}(t_k, \boldsymbol{Z}))\,. \tag{4}$$

Starting from the same state, let $\boldsymbol{X}_{k+1}^{\text{par}} = P_{k,h}(\boldsymbol{X}_k^{\text{par}})$ and $\boldsymbol{X}_{k+1}^{\text{split}} = S_{k,h}(\boldsymbol{X}_k^{\text{split}})$. Then

$$\max_{0\le k\le T} \|\boldsymbol{X}_k^{\text{split}} - \boldsymbol{X}_k^{\text{par}}\|_{\text{F}} \le \frac{\Phi(K) K_{\text{MLP}} B_{\text{A}}}{T}, \qquad \Phi(K) = \begin{cases} (e^K - 1)/K, & K > 0, \\ 1, & K = 0. \end{cases} \tag{5}$$

If the fields are also Lipschitz in time with constants Λ_j, both schemes converge with first-order global error to the solution of $\dot{\boldsymbol{X}} = F_{\text{A}}(t, \boldsymbol{X}) + F_{\text{MLP}}(t, \boldsymbol{X})$. In particular, with $\Lambda = \Lambda_{\text{A}} + \Lambda_{\text{MLP}}$,

$$\max_{0\le k\le T} \|\boldsymbol{X}_k^{\text{par}} - \boldsymbol{X}(t_k)\|_{\text{F}} \le \frac{\Phi(K)(\Lambda + KB)}{2T}, \tag{6}$$

$$\max_{0\le k\le T} \|\boldsymbol{X}_k^{\text{split}} - \boldsymbol{X}(t_k)\|_{\text{F}} \le \frac{\Phi(K)}{T}\left[\frac{1}{2}(\Lambda + KB) + K_{\text{MLP}} B_{\text{A}}\right]. \tag{7}$$

The proof and explicit constants are given in Lemma 11. Appendix A separately records what a special commutator relation implies; it produces a scalar reweighting, not a sparse attention pattern by itself.

2.3 Potential-Generated Fields

To state the algebraic classification, we first specify how a scalar potential generates a matrix field.

Definition 2 (Potential-Generated Field)**.** For $V : \mathbb{C}^{N\times C} \to \mathbb{C}$ continuously differentiable in real coordinates, define F_V by the Wirtinger differential

$$F_V = \frac{\partial V}{\partial \boldsymbol{X}^*}, \qquad d_{\boldsymbol{X}^*} V[\boldsymbol{H}^*] = \text{Tr}(\boldsymbol{H}^* F_V), \quad \boldsymbol{X} \text{ held fixed.} \tag{8}$$

A field is potential-generated if $F = F_V$. Its dynamics are $\dot{\boldsymbol{X}} = F_V(\boldsymbol{X})$, with the conjugate equation obtained by conjugation.

In general, V is a complex-valued generating potential and need not admit an energy interpretation. For a real-valued energy E defined on complex states, setting $V = -2E$ yields the generated field $F_V = -\nabla_{\mathbb{R}} E$. If E has no explicit dependence on layer time, then along the continuous flow,

$$\frac{dE}{dt} = -\|\nabla_{\mathbb{R}} E\|_F^2 \le 0.$$

3 LOW-ORDER EFFECTIVE FIELDS

We characterize the lowest Taylor orders of a complex, potential-generated content backbone, illustrated by Figure 1 (right). Positional information, causal masks, normalization, nonlinear activations, and softmax are outside this classification.

3.1 LINEARIZED TRANSFORMER FIELDS

For a complex-linear map $\mathsf{T} : M_C(\mathbb{C}) \to M_C(\mathbb{C})$, define its trace adjoint $\mathsf{T}^\sharp$ by

$$\mathrm{Tr}(\boldsymbol{H}\mathsf{T}(\boldsymbol{G})) = \mathrm{Tr}(\mathsf{T}^\sharp(\boldsymbol{H})\boldsymbol{G}) \quad \text{for all } \boldsymbol{G}, \boldsymbol{H} \in M_C(\mathbb{C}). \tag{9}$$

The bilinear trace pairing is nondegenerate, so this adjoint is unique. It is not the adjoint associated with the Hermitian Frobenius inner product.

Definition 3 (Linearized Transformer Field)**.** Let $\boldsymbol{G} = \boldsymbol{X}^*\boldsymbol{X}$. A linearized Transformer field is

$$\begin{aligned} F(\boldsymbol{X}) &= \boldsymbol{X}\boldsymbol{A} + \boldsymbol{X}\mathsf{T}(\boldsymbol{G}), \\ \mathsf{T}(\boldsymbol{G}) &= \frac{1}{\sqrt{C_{\mathrm{A}}}} \sum_{i=1}^{n} \boldsymbol{B}_i \boldsymbol{G} \boldsymbol{C}_i, \end{aligned} \tag{10}$$

where $\boldsymbol{A} = \boldsymbol{W}_{\mathrm{in}}\boldsymbol{W}_{\mathrm{out}}^*$, $\boldsymbol{B}_i = \boldsymbol{Q}_i\boldsymbol{K}_i^*$, and $\boldsymbol{C}_i = \boldsymbol{V}_i\boldsymbol{O}_i^*$. Thus the interaction is $C_{\mathrm{A}}^{-1/2}\sum_i \boldsymbol{X}\boldsymbol{Q}_i\boldsymbol{K}_i^*\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{V}_i\boldsymbol{O}_i^*$.

Lemma 3 (Potential Compatibility of Linear Attention)**.** *The Hessian operator of a homogeneous quartic Gram potential is trace-self-adjoint. Conversely, if $\mathsf{T}^\sharp = \mathsf{T}$, then*

$$V_4(\boldsymbol{X}) = \frac{1}{2}\mathrm{Tr}(\boldsymbol{G}\mathsf{T}(\boldsymbol{G})), \qquad \frac{\partial V_4}{\partial \boldsymbol{X}^*} = \boldsymbol{X}\mathsf{T}(\boldsymbol{G}). \tag{11}$$

For $\mathsf{T}(\boldsymbol{G}) = \sum_i \boldsymbol{B}_i\boldsymbol{G}\boldsymbol{C}_i$, one has $\mathsf{T}^\sharp(\boldsymbol{G}) = \sum_i \boldsymbol{C}_i\boldsymbol{G}\boldsymbol{B}_i$. The trace-self-adjoint component is therefore represented by paired terms

$$\frac{\mathsf{T} + \mathsf{T}^\sharp}{2}(\boldsymbol{G}) = \frac{1}{2}\sum_i (\boldsymbol{B}_i\boldsymbol{G}\boldsymbol{C}_i + \boldsymbol{C}_i\boldsymbol{G}\boldsymbol{B}_i). \tag{12}$$

A generic untied linear-attention parameterization need not satisfy this potential-compatibility condition. Lemma 12 supplies the differential calculation.

3.2 AXIOMS AND THE FOURTH-ORDER GRAM JET

For each layer time t, suppressed below, write the potential as a parameterized family $V_\Theta(\boldsymbol{X})$.

Axiom 1 (Local Analyticity)**.** For every Θ, V_Θ is real analytic in $(\mathrm{Re}\,\boldsymbol{X}, \mathrm{Im}\,\boldsymbol{X})$ on a neighborhood of $\boldsymbol{X} = 0$.

Axiom 2 (Token Isometry Invariance)**.** For every $\boldsymbol{P} \in U(N)$, $V_\Theta(\boldsymbol{P}\boldsymbol{X}) = V_\Theta(\boldsymbol{X})$.

Axiom 3 (Learnable Feature Covariance)**.** The parameter space carries an action $\Theta \mapsto \boldsymbol{R}\cdot\Theta$ of $U(C)$ such that, for every $\boldsymbol{R} \in U(C)$,

$$V_{\boldsymbol{R}\cdot\Theta}(\boldsymbol{X}\boldsymbol{R}^*) = V_\Theta(\boldsymbol{X}). \tag{13}$$

Axiom 4 (Optional Low-Complexity Jet)**.** The degree-one and degree-three terms of the induced field admit

$$\begin{aligned} \boldsymbol{A} &= \boldsymbol{W}_{\mathrm{in}}\boldsymbol{W}_{\mathrm{out}}^*, & C_{\mathrm{MLP}} &\geq \mathrm{rank}(\boldsymbol{A}), \\ \mathcal{L}(\boldsymbol{G}) &= \frac{1}{2\sqrt{C_{\mathrm{A}}}}\sum_{h=1}^{H}(\boldsymbol{B}_h\boldsymbol{G}\boldsymbol{C}_h + \boldsymbol{C}_h\boldsymbol{G}\boldsymbol{B}_h), \end{aligned} \tag{14}$$

where $\boldsymbol{B}_h = \boldsymbol{Q}_h\boldsymbol{K}_h^*$ and $\boldsymbol{C}_h = \boldsymbol{V}_h\boldsymbol{O}_h^*$ have rank at most C_{A}, with $C_{\mathrm{A}} \ll C$ and $H \ll C^2$.

Scope of the axioms. Axiom 1 justifies a local Taylor classification; it does not make all higher-order terms vanish. Axiom 2 is a strong sufficient idealization of a content-only backbone. Positional embeddings, rotations, relative-position biases, and causal masks generally break this $U(N)$ symmetry (Vaswani et al., 2017; Heo et al., 2024). They must remain outside the theorem or enter as additional covariant inputs. Axiom 3 is a change-of-basis covariance statement, not invariance of a fixed parameter realization; only its stabilizer remains after Θ is fixed. Axiom 4 is an independent compressibility assumption. Symmetry alone implies neither few heads nor low head rank, and does not require $C_{\mathrm{MLP}} > C$.

Lemma 4 (Low-Order Gram Normal Form)**.** *Let $N \geq 2$. Under Axioms 1 and 2, the Taylor expansion of V_Θ through total degree four is*

$$V_\Theta(\boldsymbol{X}) = c_0 + \operatorname{Tr}(\boldsymbol{G}\boldsymbol{A}) + \frac{1}{2}\operatorname{Tr}(\boldsymbol{G}\mathcal{L}(\boldsymbol{G})) + R_6(\boldsymbol{X}), \quad \boldsymbol{G} = \boldsymbol{X}^*\boldsymbol{X}, \tag{15}$$

where $R_6(\boldsymbol{X}) = O(\|\boldsymbol{X}\|_{\mathrm{F}}^6)$ and $\mathcal{L} : M_C(\mathbb{C}) \to M_C(\mathbb{C})$ is complex-linear and trace-self-adjoint, $\mathcal{L}^\sharp = \mathcal{L}$. Under Axiom 3, the coefficients transform as

$$\begin{aligned} \boldsymbol{A}_{\boldsymbol{R}\cdot\Theta} &= \boldsymbol{R}\boldsymbol{A}_\Theta\boldsymbol{R}^*, \\ \mathcal{L}_{\boldsymbol{R}\cdot\Theta}(\boldsymbol{H}) &= \boldsymbol{R}\mathcal{L}_\Theta(\boldsymbol{R}^*\boldsymbol{H}\boldsymbol{R})\boldsymbol{R}^*. \end{aligned} \tag{16}$$

This statement classifies only the fourth-order jet. It does not assert an all-order representation $V = \operatorname{Tr} f(\boldsymbol{X}\boldsymbol{W}_1\boldsymbol{X}^*, \ldots)$. Higher-order invariant algebras generally require products of trace words, not just a single trace (Procesi, 1976).

Proof sketch. Scalar token rotations eliminate Taylor monomials with unequal degrees in $\boldsymbol{X}$ and $\boldsymbol{X}^*$. Through degree four, only bidegrees $(0,0)$, $(1,1)$, and $(2,2)$ remain. Unitary-invariant contractions produce linear and quadratic polynomials in the Gram entries (Weyl, 1939, Ch. II). The nondegenerate trace pairing represents the linear term by $\boldsymbol{A}$ and the polarized quadratic term by $\mathcal{L}$. Symmetry of the polarization gives $\mathcal{L}^\sharp = \mathcal{L}$. Lemma 13 in Appendix B gives the coefficient-level proof. □

Theorem 5 (Leading Potential-Generated Transformer Field)**.** *Under Lemma 4, the induced field satisfies*

$$\frac{\partial V_\Theta}{\partial \boldsymbol{X}^*} = \boldsymbol{X}\boldsymbol{A} + \boldsymbol{X}\mathcal{L}(\boldsymbol{X}^*\boldsymbol{X}) + O(\|\boldsymbol{X}\|_{\mathrm{F}}^5). \tag{17}$$

Every complex-linear operator admits a finite left–right decomposition. For trace-self-adjoint $\mathcal{L}$, it can be symmetrized as

$$\mathcal{L}(\boldsymbol{G}) = \frac{1}{2}\sum_{h=1}^{r}(\boldsymbol{B}_h\boldsymbol{G}\boldsymbol{C}_h + \boldsymbol{C}_h\boldsymbol{G}\boldsymbol{B}_h), \qquad r \leq C^2. \tag{18}$$

Consequently, the cubic term is paired linear self-attention:

$$\boldsymbol{X}\mathcal{L}(\boldsymbol{X}^*\boldsymbol{X}) = \frac{1}{2}\sum_{h=1}^{r}(\boldsymbol{X}\boldsymbol{B}_h\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{C}_h + \boldsymbol{X}\boldsymbol{C}_h\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{B}_h). \tag{19}$$

Conversely, every field in equation 19 is generated by the corresponding quartic Gram potential in equation 11. Under Axiom 4, rank factorizations give H pairs of bottleneck heads. Indexing the two heads in each pair by h and $h+H$, their effective matrices obey

$$\begin{aligned} \boldsymbol{Q}_{h+H}\boldsymbol{K}_{h+H}^* &= \boldsymbol{V}_h\boldsymbol{O}_h^*, \\ \boldsymbol{V}_{h+H}\boldsymbol{O}_{h+H}^* &= \boldsymbol{Q}_h\boldsymbol{K}_h^*, \end{aligned} \tag{20}$$

up to shared scalar rescaling. The quadratic potential is token-separable, so quartic order is the first Taylor sector that can generate cross-token dependence.

Theorem 5 classifies leading potential-generated fields; it does not assert that every standard Transformer is conservative or statistically optimal. The admissible multi-trace term $[\operatorname{Tr}(\boldsymbol{G})]^2$ is included through $\mathcal{L}(\boldsymbol{G}) = 2\operatorname{Tr}(\boldsymbol{G})\boldsymbol{I}$. Thus equation 15 does not discard multi-trace interactions. The antisymmetric operator component $(\mathsf{T} - \mathsf{T}^\sharp)/2$ is not a scalar quartic Gram Hessian. Theorem 14 provides the full decomposition proof.

Lemma 6 (Bottleneck Realization)**.** *Under Axiom 4, the leading field in equation 17 admits H pairs of width-C_{A} attention heads and a width-C_{MLP} linear feed-forward factorization. These are sufficient realization bounds, not width inequalities implied by symmetry.*

Exchange Tying and Potential-Compatibility Breaking. Lemma 6 realizes each trace-self-adjoint interaction through a pair of bottleneck attention heads. Writing

$$\boldsymbol{B}_h = \boldsymbol{Q}_h \boldsymbol{K}_h^*, \qquad \boldsymbol{C}_h = \boldsymbol{V}_h \boldsymbol{O}_h^*,$$

the potential-compatible pair is

$$F_h^{\mathrm{pair}}(\boldsymbol{X}) = \frac{1}{2\sqrt{C_{\mathrm{A}}}} \left(\boldsymbol{X}\boldsymbol{B}_h\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{C}_h + \boldsymbol{X}\boldsymbol{C}_h\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{B}_h \right).$$

The first branch uses $\boldsymbol{B}_h$ as its score map and $\boldsymbol{C}_h$ as its value–output map, while the second branch exchanges these roles.

By contrast, a generic untied linear-attention head contains

$$F_h^{\mathrm{untied}}(\boldsymbol{X}) = \frac{1}{\sqrt{C_{\mathrm{A}}}} \boldsymbol{X}\boldsymbol{B}_h\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{C}_h$$

without requiring the trace-adjoint partner $\boldsymbol{X}\boldsymbol{C}_h\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{B}_h$. At the operator level, this distinction is expressed by

$$\mathsf{T} = \frac{\mathsf{T} + \mathsf{T}^\sharp}{2} + \frac{\mathsf{T} - \mathsf{T}^\sharp}{2}.$$

The first component is trace-self-adjoint and is compatible with a quartic Gram potential. The second component is trace-anti-self-adjoint and cannot arise as the Hessian of that scalar potential.

Relaxing the exchanged-head tying therefore opens a concrete symmetry-broken sector within the low-order theory. Here, symmetry breaking refers specifically to the loss of trace self-adjointness and potential compatibility, rather than to the loss of token permutation symmetry. Standard softmax attention in Definition 1 further enlarges this class through normalization and independently learned score and value–output maps. The full-Jacobian audit in Appendix G measures the aggregate departure from potential compatibility and does not attribute all observed skew exclusively to missing exchange tying. Appendix H provides a complementary illustration of reciprocal parameter tying in particle dynamics, where a tied pair-interaction model achieves lower field and rollout errors than its endpoint-untied counterpart under the tested protocol.

3.3 Finite Depth–Width Approximation of Effective Fields

Theorem 5 identifies the leading low-order form within the potential-compatible class. Lemma 2 concerns split and parallel implementations in the continuous-depth limit. Neither result quantifies how a finite Transformer-like model tracks a prescribed field. We now separate two approximation mechanisms: depth resolves the flow, whereas width controls branch-wise field approximation. The following result applies to any fixed target field satisfying its assumptions, not only the fourth-order truncation.

Theorem 7 (Depth–Width Refinement, Informal)**.** *Consider a fixed potential-generated flow whose attention and MLP branches are uniformly bounded and Lipschitz on a common control region. Let $h = 1/T$ and use a depth-T attention-then-MLP split Transformer. If its learned branch fields approximate the two target branch fields uniformly with errors $\delta_{\mathrm{A},T}$ and $\delta_{\mathrm{MLP},T}$, then, for equal initial states,*

$$\max_{0 \le k \le T} \|\widehat{\boldsymbol{X}}_k - \boldsymbol{X}(k/T)\|_{\mathrm{F}} \le C_{\mathcal{K}} \big(T^{-1} + \delta_{\mathrm{A},T} + \delta_{\mathrm{MLP},T} \big). \tag{21}$$

If the width-dependent potential classes are nested and their best uniform field errors are $\delta_{\mathrm{A}}^*(C_{\mathrm{A}})$ and $\delta_{\mathrm{MLP}}^*(C_{\mathrm{MLP}})$, the best attainable trajectory error is therefore

$$O\big(T^{-1} + \delta_{\mathrm{A}}^*(C_{\mathrm{A}}) + \delta_{\mathrm{MLP}}^*(C_{\mathrm{MLP}}) \big). \tag{22}$$

Here $C_{\mathcal{K}}$ is independent of depth and width when the same uniform regularity bounds hold. Field density then gives convergence as depth and width grow. These are best-in-class statements for a fixed target flow; they do not imply monotone improvement for independently trained Transformers. Appendix C gives the full hypotheses, constants, statement, and proof.

Theorem 7 is a field-approximation result, separate from the low-order classification in Theorem 5. Greater depth improves numerical resolution of a fixed field. Greater width cannot worsen the best approximation over nested branch classes, although improvement need not be strict.

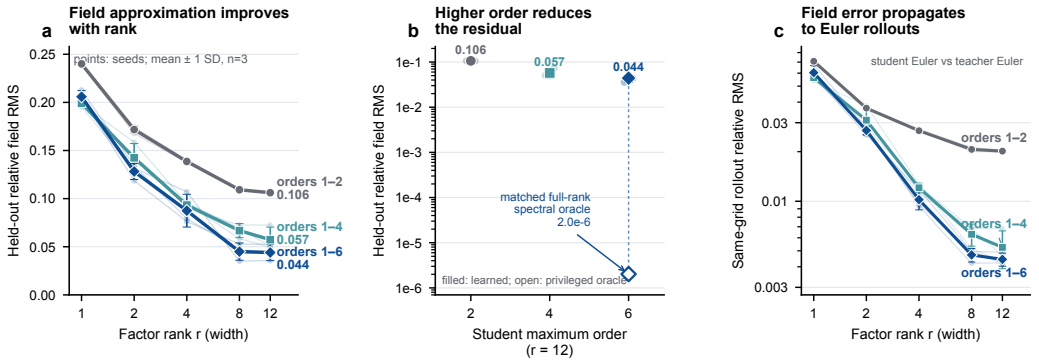

Figure 2: **Effect of factor rank and interaction order on field approximation. a**, Held-out relative field RMS versus factor rank r for students retaining orders 1–2, 1–4 or 1–6. **b**, Field error at $r = 12$; the open diamond denotes the parameter-privileged spectral oracle. **c**, Relative path RMS between student and teacher 32-step Euler trajectories. Dark curves and symbols show means, pale marks show individual runs, and error bars indicate ± 1 s.d. across $n = 3$ optimization seeds.

4 Empirical Validations of Depth and Width

Theorem 7 separates two mechanisms that should be tested separately: refining the numerical solution of a *fixed* field, and enlarging a model class used to approximate a *fixed* target field. We therefore report one controlled experiment for each mechanism. These experiments are diagnostic rather than equivalence claims: the depth test uses a parameter-shared, potential-generated model, whereas the width test measures held-out distributional errors of learned surrogate fields rather than the theorem's uniform best field errors. Full protocols and limitations are given in Appendix E.

4.1 Width-dependent approximation

We first examined whether increasing the width of a conservative network improves its approximation of a known vector field. For $X \in \mathbb{R}^{4\times 12}$, we considered the structured potential and its induced field

$$V_\theta^{P,r}(X) = \frac{\kappa}{2}\|X\|_F^2 + \sum_{p=1}^{P}\sum_{h=1}^{4} s_p c_{p,h} \frac{N}{2p} \operatorname{tr}\left(A_{p,h}^p\right), \quad A_{p,h} = \frac{X B_{p,h} B_{p,h}^\top X^\top}{N},$$

where

$$F_\theta^{P,r}(X) = -\nabla_X V_\theta^{P,r}(X), \qquad B_{p,h} \in \mathbb{R}^{12\times r}.$$

The teacher was a fixed full-rank model with maximum order $P_\star = 6$ and $r_\star = 12$, yielding an analytically known target field $F_\star$. The students used the same conservative parameterization but varied in factor-rank width and maximum interaction order: $r \in \{1, 2, 4, 8, 12\}$ and $P \in \{2, 4, 6\}$. At fixed P, increasing r enlarges a strictly nested model class because a narrow factor can be embedded into a wider factor by appending zero columns. By contrast, P controls structural interaction order. The $p = 1$ term corresponds to an MLP-like linear transformation, whereas the $p = 2$ term corresponds to a parameter-tied linear self-attention operation.

The students were trained directly on exact teacher-field evaluations using normalized field mean-squared error:

$$\mathcal{L}_{\text{field}} = \frac{\mathbb{E}_{X\in\mathcal{B}}\left\|\widehat{F}_{P,r}(X) - F_\star(X)\right\|_F^2}{\mathbb{E}_{X\in\mathcal{D}_{\text{train}}}\left\|F_\star(X)\right\|_F^2}.$$

We evaluated field approximation on held-out trajectory, perturbed-trajectory, and control-region states using relative field RMSE. Three optimization seeds were used for each combination of P and r. This design separates errors caused by insufficient factor-rank width from those caused by interaction orders absent from the student model.

4.2 Depth-dependent approximation

We next isolated the effect of network depth by freezing the trained rank-12 student fields and varying only the number of shared-weight forward-Euler steps. For a fixed student field $\widehat{F}_{P,12}$, the depth-L network was

$$\widehat{X}^{\,L}_{\ell+1} = \widehat{X}^{\,L}_{\ell} + \frac{1}{L}\widehat{F}_{P,12}\Big(\widehat{X}^{\,L}_{\ell}\Big), \qquad \ell = 0, \dots, L-1,$$

with $L \in \{4, 8, 16, 32, 64, 128, 256, 512\}$ over the fixed interval $[0, 1]$. The field parameters were shared across all layers, so increasing L reduced the Euler step size without changing the model capacity.

High-accuracy RK4 solutions were computed for both the teacher and every frozen student. Let $X^{\mathrm{cont}}_{\star}$ and $\widehat{X}^{\mathrm{cont}}_{P,12}$ denote these continuous-flow references. Using aggregate relative path RMSE d, we measured

$$E_{\mathrm{total}}(L) = d\Big(\widehat{X}^{\,L}_{P,12}, X^{\mathrm{cont}}_{\star}\Big), \quad E_{\mathrm{num}}(L) = d\Big(\widehat{X}^{\,L}_{P,12}, \widehat{X}^{\mathrm{cont}}_{P,12}\Big), \quad E_{\mathrm{model}} = d\Big(\widehat{X}^{\mathrm{cont}}_{P,12}, X^{\mathrm{cont}}_{\star}\Big).$$

Here, E_{num} measures the error removed by increasing Euler depth, whereas E_{model} measures the depth-independent discrepancy between the learned and teacher continuous flows. Therefore, increasing L can reduce time-discretization error but cannot eliminate an inaccurate learned vector field. Detailed data generation, optimization settings, RK4 accuracy checks, and metric definitions are provided in Appendix E, with the corresponding curves in Appendix Figure 3.

5 Related Works

Energy and optimization perspectives provide several routes to understanding Transformer computations. Modern Hopfield networks connect associative-memory updates to attention (Ramsauer et al., 2021), while Yang et al. (2022) relate Transformer layers to energy-minimization steps under suitable conditions. Along a constructive direction, Energy Transformer designs attention layers from an explicit energy function (Hoover et al., 2023), and CRATE develops Transformer-like blocks by unfolding optimization of a sparse rate-reduction objective (Yu et al., 2023). More recently, Hyper-SET derives symmetric attention and feedforward modules with normalization through constrained energy minimization on the hypersphere (Hu et al., 2026).

Our work contributes a complementary structural classification. Starting from local analyticity and token-isometry invariance, we characterize matrix potentials through fourth order and show that their induced leading fields admit token-wise linear feature mixing and paired linear-attention representations. This identifies the computational forms permitted by a class of potentials without selecting a specific energy or representation objective in advance. The focus is on the resulting theoretical correspondence and its approximation properties, rather than developing a new architecture or training algorithm for downstream tasks. The model realizations and controlled experiments illustrate this framework.

6 Conclusion and Perspectives

We classify the fourth-order Taylor expansion of analytic token-isometry-invariant matrix potentials. The quadratic term generates token-wise linear feature mixing, while the trace-self-adjoint quartic Hessian generates paired linear attention. A small number of low-rank heads requires a separate compressibility assumption. This low-order result distinguishes potential-compatible fields from generic untied attention without asserting an all-order single-trace representation. The depth–width theorem separately controls flow resolution and branch-wise field approximation, and the experiments diagnose these mechanisms within their stated protocols. Extensions to positional structure, mixture-of-experts, and non-attentional sequence models require additional assumptions.

AI Use Statement

Generative AI tools were used to assist with the checking of the depth–width approximation theorem and its proof, and with editing and LaTeX preparation. The authors remain responsible for independently verifying every mathematical statement and for the final content of the paper.

REFERENCES

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Corneel Casert, Isaac Tamblyn, and Stephen Whitelam. Learning stochastic dynamics and predicting emergent behavior using transformers. *Nature Communications*, 15:1875, 2024. doi: 10.1038/s41467-024-45629-w. URL https://doi.org/10.1038/s41467-024-45629-w.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 31, 2018.

Pierre Curie. Sur la symétrie dans les phénomènes physiques, symétrie d'un champ électrique et d'un champ magnétique. *Journal de physique théorique et appliquée*, 3(1):393–415, 1894.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyzdRiR9Y7.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, 2023. URL https://openreview.net/forum?id=Lhyy8H75KA.

Weinan E. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 5(1):1–11, 2017.

Changqing Fu and Laurent D Cohen. Conic activation functions. In *UniReps: the Second Workshop on Unifying Representations in Neural Models*, Proceedings of Machine Learning Research. PMLR, 2024. URL https://openreview.net/forum?id=6EDbuqER4p.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *Bulletin of the American Mathematical Society*, 62(3):427–479, 2025. doi: 10.1090/bull/1863.

Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34 (1):014004, 2017. doi: 10.1088/1361-6420/aa9a90.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, pp. 289–305. Springer, 2024.

Benjamin Hoover, Yuchen Liang, Bao Pham, Rameswar Panda, Hendrik Strobelt, Duen Horng Chau, Mohammed J. Zaki, and Dmitry Krotov. Energy transformer. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27532–27559, 2023. doi: 10.52202/075280-1197.

Yunzhe Hu, Difan Zou, and Dong Xu. Hyper-SET: Designing transformers via hyperspherical energy minimization. In *International Conference on Learning Representations*, 2026.

Laura Krick, Mireille E. Broucke, and Bruce A. Francis. Stabilisation of infinitesimally rigid formations of multi-robot networks. *International Journal of Control*, 82(3):423–439, 2009. doi: 10.1080/00207170802108441. URL https://doi.org/10.1080/00207170802108441.

Fei Lu, Ming Zhong, Sui Tang, and Mauro Maggioni. Nonparametric inference of interaction laws in systems of agents from trajectory data. *Proceedings of the National Academy of Sciences*, 116(29): 14424–14433, 2019. doi: 10.1073/pnas.1822012116. URL https://doi.org/10.1073/pnas.1822012116.

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 of *Proceedings of Machine Learning Research*, pp. 3276–3285. PMLR, 2018. URL https://proceedings.mlr.press/v80/lu18d.html.

Sharan Narang, Hyung Won Chung, Yi Tay, Liam Fedus, Thibault Févry, Michael Matena, Karishma Malkan, Noah Fiedel, Noam Shazeer, Zhenzhong Lan, et al. Do transformer modifications transfer across implementations and applications? In *Conference on Empirical Methods in Natural Language Processing*, pp. 5758–5773, 2021.

Claudio Procesi. The invariant theory of $n \times n$ matrices. *Advances in Mathematics*, 19(3):306–381, 1976. doi: 10.1016/0001-8708(76)90027-X.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021.

Michael Sander, Pierre Ablin, and Gabriel Peyré. Do residual neural networks discretize neural ordinary differential equations? *Advances in Neural Information Processing Systems*, 35:36520–36532, 2022.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Veeravalli S. Varadarajan. *Lie Groups, Lie Algebras, and Their Representations*, volume 102. Springer, 2013.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Hermann Weyl. *The Classical Groups: Their Invariants and Representations*. Princeton University Press, 1939.

Yongyi Yang, Zengfeng Huang, and David P. Wipf. Transformers from an optimization perspective. *Advances in Neural Information Processing Systems*, 35:36958–36971, 2022. doi: 10.52202/068431-2678.

Yaodong Yu, Sam Buchanan, Druv Pai, Tianzhe Chu, Ziyang Wu, Shengbang Tong, Benjamin D. Haeffele, and Yi Ma. White-box transformers via sparse rate reduction. In *Advances in Neural Information Processing Systems*, volume 36, pp. 9422–9457, 2023. doi: 10.52202/075280-0413.

A Engineering Implications

Theorem 5 concerns an idealized content backbone. Softmax, nonlinear activations, positional tensors, and masks require different potential classes or additional covariant inputs; Lemma 4 does not classify these extensions. The constructions below are separate engineering considerations, not consequences of the fourth-order theorem. In particular, geometric symmetry alone does not establish potential compatibility of an untied branch.

ReLU Projection in Activation Coordinates For a real activation matrix $\boldsymbol{U}$, ReLU is exactly the Euclidean projection

$$\mathrm{ReLU}(\boldsymbol{U}) = \Pi_{\mathbb{R}_+}(\boldsymbol{U}) := \operatorname*{argmin}_{\boldsymbol{Y}\in\mathbb{R}_+} \|\boldsymbol{U}-\boldsymbol{Y}\|_{\mathrm{F}}^2. \tag{23}$$

This statement holds in activation coordinates. In general, $\mathrm{ReLU}(\boldsymbol{X}\boldsymbol{W}_{\mathrm{in}})\boldsymbol{W}_{\mathrm{out}}^\top$ is neither idempotent nor the metric projection of $\boldsymbol{X}$ onto an affine set, so we do not identify an untied ReLU MLP with a projected gradient flow. An exact tied example is

$$\Psi_{\boldsymbol{W}}(\boldsymbol{X}) = \frac{1}{2}\|\mathrm{ReLU}(\boldsymbol{X}\boldsymbol{W})\|_{\mathrm{F}}^2, \qquad \nabla_{\boldsymbol{X}}\Psi_{\boldsymbol{W}}(\boldsymbol{X}) = \mathrm{ReLU}(\boldsymbol{X}\boldsymbol{W})\boldsymbol{W}^\top, \tag{24}$$

where the scalar squared-ReLU primitive is continuously differentiable, including at zero. Thus weight tying supplies a sufficient potential-compatible construction, but it is not a property of a general two-matrix ReLU branch. Geometric characterizations of ReLU itself, such as the one in Fu & Cohen (2024), concern the activation map and do not by themselves turn its learned linear composition into a projection in state space.

Gated MLPs and Integrability The token-isometry invariance in Axiom 2 is stronger than token exchangeability. Relaxing $U(N)$ to S_N permits row-wise nonlinear potentials, but permutation equivariance alone does not imply that a generic gated branch

$$F(\boldsymbol{X}) = (\boldsymbol{X}\boldsymbol{W}_{\mathrm{gate}}) \odot (\boldsymbol{X}\boldsymbol{W}_{\mathrm{in}})\boldsymbol{W}_{\mathrm{out}}^\top \tag{25}$$

is a gradient field. Differentiating the previously proposed triple product produces three product-rule terms; it does not recover only the displayed untied GLU term. A valid tied example is

$$\Psi_{\boldsymbol{W}}^{(3)}(\boldsymbol{X}) = \frac{1}{3}\sum_{\omega,r}(\boldsymbol{X}\boldsymbol{W})_{\omega r}^3, \qquad \nabla_{\boldsymbol{X}}\Psi_{\boldsymbol{W}}^{(3)}(\boldsymbol{X}) = (\boldsymbol{X}\boldsymbol{W})^{\odot 2}\boldsymbol{W}^\top. \tag{26}$$

More general GLU parameterizations (Shazeer, 2020) require explicit weight relations or a direct Jacobian-symmetry check before a scalar-potential interpretation is valid.

Row-Softmax Log-Partition: What It Does and Does Not Generate For real score rows and $\tau > 0$, the log-partition has the entropy-regularized variational identity

$$\tau \log \sum_k e^{s_k/\tau} = \max_{\boldsymbol{p}\in\Delta_N} \{\langle \boldsymbol{p}, \boldsymbol{s}\rangle + \tau H(\boldsymbol{p})\}, \qquad H(\boldsymbol{p}) = -\sum_k p_k \log p_k, \tag{27}$$

whose unique optimizer is $\boldsymbol{p} = \mathrm{softmax}(\boldsymbol{s}/\tau)$. This identity characterizes the weights. Differentiating the corresponding state-dependent log-partition is a separate calculation.

Lemma 8 (Row-softmax log-sum-exp derivative)**.** *Let*

$$\boldsymbol{S} = \frac{1}{\sqrt{C_{\mathrm{A}}}}\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^*, \qquad P_{jk} = \frac{e^{S_{jk}}}{\sum_{\ell=1}^N e^{S_{j\ell}}}, \qquad V_{\boldsymbol{W}}(\boldsymbol{X}) = \sqrt{C_{\mathrm{A}}}\sum_{j=1}^N \log \sum_{k=1}^N e^{S_{jk}}. \tag{28}$$

On any domain on which the logarithm is defined, with $\boldsymbol{X}$ held fixed in the Wirtinger differential of Definition 2,

$$\frac{\partial V_{\boldsymbol{W}}}{\partial \boldsymbol{X}^*} = \boldsymbol{P}^\top \boldsymbol{X}\boldsymbol{W}. \tag{29}$$

For real $\boldsymbol{X}$ and $\boldsymbol{W}$, the full Euclidean gradient is $\nabla_{\boldsymbol{X}} V_{\boldsymbol{W}} = \boldsymbol{P}^\top \boldsymbol{X}\boldsymbol{W} + \boldsymbol{P}\boldsymbol{X}\boldsymbol{W}^\top$.

Thus the natural log-sum-exp potential generates transpose-aggregated softmax, not the standard row-softmax field $\boldsymbol{P}\boldsymbol{X}\boldsymbol{W}$. Equality requires additional special structure and does not hold for generic tied or untied attention parameters. Lemma 16 gives the index-level differential.

Commutator-Induced Scalar Reweighting

Lemma 9 (BCH scalar reweighting)**.** *Under the special commutator relation* $[\lambda_{\mathrm{MLP}}, \lambda_{\mathrm{A}}] = \beta\lambda_{\mathrm{A}}$*, in a setting where the exponentials and Baker–Campbell–Hausdorff series are defined,*

$$\exp(\lambda_{\mathrm{MLP}})\exp(\lambda_{\mathrm{A}}) = \exp(\lambda_{\mathrm{MLP}} + G(\beta)\lambda_{\mathrm{A}}), \qquad G(\beta) = \frac{\beta}{1-e^{-\beta}}. \tag{30}$$

The correction is a scalar multiple of the existing attention field. It does not create a sparse token-to-token pattern unless sparsity is already present through an additional mechanism such as a mask or a structured generator.

Positional Encodings as Additional Inputs Rotary positional encodings (Heo et al., 2024) introduce index-dependent feature rotations that carry relative-position information. Jointly relabeling tokens and their positional inputs is distinct from transforming token states while holding positional tensors fixed. The latter generally breaks the token-isometry assumption in Axiom 2. Consequently, positional structure lies outside Theorem 5 unless the formulation is extended to include the corresponding covariant inputs.

B Expanded Proofs

Lemma 10 (Lie–Trotter for autonomous linear generators)**.** *For bounded linear generators* A *and* M*,* $e^{M+A} = \lim_{T\to\infty}(e^{M/T}e^{A/T})^T$.

Proof. Applying the Baker–Campbell–Hausdorff expansion (Varadarajan, 2013) $e^{Mt}e^{At} = e^{(M+A)t+\frac{1}{2}[M,A]t^2+O(t^3)}$ gives the leading non-commutative correction. □

Lemma 11 (Expanded split–parallel consistency)**.** *Under the assumptions of Lemma 2, the bounds equation 5–equation 7 hold.*

Proof. For one grid point t_k, use the maps $P_{k,h}$ and $S_{k,h}$ from equation 3–equation 4. Their one-step splitting defect satisfies

$$\begin{aligned}\|S_{k,h}(\boldsymbol{Z}) - P_{k,h}(\boldsymbol{Z})\|_{\mathrm{F}} &= h\,\|F_{\mathrm{MLP}}(t_k, \boldsymbol{Z} + hF_{\mathrm{A}}(t_k,\boldsymbol{Z})) - F_{\mathrm{MLP}}(t_k,\boldsymbol{Z})\|_{\mathrm{F}} \\ &\le h^2 K_{\mathrm{MLP}} B_{\mathrm{A}}.\end{aligned}$$

Moreover,

$$\|P_{k,h}(\boldsymbol{Z}) - P_{k,h}(\boldsymbol{Z}')\|_{\mathrm{F}} \le (1+hK)\|\boldsymbol{Z}-\boldsymbol{Z}'\|_{\mathrm{F}}.$$

Hence, for $e_k = \|\boldsymbol{X}_k^{\mathrm{split}} - \boldsymbol{X}_k^{\mathrm{par}}\|_{\mathrm{F}}$,

$$e_{k+1} \le (1+hK)e_k + h^2 K_{\mathrm{MLP}} B_{\mathrm{A}}, \qquad e_0 = 0.$$

Because $kh \le 1$ and $(1+hK)^k \le e^K$, summing the geometric series gives

$$e_k \le h\Phi(K) K_{\mathrm{MLP}} B_{\mathrm{A}},$$

which proves equation 5 with the required factor $h = 1/T$.

For the exact flow, its integral form and the state- and time-Lipschitz bounds give

$$\|\boldsymbol{X}(t_{k+1}) - P_{k,h}(\boldsymbol{X}(t_k))\|_{\mathrm{F}} \le \frac{1}{2}(\Lambda + KB)h^2.$$

Applying the same recurrence and discrete Grönwall estimate proves equation 6. The split-flow bound equation 7 follows from the triangle inequality and equation 5. This direct Euler comparison, rather than the autonomous linear Lie–Trotter formula, covers the nonlinear time-dependent fields used here. □

Lemma 12 (Potential Compatibility of Linear Attention)**.** *Let* $\mathsf{T} : M_C(\mathbb{C}) \to M_C(\mathbb{C})$ *be complex-linear, with trace adjoint defined by equation 9. A homogeneous quartic Gram potential has a trace-self-adjoint Hessian. Conversely, if* $\mathsf{T}^\sharp = \mathsf{T}$*, then equation 11 holds.*

Proof. The polarization of a homogeneous quadratic function of $\boldsymbol{G}$ is symmetric. Nondegeneracy of the bilinear trace pairing therefore represents its Hessian by $\mathsf{T} = \mathsf{T}^\sharp$. Conversely, hold $\boldsymbol{X}$ fixed, so $d\boldsymbol{G} = d\boldsymbol{X}^*\boldsymbol{X}$. Then

$$\begin{aligned} d_{\boldsymbol{X}^*} V_4 &= \frac{1}{2}\operatorname{Tr}(d\boldsymbol{G}\,\mathsf{T}(\boldsymbol{G})) + \frac{1}{2}\operatorname{Tr}(\boldsymbol{G}\,\mathsf{T}(d\boldsymbol{G})) \\ &= \operatorname{Tr}(d\boldsymbol{G}\,\mathsf{T}(\boldsymbol{G})) = \operatorname{Tr}(d\boldsymbol{X}^*\boldsymbol{X}\mathsf{T}(\boldsymbol{G})). \end{aligned}$$

Definition 2 gives $\partial V_4/\partial \boldsymbol{X}^* = \boldsymbol{X}\mathsf{T}(\boldsymbol{G})$. For $\mathsf{T}(\boldsymbol{G}) = \sum_i \boldsymbol{B}_i\boldsymbol{G}\boldsymbol{C}_i$, cyclicity gives

$$\operatorname{Tr}(\boldsymbol{H}\boldsymbol{B}_i\boldsymbol{G}\boldsymbol{C}_i) = \operatorname{Tr}(\boldsymbol{C}_i\boldsymbol{H}\boldsymbol{B}_i\boldsymbol{G}).$$

Hence $\mathsf{T}^\sharp(\boldsymbol{G}) = \sum_i \boldsymbol{C}_i\boldsymbol{G}\boldsymbol{B}_i$, proving equation 12 without an action principle. □

Lemma 13 (Low-Order Gram Normal Form)**.** *Let $N \geq 2$. Under Axioms 1 and 2, the fourth-order Taylor jet is equation 15, with $\mathcal{L}^\sharp = \mathcal{L}$. Under Axiom 3, its coefficients satisfy equation 16.*

Proof. Write the real-analytic expansion in complex coordinates as

$$V_\Theta(\boldsymbol{X}) = \sum_{p,q\geq 0} V_{p,q}(\boldsymbol{X}, \boldsymbol{X}^*),$$

where $V_{p,q}$ has degree p in $\boldsymbol{X}$ and degree q in $\boldsymbol{X}^*$. Scalar token rotations give

$$V_{p,q}(e^{i\varphi}\boldsymbol{X}, e^{-i\varphi}\boldsymbol{X}^*) = e^{i(p-q)\varphi}V_{p,q}(\boldsymbol{X}, \boldsymbol{X}^*) = V_{p,q}(\boldsymbol{X}, \boldsymbol{X}^*)$$

for every φ. Thus $V_{p,q} = 0$ unless $p = q$, leaving only bidegrees $(0,0)$, $(1,1)$, and $(2,2)$ through degree four.

Write the columns of $\boldsymbol{X}$ as $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_C \in \mathbb{C}^N$. At bidegree $(1,1)$, unitary invariance forces the token-index contraction to be proportional to the identity. Its contractions are $\boldsymbol{x}_a^*\boldsymbol{x}_b = G_{ab}$, so $V_{1,1} = \ell(\boldsymbol{G})$ for a linear functional ℓ. The nondegenerate trace pairing gives a unique matrix $\boldsymbol{A}$ with $\ell(\boldsymbol{G}) = \operatorname{Tr}(\boldsymbol{G}\boldsymbol{A})$.

At bidegree $(2,2)$, invariant contractions of two defining and two dual token indices are generated by the identity and swap pairings for $N \geq 2$ (Weyl, 1939, Ch. II). Restoring feature indices makes both pairings quadratic polynomials in the Gram entries. Hence $V_{2,2} = q(\boldsymbol{G})$ for a homogeneous quadratic polynomial q. Its polarization

$$\mathcal{B}(\boldsymbol{H}, \boldsymbol{G}) = q(\boldsymbol{H} + \boldsymbol{G}) - q(\boldsymbol{H}) - q(\boldsymbol{G})$$

is symmetric and bilinear. A unique complex-linear map $\mathcal{L}$ satisfies $\mathcal{B}(\boldsymbol{H}, \boldsymbol{G}) = \operatorname{Tr}(\boldsymbol{H}\mathcal{L}(\boldsymbol{G}))$. Symmetry gives $\mathcal{L}^\sharp = \mathcal{L}$, while homogeneity gives $q(\boldsymbol{G}) = \frac{1}{2}\operatorname{Tr}(\boldsymbol{G}\mathcal{L}(\boldsymbol{G}))$. The next permitted bidegree is $(3,3)$, so the remainder is $O(\|\boldsymbol{X}\|_{\mathrm{F}}^6)$.

Finally, under $\boldsymbol{X} \mapsto \boldsymbol{X}\boldsymbol{R}^*$, the Gram matrix becomes $\boldsymbol{R}\boldsymbol{G}\boldsymbol{R}^*$. Substitution into equation 13 and comparison of degree-two and degree-four coefficients yield equation 16. The assumption $N \geq 2$ avoids degree-two identities among Gram entries, so these coefficients are identifiable. No step classifies the full invariant algebra; higher orders may contain products of trace contractions. □

Theorem 14 (Leading Potential-Generated Transformer Field)**.** *Under Lemma 4, the leading field is given by Equations 17–19. Under Axiom 4, it has the bottleneck realization in equation 20.*

Proof. Differentiating $\operatorname{Tr}(\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{A})$ with respect to $\boldsymbol{X}^*$ gives $\boldsymbol{X}\boldsymbol{A}$. Lemma 3 gives the quartic derivative, and analyticity gives $\partial R_6/\partial \boldsymbol{X}^* = O(\|\boldsymbol{X}\|_{\mathrm{F}}^5)$. This proves equation 17.

To decompose $\mathcal{L}$, use

$$\operatorname{End}(M_C(\mathbb{C})) \simeq M_C(\mathbb{C}) \otimes M_C(\mathbb{C})^{\mathrm{op}}.$$

Equivalently, write $[\mathcal{L}(\boldsymbol{G})]_{ab} = \sum_{c,d} L_{ab,cd}G_{cd}$ and reshape the coefficient tensor as $\widetilde{L}_{(a,c),(d,b)} = L_{ab,cd}$. A rank factorization of this $C^2 \times C^2$ matrix gives

$$\mathcal{L}(\boldsymbol{G}) = \sum_{h=1}^{r} \boldsymbol{B}_h\boldsymbol{G}\boldsymbol{C}_h, \qquad r \leq C^2.$$

Here r bounds the number of left–right summands, not the ordinary matrix rank of the superoperator in a different index grouping.

The trace adjoint swaps $\boldsymbol{B}_h$ and $\boldsymbol{C}_h$ without conjugation. Since $\mathcal{L} = \mathcal{L}^\sharp$, averaging the two decompositions proves equation 18 and equation 19. Writing $\boldsymbol{x}_\omega$ for row ω of $\boldsymbol{X}$, a summand satisfies

$$[\boldsymbol{X}\boldsymbol{B}_h\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{C}_h]_{\omega:} = \sum_{\nu=1}^{N} (\boldsymbol{x}_\omega \boldsymbol{B}_h \boldsymbol{x}_\nu^*)\, \boldsymbol{x}_\nu \boldsymbol{C}_h.$$

This is unnormalized linear self-attention with score matrix $\boldsymbol{B}_h$ and value–output matrix $\boldsymbol{C}_h$. The swapped summand is its trace-adjoint partner. The converse follows from Lemma 3.

Under Axiom 4, factor $\boldsymbol{B}_h = \boldsymbol{Q}_h\boldsymbol{K}_h^*$ and $\boldsymbol{C}_h = \boldsymbol{V}_h\boldsymbol{O}_h^*$ with factors in $\mathbb{C}^{C\times C_\mathrm{A}}$. Using two heads for the paired summands yields equation 20; scalar normalizations can be absorbed into the factors. Finally, $\boldsymbol{X}\boldsymbol{A}$ acts independently on each row, while $\boldsymbol{G} = \sum_\nu \boldsymbol{x}_\nu^*\boldsymbol{x}_\nu$ permits cross-token dependence in the cubic field. Quartic order is therefore the first potential order at which this dependence can occur. □

Lemma 15 (Bottleneck Realization)**.** *Under Axiom 4, the leading field admits H pairs of width-C_A attention heads and a width-C_MLP linear feed-forward factorization.*

Proof. The attention realization follows from the rank factorizations in Axiom 4. A matrix $\boldsymbol{A}$ has a factorization $\boldsymbol{A} = \boldsymbol{W}_\mathrm{in}\boldsymbol{W}_\mathrm{out}^*$ whenever $C_\mathrm{MLP} \geq \mathrm{rank}(\boldsymbol{A})$. These are sufficient realization bounds; they do not imply an ordering between C and C_MLP. □

Lemma 16 (Differential of the row-softmax log-sum-exp potential)**.** *With $\boldsymbol{S}$, $\boldsymbol{P}$, and $V_{\boldsymbol{W}}$ as in Lemma 8,*

$$\frac{\partial V_{\boldsymbol{W}}}{\partial \boldsymbol{X}^*} = \boldsymbol{P}^\top \boldsymbol{X}\boldsymbol{W}.$$

Consequently, this potential does not generate the standard row-softmax update $\boldsymbol{P}\boldsymbol{X}\boldsymbol{W}$ in general.

Proof. For an arbitrary perturbation $\boldsymbol{H}^*$ and with $\boldsymbol{X}$ held fixed,

$$\begin{aligned} d_{\boldsymbol{X}^*} V_{\boldsymbol{W}}[\boldsymbol{H}^*] &= \sqrt{C_\mathrm{A}} \sum_{j,k} P_{jk} \left[\frac{1}{\sqrt{C_\mathrm{A}}} \boldsymbol{X}\boldsymbol{W}\boldsymbol{H}^*\right]_{jk} \\ &= \sum_{j,k,a} P_{jk} (\boldsymbol{X}\boldsymbol{W})_{ja} H^*_{ak} \\ &= \sum_{k,a} H^*_{ak} (\boldsymbol{P}^\top \boldsymbol{X}\boldsymbol{W})_{ka} \\ &= \mathrm{Tr}\left(\boldsymbol{H}^* \boldsymbol{P}^\top \boldsymbol{X}\boldsymbol{W}\right). \end{aligned}$$

Definition 2 gives the claimed derivative. For real variables, varying both occurrences of $\boldsymbol{X}$ adds the reciprocal term and yields $\nabla_{\boldsymbol{X}} V_{\boldsymbol{W}} = \boldsymbol{P}^\top \boldsymbol{X}\boldsymbol{W} + \boldsymbol{P}\boldsymbol{X}\boldsymbol{W}^\top$. Neither an even number of heads nor pairing $\boldsymbol{W}$ with $\boldsymbol{W}^*$ removes the transpose for generic scores. □

Lemma 17 (BCH scalar reweighting)**.** *Under the commutator relation $[\lambda_\mathrm{MLP}, \lambda_\mathrm{A}] = \beta\lambda_\mathrm{A}$, in a setting where the relevant series converges,*

$$\exp(\lambda_\mathrm{MLP})\exp(\lambda_\mathrm{A}) = \exp(\lambda_\mathrm{MLP} + G(\beta)\lambda_\mathrm{A}), \qquad G(\beta) = \frac{\beta}{1-e^{-\beta}}.$$

Proof. The assumption implies $\mathrm{ad}^k_{\lambda_\mathrm{MLP}}(\lambda_\mathrm{A}) = \beta^k\lambda_\mathrm{A}$, while every BCH term containing a further bracket with λ_A vanishes. Therefore the BCH series reduces to

$$\lambda_\mathrm{MLP} + \left(1 + \frac{\beta}{2} + \frac{\beta^2}{12} - \frac{\beta^4}{720} + \cdots\right)\lambda_\mathrm{A} = \lambda_\mathrm{MLP} + \frac{\beta}{1-e^{-\beta}}\lambda_\mathrm{A}.$$

The fourth-order coefficient is negative. Because $G(\beta)$ is a scalar, this identity only rescales the existing attention field and supplies no token-wise sparsification mechanism. □

C Complete Proof of the Depth–Width Refinement Theorem

Write the target potential-generated field as

$$F(t, \boldsymbol{X}) = \lambda_{\mathrm{A}}^{\boldsymbol{W}^{(t)}}(\boldsymbol{X}) + \lambda_{\mathrm{MLP}}^{\boldsymbol{W}^{(t)}}(\boldsymbol{X}) = \frac{\partial V_{\mathrm{A}}(t, \boldsymbol{X})}{\partial \boldsymbol{X}^*} + \frac{\partial V_{\mathrm{MLP}}(t, \boldsymbol{X})}{\partial \boldsymbol{X}^*}. \tag{31}$$

For $j \in \{\mathrm{A}, \mathrm{MLP}\}$, let $\mathcal{V}_{j,C_j}$ denote the branch-potential class with width C_j, where C_{A} is the attention width and C_{MLP} is the MLP width. On a control region $\mathcal{K} \subset \mathbb{C}^{N\times C}$, define the best uniform field error

$$\delta_j^*(C_j) := \sup_{t\in[0,1]} \inf_{\widehat{V}\in\mathcal{V}_{j,C_j}} \sup_{\boldsymbol{Z}\in\mathcal{K}} \left\| \frac{\partial \widehat{V}(\boldsymbol{Z})}{\partial \boldsymbol{Z}^*} - \lambda_j^{\boldsymbol{W}^{(t)}}(\boldsymbol{Z}) \right\|_{\mathrm{F}}. \tag{32}$$

This derivative-based metric is insensitive to additive constants in the potential and directly controls the induced dynamics.

Theorem (Full version of Theorem 7). Let $\boldsymbol{X} : [0,1] \to \mathbb{C}^{N\times C}$ solve $\dot{\boldsymbol{X}} = F(t, \boldsymbol{X})$, and suppose the exact trajectory, the states $\widehat{\boldsymbol{X}}_k$, and the split intermediate states $\boldsymbol{Y}_k$ below remain in $\mathcal{K}$. Assume that, for $j \in \{\mathrm{A}, \mathrm{MLP}\}$ and all $t, s \in [0,1]$ and $\boldsymbol{Z}, \boldsymbol{Z}' \in \mathcal{K}$,

$$\|\lambda_j^{\boldsymbol{W}^{(t)}}(\boldsymbol{Z})\|_{\mathrm{F}} \le B_j, \tag{33}$$

$$\|\lambda_j^{\boldsymbol{W}^{(t)}}(\boldsymbol{Z}) - \lambda_j^{\boldsymbol{W}^{(t)}}(\boldsymbol{Z}')\|_{\mathrm{F}} \le K_j \|\boldsymbol{Z} - \boldsymbol{Z}'\|_{\mathrm{F}}, \tag{34}$$

$$\|\lambda_j^{\boldsymbol{W}^{(t)}}(\boldsymbol{Z}) - \lambda_j^{\boldsymbol{W}^{(s)}}(\boldsymbol{Z})\|_{\mathrm{F}} \le \Lambda_j |t - s|. \tag{35}$$

Set $B = B_{\mathrm{A}} + B_{\mathrm{MLP}}$, $K = K_{\mathrm{A}} + K_{\mathrm{MLP}}$, and $\Lambda = \Lambda_{\mathrm{A}} + \Lambda_{\mathrm{MLP}}$. For depth T, let $h = 1/T$, $t_k = kh$, and consider an attention-then-MLP split Transformer

$$\boldsymbol{Y}_k = \widehat{\boldsymbol{X}}_k + h\widehat{\lambda}_{\mathrm{A},k}(\widehat{\boldsymbol{X}}_k), \tag{36}$$

$$\widehat{\boldsymbol{X}}_{k+1} = \boldsymbol{Y}_k + h\widehat{\lambda}_{\mathrm{MLP},k}(\boldsymbol{Y}_k), \qquad 0 \le k < T. \tag{37}$$

Suppose each learned branch is potential-generated, $\widehat{\lambda}_{j,k} = \partial \widehat{V}_{j,k}/\partial \boldsymbol{X}^*$ with $\widehat{V}_{j,k} \in \mathcal{V}_{j,C_j}$, and has uniform errors

$$\delta_{j,T} := \max_{0\le k<T} \sup_{\boldsymbol{Z}\in\mathcal{K}} \|\widehat{\lambda}_{j,k}(\boldsymbol{Z}) - \lambda_j^{\boldsymbol{W}^{(t_k)}}(\boldsymbol{Z})\|_{\mathrm{F}}, \qquad \delta_T = \delta_{\mathrm{A},T} + \delta_{\mathrm{MLP},T}. \tag{38}$$

Then, with

$$\Phi(K) = \begin{cases} (\exp(K) - 1)/K, & K > 0, \\ 1, & K = 0, \end{cases} \qquad D_T = \frac{1}{2}(\Lambda + KB) + K_{\mathrm{MLP}}(B_{\mathrm{A}} + \delta_{\mathrm{A},T}), \tag{39}$$

the finite-depth trajectory satisfies

$$\max_{0\le k\le T} \|\widehat{\boldsymbol{X}}_k - \boldsymbol{X}(t_k)\|_{\mathrm{F}} \le \exp(K)\|\widehat{\boldsymbol{X}}_0 - \boldsymbol{X}(0)\|_{\mathrm{F}} + \Phi(K)\left(\delta_T + \frac{D_T}{T}\right). \tag{40}$$

Moreover, assume the width classes are nested: $\mathcal{V}_{j,C_j} \subseteq \mathcal{V}_{j,\widetilde{C}_j}$ whenever $C_j \le \widetilde{C}_j$. Then $\delta_j^*(\widetilde{C}_j) \le \delta_j^*(C_j)$. For equal initial states and any $\eta > 0$, suppose width-C_j near-minimizers can be selected at the grid points so that their split trajectories stay in the same region $\mathcal{K}$ and the preceding regularity constants and D_T are bounded uniformly in depth and width. Then

$$\max_{0\le k\le T} \|\widehat{\boldsymbol{X}}_k - \boldsymbol{X}(t_k)\|_{\mathrm{F}} \le C_{\mathcal{K}}\left[T^{-1} + \delta_{\mathrm{A}}^*(C_{\mathrm{A}}) + \delta_{\mathrm{MLP}}^*(C_{\mathrm{MLP}}) + \eta\right], \tag{41}$$

where $C_{\mathcal{K}}$ is independent of $T, C_{\mathrm{A}}, C_{\mathrm{MLP}}$. Consequently, if the two width families are field-dense on $\mathcal{K}$ uniformly in t, then the best attainable trajectory error converges to zero as $T, C_{\mathrm{A}}, C_{\mathrm{MLP}} \to \infty$. If $\delta_j^*(C_j) = O(C_j^{-\alpha_j})$ for some $\alpha_j > 0$, the joint rate is

$$O\big(T^{-1} + C_{\mathrm{A}}^{-\alpha_{\mathrm{A}}} + C_{\mathrm{MLP}}^{-\alpha_{\mathrm{MLP}}}\big). \tag{42}$$

For the proof, abbreviate

$$F_j(t, \boldsymbol{Z}) = \lambda_j^{\boldsymbol{W}(t)}(\boldsymbol{Z}), \qquad F = F_{\mathrm{A}} + F_{\mathrm{MLP}}.$$

For one grid point t_k, define the exact parallel Euler map and the learned split map by

$$P_{k,h}(\boldsymbol{Z}) = \boldsymbol{Z} + hF_{\mathrm{A}}(t_k, \boldsymbol{Z}) + hF_{\mathrm{MLP}}(t_k, \boldsymbol{Z}), \tag{43}$$

$$\widehat{S}_{k,h}(\boldsymbol{Z}) = \boldsymbol{Z} + h\widehat{\lambda}_{\mathrm{A},k}(\boldsymbol{Z}) + h\widehat{\lambda}_{\mathrm{MLP},k}\big(\boldsymbol{Z} + h\widehat{\lambda}_{\mathrm{A},k}(\boldsymbol{Z})\big). \tag{44}$$

Lemma 18 (Local truncation of the exact effective flow)**.** *Under Equations equation 33–equation 35, for $\boldsymbol{x}_k = \boldsymbol{X}(t_k)$,*

$$\|\boldsymbol{X}(t_{k+1}) - P_{k,h}(\boldsymbol{x}_k)\|_{\mathrm{F}} \le \frac{1}{2}(\Lambda + KB)h^2. \tag{45}$$

Proof. The integral form of the ODE gives

$$\boldsymbol{X}(t_k + h) = \boldsymbol{x}_k + hF(t_k, \boldsymbol{x}_k) + \boldsymbol{r}_k,$$

where

$$\boldsymbol{r}_k = \int_0^h \big[F(t_k + \tau, \boldsymbol{X}(t_k + \tau)) - F(t_k, \boldsymbol{x}_k)\big]\, \mathrm{d}\tau.$$

The field bound implies

$$\|\boldsymbol{X}(t_k + \tau) - \boldsymbol{x}_k\|_{\mathrm{F}} \le \int_0^\tau \|F(t_k + s, \boldsymbol{X}(t_k + s))\|_{\mathrm{F}}\, \mathrm{d}s \le B\tau.$$

Adding and subtracting $F(t_k, \boldsymbol{X}(t_k + \tau))$, then applying the time and state Lipschitz bounds, yields

$$\|F(t_k + \tau, \boldsymbol{X}(t_k + \tau)) - F(t_k, \boldsymbol{x}_k)\|_{\mathrm{F}} \le \Lambda\tau + KB\tau.$$

Integration over $[0, h]$ proves Equation 45. □

Lemma 19 (Local error of the learned split block)**.** *For every $\boldsymbol{Z} \in \mathcal{K}$ whose split intermediate state also lies in $\mathcal{K}$,*

$$\|\widehat{S}_{k,h}(\boldsymbol{Z}) - P_{k,h}(\boldsymbol{Z})\|_{\mathrm{F}} \le h\delta_T + h^2 K_{\mathrm{MLP}}(B_{\mathrm{A}} + \delta_{\mathrm{A},T}). \tag{46}$$

Proof. Set $\widehat{\boldsymbol{Y}} = \boldsymbol{Z} + h\widehat{\lambda}_{\mathrm{A},k}(\boldsymbol{Z})$. Subtracting Equation 43 from Equation 44 and applying the triangle inequality gives

$$\begin{aligned}\|\widehat{S}_{k,h}(\boldsymbol{Z}) - P_{k,h}(\boldsymbol{Z})\|_{\mathrm{F}} \le\ & h\|\widehat{\lambda}_{\mathrm{A},k}(\boldsymbol{Z}) - F_{\mathrm{A}}(t_k, \boldsymbol{Z})\|_{\mathrm{F}} \\ & + h\|\widehat{\lambda}_{\mathrm{MLP},k}(\widehat{\boldsymbol{Y}}) - F_{\mathrm{MLP}}(t_k, \widehat{\boldsymbol{Y}})\|_{\mathrm{F}} \\ & + h\|F_{\mathrm{MLP}}(t_k, \widehat{\boldsymbol{Y}}) - F_{\mathrm{MLP}}(t_k, \boldsymbol{Z})\|_{\mathrm{F}}.\end{aligned}$$

The first two terms are bounded by $h\delta_{\mathrm{A},T}$ and $h\delta_{\mathrm{MLP},T}$. For the last term,

$$\|\widehat{\boldsymbol{Y}} - \boldsymbol{Z}\|_{\mathrm{F}} = h\|\widehat{\lambda}_{\mathrm{A},k}(\boldsymbol{Z})\|_{\mathrm{F}} \le h(B_{\mathrm{A}} + \delta_{\mathrm{A},T}),$$

so state Lipschitz continuity of the exact MLP field gives the remaining term in Equation 46. □

Lemma 20 (Stability of parallel Euler)**.** *For all $\boldsymbol{Z}, \boldsymbol{Z}' \in \mathcal{K}$,*

$$\|P_{k,h}(\boldsymbol{Z}) - P_{k,h}(\boldsymbol{Z}')\|_{\mathrm{F}} \le (1 + hK)\|\boldsymbol{Z} - \boldsymbol{Z}'\|_{\mathrm{F}}. \tag{47}$$

Proof. Expand the difference of the two parallel Euler maps. The identity part contributes $\|\boldsymbol{Z} - \boldsymbol{Z}'\|_{\mathrm{F}}$, and the two field differences contribute at most $h(K_{\mathrm{A}} + K_{\mathrm{MLP}})\|\boldsymbol{Z} - \boldsymbol{Z}'\|_{\mathrm{F}}$. □

Proof of the full version of Theorem 7. Let $\boldsymbol{x}_k = \boldsymbol{X}(t_k)$ and $e_k = \|\widehat{\boldsymbol{X}}_k - \boldsymbol{x}_k\|_{\mathrm{F}}$. Insert $P_{k,h}(\widehat{\boldsymbol{X}}_k)$ and $P_{k,h}(\boldsymbol{x}_k)$ between $\widehat{\boldsymbol{X}}_{k+1} = \widehat{S}_{k,h}(\widehat{\boldsymbol{X}}_k)$ and $\boldsymbol{x}_{k+1}$. Lemmas 18–20 yield

$$e_{k+1} \le (1 + hK)e_k + h\delta_T + h^2 D_T. \tag{48}$$

Iteration gives, for every $k \le T$,

$$e_k \le (1+hK)^k e_0 + (h\delta_T + h^2 D_T) \sum_{r=0}^{k-1} (1+hK)^r. \tag{49}$$

Because $kh \le 1$, $(1+hK)^k \le \exp(K)$. If $K > 0$, the geometric sum in Equation 49 is at most $(\exp(K)-1)/(hK) = \Phi(K)/h$. If $K = 0$, it is $k \le 1/h = \Phi(0)/h$. Substitution proves Equation 40.

It remains to justify the width statements. If $C_j \le \widetilde{C}_j$ and $\mathcal{V}_{j,C_j} \subseteq \mathcal{V}_{j,\widetilde{C}_j}$, then for each fixed t the infimum of the same nonnegative field-error functional over the larger set cannot increase. Taking the supremum over t preserves this inequality, hence

$$\delta_j^*(\widetilde{C}_j) \le \delta_j^*(C_j).$$

For any $\eta > 0$ and each of the finitely many grid points t_k, choose $\widehat{V}_{j,k} \in \mathcal{V}_{j,C_j}$ whose error is at most $\delta_j^*(C_j) + \eta/2$. Thus $\delta_T \le \delta_{\mathrm{A}}^*(C_{\mathrm{A}}) + \delta_{\mathrm{MLP}}^*(C_{\mathrm{MLP}}) + \eta$. Under the theorem's width-independent bounds, the same substitution controls D_T uniformly, so all fixed factors can be absorbed into $C_{\mathcal{K}}$ and Equation 41 follows.

If both width families are field-dense, the two best errors tend to zero. Taking finite near-minimizers and then $T, C_{\mathrm{A}}, C_{\mathrm{MLP}} \to \infty$ proves convergence of the best attainable trajectory error. Finally, substituting $\delta_j^*(C_j) = O(C_j^{-\alpha_j})$ into Equation 41 gives Equation 42. □

Lemma 21 (Nesting for the linearized branch classes)**.** *The quadratic MLP and paired quartic attention potential classes associated with Definition 3 and Lemma 3 are nested under increases of C_{MLP} and C_{A}, respectively, at a fixed head-pair budget.*

Proof. For the MLP term, zero-pad the columns of $\boldsymbol{W}_{\mathrm{in}}$ and $\boldsymbol{W}_{\mathrm{out}}$; their product, and therefore the potential, is unchanged. For an attention factorization, every matrix of rank at most C_{A} also has a factorization of width $\widetilde{C}_{\mathrm{A}} \ge C_{\mathrm{A}}$ by zero-padding its factors. To compensate for the normalization, replace $\boldsymbol{B}_h$ by $\sqrt{\widetilde{C}_{\mathrm{A}}/C_{\mathrm{A}}}\,\boldsymbol{B}_h$ in both occurrences of each paired term. This rescaling preserves the tied realization and the potential. Thus every narrower potential is represented by the wider class, proving the inclusions in Theorem 7. □

D Quantitative Rank Approximation and an Interaction-Order Obstruction

This section sharpens Theorem 7 for an explicitly chosen finite trace-polynomial family. Its quadratic and quartic terms fall within Lemma 4, but its higher-order terms are additional modeling choices, not consequences of that lemma. A spectral-tail bound gives a rank-dependent approximation rate, whereas omitted interaction orders create an error floor. Neither result asserts an all-order canonical representation for arbitrary analytic potentials. The factor rank r below is distinct from the left–right summand count in equation 18.

For $R > 0$, write

$$\mathcal{K}_R := \{\boldsymbol{X} \in \mathbb{C}^{N\times C} : \|\boldsymbol{X}\|_{\mathrm{F}} \le R\}.$$

For fixed integers $P, H_1, \dots, H_P$, let $\mathcal{V}_r$ be the class of potentials

$$\widehat{V}(\boldsymbol{X}) = \sum_{p=1}^{P} \sum_{h=1}^{H_p} \frac{b_{p,h}}{p} \operatorname{Tr}[(\boldsymbol{X}\boldsymbol{U}_{p,h}\boldsymbol{X}^*)^p], \qquad \operatorname{rank}(\boldsymbol{U}_{p,h}) \le r, \tag{50}$$

where $b_{p,h} \in \mathbb{R}$, $\boldsymbol{U}_{p,h} = \boldsymbol{U}_{p,h}^*$, and every $\boldsymbol{U}_{p,h}$ has an r-column two-factor representation. Thus r is also the factor width. Define

$$E_r(V_\star; R) := \inf_{\widehat{V}\in\mathcal{V}_r} \sup_{\boldsymbol{X}\in\mathcal{K}_R} \left\| \frac{\partial V_\star(\boldsymbol{X})}{\partial \boldsymbol{X}^*} - \frac{\partial \widehat{V}(\boldsymbol{X})}{\partial \boldsymbol{X}^*} \right\|_{\mathrm{F}}. \tag{51}$$

Theorem 22 (Spectral-tail rate for rank-r canonical fields)**.** *Let*

$$V_\star(\boldsymbol{X}) = \sum_{p=1}^{P} \sum_{h=1}^{H_p} \frac{a_{p,h}}{p} \operatorname{Tr}[(\boldsymbol{X}\boldsymbol{W}_{p,h}\boldsymbol{X}^*)^p], \tag{52}$$

where $a_{p,h} \in \mathbb{R}$ *and* $\boldsymbol{W}_{p,h} = \boldsymbol{W}_{p,h}^* \in \mathbb{C}^{C\times C}$*. Order the singular values of each* $\boldsymbol{W}_{p,h}$ *non-increasingly and set* $\sigma_{C+1}(\boldsymbol{W}_{p,h}) = 0$*. Then, for* $0 \le r < C$*,*

$$E_{r+1}(V_\star; R) \le E_r(V_\star; R), \tag{53}$$

$$E_r(V_\star; R) \le \sum_{p=1}^{P} \sum_{h=1}^{H_p} |a_{p,h}|\, pR^{2p-1} \|\boldsymbol{W}_{p,h}\|_2^{p-1} \sigma_{r+1}(\boldsymbol{W}_{p,h}). \tag{54}$$

In particular, if for some common $\alpha > 0$*,*

$$\sigma_k(\boldsymbol{W}_{p,h}) \le D_{p,h} k^{-\alpha} \qquad (1 \le k \le C), \tag{55}$$

then

$$E_r(V_\star; R) \le (r+1)^{-\alpha} \sum_{p=1}^{P} \sum_{h=1}^{H_p} |a_{p,h}|\, pR^{2p-1} \|\boldsymbol{W}_{p,h}\|_2^{p-1} D_{p,h} = O(r^{-\alpha}) \tag{56}$$

before finite-dimensional saturation. Moreover, $E_r(V_\star; R) = 0$ *once* $r \ge \max_{p,h} \operatorname{rank}(\boldsymbol{W}_{p,h})$*.*

Proof. The monotonicity in equation 53 follows from $\mathcal{V}_r \subseteq \mathcal{V}_{r+1}$; it does not require a unique choice of spectral truncation. To prove the quantitative bound, let $\boldsymbol{W}_{p,h}^{(r)}$ be the Hermitian spectral truncation that retains the r eigenvalues of largest absolute value, and let V_r be obtained from equation 52 by replacing every $\boldsymbol{W}_{p,h}$ by $\boldsymbol{W}_{p,h}^{(r)}$. Then $V_r \in \mathcal{V}_r$ and

$$\frac{\partial}{\partial \boldsymbol{X}^*} \left\{ \frac{a}{p} \operatorname{Tr}[(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^*)^p] \right\} = a(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^*)^{p-1}\boldsymbol{X}\boldsymbol{W}. \tag{57}$$

It therefore suffices to bound one summand. Set $\widetilde{\boldsymbol{W}} = \boldsymbol{W}^{(r)}$, $\boldsymbol{A} = \boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^*$, $\widetilde{\boldsymbol{A}} = \boldsymbol{X}\widetilde{\boldsymbol{W}}\boldsymbol{X}^*$, $M = \|\boldsymbol{W}\|_2$, and $\tau = \|\boldsymbol{W} - \widetilde{\boldsymbol{W}}\|_2 = \sigma_{r+1}(\boldsymbol{W})$. Spectral truncation gives $\|\widetilde{\boldsymbol{W}}\|_2 \le M$. Hence, on $\mathcal{K}_R$,

$$\|\boldsymbol{A}\|_2, \|\widetilde{\boldsymbol{A}}\|_2 \le R^2 M, \qquad \|\boldsymbol{A} - \widetilde{\boldsymbol{A}}\|_2 \le R^2 \tau.$$

For $p \ge 2$, the telescoping identity for matrix powers gives

$$\|\boldsymbol{A}^{p-1} - \widetilde{\boldsymbol{A}}^{p-1}\|_2 \le (p-1)(R^2 M)^{p-2} R^2 \tau.$$

Using

$$\begin{aligned}\boldsymbol{A}^{p-1}\boldsymbol{X}\boldsymbol{W} - \widetilde{\boldsymbol{A}}^{p-1}\boldsymbol{X}\widetilde{\boldsymbol{W}} &= (\boldsymbol{A}^{p-1} - \widetilde{\boldsymbol{A}}^{p-1})\boldsymbol{X}\boldsymbol{W} \\ &\quad + \widetilde{\boldsymbol{A}}^{p-1}\boldsymbol{X}(\boldsymbol{W} - \widetilde{\boldsymbol{W}})\end{aligned}$$

and $\|\boldsymbol{B}\boldsymbol{C}\boldsymbol{D}\|_{\mathrm{F}} \le \|\boldsymbol{B}\|_2 \|\boldsymbol{C}\|_{\mathrm{F}} \|\boldsymbol{D}\|_2$, we obtain

$$\|\boldsymbol{A}^{p-1}\boldsymbol{X}\boldsymbol{W} - \widetilde{\boldsymbol{A}}^{p-1}\boldsymbol{X}\widetilde{\boldsymbol{W}}\|_{\mathrm{F}} \le pR^{2p-1} M^{p-1} \tau.$$

The same estimate holds directly for $p = 1$. Multiplication by $|a_{p,h}|$, summation, and the triangle inequality prove equation 54. Substitution of equation 55 proves equation 56; spectral truncation is exact once r reaches every parameter rank. □

Interpretation. This concerns best attainable, rather than individually trained, models. A time-dependent version follows by taking the right-hand side uniformly over $t \in [0, 1]$.

The next result distinguishes factor rank from interaction order. Identify $\mathbb{C}^{N\times C}$ with $\mathbb{R}^{2NC}$ and let μ_R be normalized Lebesgue measure on the corresponding real ball $\mathcal{K}_R$. Let $\mathcal{P}_m$ denote the finite-dimensional closed subspace of $L^2(\mu_R; \mathbb{R}^{2NC})$ consisting of vector-valued real polynomials of total degree at most m, and let $\Pi_{\le m}$ be its orthogonal projection.

Theorem 23 (Error floor from omitted higher-order interactions)**.** *Fix an interaction order $q \geq 1$. Let $\mathcal{F}_{q,r}$ be any class of rank-r fields generated by real-valued canonical polynomial potentials that are finite linear combinations of Gram monomials*

$$\operatorname{Tr}\left[\prod_{\ell=1}^{p}(\boldsymbol{X}\boldsymbol{U}_\ell\boldsymbol{X}^*)\right], \qquad p \leq q, \tag{58}$$

with parameter ranks at most r. For a target field $F_\star \in L^2(\mu_R;\mathbb{R}^{2NC})$, define

$$\varepsilon_q(F_\star;R) := \|(I-\Pi_{\leq 2q-1})F_\star\|_{L^2(\mu_R)}. \tag{59}$$

Then, for every rank r,

$$\inf_{G\in\mathcal{F}_{q,r}}\sup_{\boldsymbol{X}\in\mathcal{K}_R}\|F_\star(\boldsymbol{X})-G(\boldsymbol{X})\|_{\mathrm{F}} \geq \varepsilon_q(F_\star;R). \tag{60}$$

Consequently, if $\varepsilon_q(F_\star;R) > 0$, increasing factor rank alone cannot make an order-q student approximate $F_\star$ arbitrarily well.

In particular, suppose $F_\star$ is a finite canonical polynomial field. After collecting all heads of the same degree, if its highest nonzero homogeneous field component has degree $K > 2q-1$, then $\varepsilon_q(F_\star;R) > 0$. Equivalently, if the largest non-cancelling Gram-interaction order of the target is $P_\star > q$, then $K = 2P_\star - 1$ and the positive error floor in equation 60 remains even at full parameter rank.

Proof. Each entry of $\boldsymbol{X}\boldsymbol{U}\boldsymbol{X}^*$ is quadratic in the real and imaginary parts of $\boldsymbol{X}$. A monomial of interaction order p in equation 58 therefore has total degree $2p$, and its field has degree at most $2p-1$. Thus every $G \in \mathcal{F}_{q,r}$ belongs to $\mathcal{P}_{2q-1}$, independently of r. Orthogonality gives

$$\begin{aligned}\|F_\star - G\|^2_{L^2(\mu_R)} &= \|\Pi_{\leq 2q-1}F_\star - G\|^2_{L^2(\mu_R)} \\ &\quad + \|(I-\Pi_{\leq 2q-1})F_\star\|^2_{L^2(\mu_R)} \geq \varepsilon_q(F_\star;R)^2.\end{aligned}$$

Because μ_R is a probability measure, the uniform norm on $\mathcal{K}_R$ dominates the $L^2(\mu_R)$ norm, proving equation 60.

Finally, $\mathcal{P}_{2q-1}$ is finite-dimensional and closed, and a polynomial of degree $K > 2q-1$ cannot agree almost everywhere on a ball with an element of this subspace. Its distance is therefore positive. A non-cancelling order-$P_\star$ potential produces degree $2P_\star - 1$, proving the final claim. □

Scope. This is an instantaneous-field obstruction for fixed polynomial order, not automatically a trajectory bound; it need not survive growing order, higher-degree nonlinearities, layer composition, or cancellation of the summed high-order field.

E Detailed Protocols for the Canonical Width and Euler-Depth Experiments

This appendix expands the two diagnostics in Section 4. Both use the same fixed, analytically differentiable teacher field. The width experiment fits students of different factor ranks and interaction orders; the depth experiment then freezes the fitted rank-12 students and changes only the number of Euler steps. Thus width/order affect the represented field, whereas Euler depth affects only its numerical integration.

E.1 Shared teacher and student family

All states are real matrices $\boldsymbol{X} \in \mathbb{R}^{N\times C}$ with $N = 4$ tokens and $C = 12$ features. For maximum interaction order P and factor rank r, the implemented potential is

$$V_\theta^{P,r}(\boldsymbol{X}) = \frac{\kappa}{2}\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \sum_{p=1}^{P}\sum_{h=1}^{4} s_p c_{p,h}\frac{N}{2p}\operatorname{Tr}\left(\boldsymbol{A}_{p,h}^p\right), \qquad \boldsymbol{A}_{p,h} = \frac{\boldsymbol{X}\boldsymbol{B}_{p,h}\boldsymbol{B}_{p,h}^\top\boldsymbol{X}^\top}{N}, \tag{61}$$

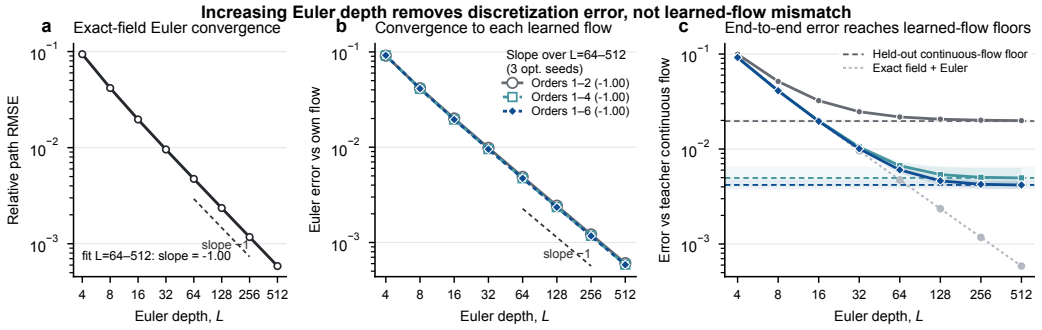

Figure 3: **Euler depth reduces discretization error but not learned-flow mismatch.** **a**, Euler integration of the exact teacher field shows first-order convergence to the RK4 reference. **b**, Frozen rank-12 students converge to their respective continuous flows as the Euler depth L increases. **c**, End-to-end student error approaches the learned-flow error floors shown by the horizontal dashed lines; the grey dotted curve is the exact-field Euler control. Curves show means and shaded regions show min–max ranges across $n = 3$ optimization seeds and 24 held-out trajectories.

where $\boldsymbol{B}_{p,h} \in \mathbb{R}^{12\times r}$ and $\kappa = 0.8$. The runs are autonomous (the number of time Fourier modes is zero). The field is evaluated by the closed-form derivative

$$F_\theta^{P,r}(\boldsymbol{X}) = -\nabla_{\boldsymbol{X}} V_\theta^{P,r}(\boldsymbol{X}) = -\kappa \boldsymbol{X} - \sum_{p=1}^{P}\sum_{h=1}^{4} s_p c_{p,h} \boldsymbol{A}_{p,h}^{p-1} \boldsymbol{X} \boldsymbol{B}_{p,h} \boldsymbol{B}_{p,h}^\top. \tag{62}$$

Consequently, training optimizes the explicit factors $\boldsymbol{B}_{p,h}$ and coefficients $c_{p,h}$; it does not fit an unrelated scalar MLP or use a second derivative. At fixed P, a rank-r model embeds exactly into every larger-rank class by appending zero columns to each factor.

The teacher is one fixed realization with $P_\star = 6$, $r_\star = 12$, four heads per order, teacher seed 1729, and positive coefficients. Each teacher factor is initialized in random orthogonal directions with singular values $j^{-0.4}$, so the eigenvalues of $\boldsymbol{B}_{p,h}\boldsymbol{B}_{p,h}^\top$ decay as $j^{-0.8}$. Coefficient amplitudes are initialized at scale 0.3 and attenuated across heads as $h^{-0.35}$. To prevent the higher powers from becoming numerically invisible, the fixed order multiplier s_p is calibrated on 2,048 annular control states. If $R_0 = 0.239365$ is the elementwise RMS of the quadratic base field, the target RMS for order p is $R_0(0.22)(0.85)^{p-1}$. The realized calibration is listed below.

Table 1: Calibration of the six teacher interaction orders. "Raw RMS" is the elementwise field RMS before multiplying by s_p on the fixed calibration sample.

Order p	Raw RMS	Target RMS	Multiplier s_p
1	0.179006	0.052660	0.294182
2	0.038382	0.044761	1.166198
3	0.015643	0.038047	2.432182
4	0.008104	0.032340	3.990849
5	0.004204	0.027489	6.539247
6	0.001741	0.023366	13.419535

E.2 Leakage-free state–field data

Training, validation, and test splits use independent random streams with seeds 2819, 2920, and 3021, respectively. In each split, initial states are sampled as $\boldsymbol{X}_0 = 0.35\boldsymbol{Z}$ with independent standard-normal entries. The teacher ODE is integrated over $[0, 1]$ by classical RK4 at 12 recorded intervals and four RK4 substeps per interval. The resulting 13 states per trajectory form the first part of the split. A second, equally sized tube sample is obtained by adding independent Gaussian noise with standard deviation 0.025 to every recorded state.

The third part probes a broader control region. A standard-normal matrix is normalized in Frobenius norm and rescaled to

$$\boldsymbol{X} = \rho\sqrt{N}\frac{\boldsymbol{Z}}{\|\boldsymbol{Z}\|_{\mathrm{F}}}, \qquad \rho \sim \mathrm{Unif}[0.5, 1.5]. \tag{63}$$

An independent $t \sim \mathrm{Unif}[0, 1]$ is stored for each control state, although the present teacher is autonomous. Every label is the exact analytic field value in Equation 62; no finite-difference labels are used. The sample accounting is

Table 2: State–field pairs in each split. Trajectory and tube columns each contain 13 states per initial condition.

Split	Initial paths	Path states	Tube states	Controls	Total
Train	96	1,248	1,248	2,048	4,544
Validation	24	312	312	512	1,136
Test	24	312	312	512	1,136

Disjoint initial conditions and random streams prevent a trajectory or its perturbed copies from crossing split boundaries. The same fixed dataset is used for all ranks, orders, and optimization seeds.

E.3 Width/order sweep: optimization and evaluation

Students retain orders $P \in \{2, 4, 6\}$ and use factor ranks $r \in \{1, 2, 4, 8, 12\}$. Three optimization seeds (0, 1, and 2) are run for every pair, giving 45 fitted checkpoints. The quadratic coefficient κ and the calibrated order scales s_p are fixed to the teacher values; all student factors and head coefficients are trainable. At the smallest rank, factor entries use scale $1/\sqrt{12}$ and coefficient entries use standard deviation 0.08. Student coefficients are unconstrained, so every student remains a gradient field but is not necessarily dissipative.

Ranks are optimized in ascending order separately for each (P, seed) pair. The preceding checkpoint is embedded into the wider model, while newly available columns receive a small random perturbation of scale 0.05. This perturbation is an optimization device: exact zero padding certifies nesting, but the derivative of $\boldsymbol{B}\boldsymbol{B}^{\top}$ with respect to a newly zero column would also be zero.

For a minibatch $\mathcal{B}$, the optimized objective is

$$\mathcal{L}_{\mathrm{field}}(\theta) = \frac{|\mathcal{B}|^{-1}\sum_{i\in\mathcal{B}}\|F_\theta^{P,r}(\boldsymbol{X}_i) - F_\star(\boldsymbol{X}_i)\|_{\mathrm{F}}^2}{|\mathcal{D}_{\mathrm{train}}|^{-1}\sum_{i\in\mathcal{D}_{\mathrm{train}}}\|F_\star(\boldsymbol{X}_i)\|_{\mathrm{F}}^2}. \tag{64}$$

Optimization uses AdamW, learning rate 2×10^{-3}, weight decay 10^{-6}, batch size 256, and gradient-norm clipping at 10. Training lasts at most 140 epochs. Validation relative RMS is evaluated every two epochs; the best checkpoint is retained, and training stops after 18 validation evaluations without an improvement greater than 10^{-6}. Evaluation uses batches of 512. The epoch-0 initialization is eligible for selection, no learning-rate scheduler is used, and test metrics are evaluated only after restoring the best-validation checkpoint. Width training uses PyTorch 2.0.0+cu118 in float32 on CUDA. All 45 runs, including early-stopped runs, enter the reported summary; no run or seed is excluded.

For a held-out collection $\mathcal{S}$, field error is the ratio-of-sums quantity

$$\mathcal{E}_{\mathrm{field}}(\mathcal{S}) = \left(\frac{\sum_{i\in\mathcal{S}}\|F_\theta^{P,r}(\boldsymbol{X}_i) - F_\star(\boldsymbol{X}_i)\|_{\mathrm{F}}^2}{\sum_{i\in\mathcal{S}}\|F_\star(\boldsymbol{X}_i)\|_{\mathrm{F}}^2}\right)^{1/2}. \tag{65}$$

Figure 2a evaluates this metric on all 1,136 test states. Figure 2c starts from the 24 held-out test initial conditions and compares student and teacher explicit-Euler rollouts on the same 32-step grid, including all 33 time nodes. Both summaries show the mean and one sample standard deviation across the three optimization seeds; the seed, rather than an individual state or trajectory, is the replicate unit.

The open diamond in Figure 2b is a parameter-privileged realizability control. For every teacher head, it keeps the leading r singular triplets of the teacher factor $\boldsymbol{B}_{p,h}$ and copies the teacher coefficients.

At matched order and full rank it has test relative field RMS 2.02×10^{-6} (floating-point residual). Because this construction accesses the teacher parameters, it is neither a trained baseline nor a lower bound on the error attainable by optimization.

E.4 Euler-depth sweep with frozen fields

The depth sweep uses the nine rank-12 checkpoints: three maximum orders and three optimization seeds. No parameter is retrained. For each frozen field, explicit Euler is applied with shared parameters at

$$L \in \{4, 8, 16, 32, 64, 128, 256, 512\}, \qquad h = L^{-1},$$

on $[0, 1]$, always from the same 24 held-out initial states. Hence increasing L adds repeated evaluations of one field but no trainable parameters.

Teacher and student continuous-flow references are computed in float64 with classical RK4 and 2,048 uniform steps. A reference-convergence audit compares 1,024-step and 2,048-step RK4 trajectories: their relative path RMS is 4.61×10^{-13} for the teacher and at most 4.30×10^{-13} over the nine students, below the prespecified 10^{-8} tolerance. All whole-path comparisons use the same 257-point uniform time grid. An Euler path with a coarser native grid is mapped to this grid by piecewise-linear interpolation.

For paths Y and Z sampled at the common times t_m, define

$$d(Y, Z) = \left(\frac{\sum_{i=1}^{24} \sum_{m=0}^{256} \|Y_i(t_m) - Z_i(t_m)\|_{\mathrm{F}}^2}{\sum_{i=1}^{24} \sum_{m=0}^{256} \|Z_i(t_m)\|_{\mathrm{F}}^2} \right)^{1/2}. \tag{66}$$

With $X_\star^{\mathrm{cont}}$ denoting the fine-RK4 teacher path, $\widehat{X}_{P,12}^{\mathrm{cont}}$ the fine-RK4 path of a frozen student, and $\widehat{X}_{P,12}^{L}$ its Euler path, the three reported errors are

$$\begin{aligned} E_{\mathrm{total}}(L) &= d(\widehat{X}_{P,12}^{L}, X_\star^{\mathrm{cont}}), & E_{\mathrm{num}}(L) &= d(\widehat{X}_{P,12}^{L}, \widehat{X}_{P,12}^{\mathrm{cont}}), \\ E_{\mathrm{model}} &= d(\widehat{X}_{P,12}^{\mathrm{cont}}, X_\star^{\mathrm{cont}}). \end{aligned} \tag{67}$$

The exact-field control replaces the student Euler path by a teacher Euler path. We also record student Euler versus teacher Euler on the identical L-step grid. This latter quantity removes a solver mismatch, but it is a same-integrator discrepancy rather than an isolated field-error term: nonlinear interaction between field mismatch and discretization remains.

Figure 3a pools the deterministic exact-teacher error over the 24 paths. Figures 3b–c first compute each path metric for a checkpoint and then show the mean and min–max range over the three optimization seeds within each order family. The 24 paired trajectories are not treated as 24 independent optimization replicates. Straight lines are least-squares fits of $\log E_{\mathrm{num}}$ against $\log L$ over $L \in \{64, 128, 256, 512\}$. The fitted slopes are -1.0043 for the exact teacher and -1.0031, -1.0038, and -1.0040 for the order-2, order-4, and order-6 students. The corresponding mean continuous-flow discrepancies E_{model} are 0.01970, 0.004959, and 0.004189.

The three quantities in Equation 67 are norms with different reference paths and do not form an additive scalar decomposition; in particular, $E_{\mathrm{total}} - E_{\mathrm{model}}$ is not interpreted as numerical error. No hypothesis test is used in either sweep. The conclusions are conditional on one constructed teacher, one fixed data split, the stated optimization budget, and the tested ranks and depths. Factor rank r is the width of the canonical factors, P is interaction order, and L is shared-weight numerical unrolling depth; these three axes should not be conflated. Finally, the held-out metrics are distributional diagnostics, not estimators of the uniform best-approximation errors in Theorem 7.

F Experimental Protocols and Additional Results

F.1 Fixed-16 training and frozen-field refinement

Model and training. The depth experiment uses CIFAR-10 with 45,000/5,000/10,000 train/validation/test images, 65 tokens (64 image patches and one classification token), feature dimension $C = 96$, three attention heads, MLP hidden dimension 192, and patch size 4. The

Table 3: Frozen-field depth scan. Errors are relative RMS values against the RK4-512 reference. Centered logits remove each example's common logit offset before comparison. Accuracy is reported to emphasize that decision accuracy need not be monotone even when state and logit errors are.

L	Terminal state	Full path	Centered logits	Accuracy (%)
1	0.314477	0.263414	0.722406	30.86
2	0.202163	0.159437	0.455847	48.30
4	0.113611	0.081698	0.265135	54.49
8	0.067113	0.044509	0.163826	55.63
16	0.037638	0.023521	0.091658	55.96
32	0.020093	0.012136	0.048605	55.90
64	0.010401	0.006171	0.025042	55.74
128	0.005293	0.003112	0.012712	55.66
RK4-512	0	0	0	55.66

learned vector field is potential-generated and explicitly time-dependent. Three bounded Fourier gates modulate its quadratic, attention, and local/MLP potential components; four harmonics per gate give 24 trainable time coefficients. Spatial field parameters, gates, and the classifier are trained jointly for 120 epochs with one training seed. The forward pass always uses 16 explicit-Euler steps during training and validation, and the checkpoint is selected by validation accuracy at this same depth.

The implementation uses the ordinary real-gradient field $F_\theta = -\nabla_{\boldsymbol{X}} V_\theta$. This follows the real-state convention stated after Definition 2. The depth-refinement argument applies directly to this field; it requires neither an action principle nor a complex-gradient sign reinterpretation.

Frozen evaluation and reference audit. After selection, all parameters and inputs are frozen and the same field is integrated with

$$L \in \{1, 2, 4, 8, 16, 32, 64, 128\}, \qquad h = L^{-1}.$$

Here L is the notation retained in the archived plots and equals the depth T used in the theory. The numerical reference is FP64 RK4 with 512 steps. Its terminal difference from RK4-1024 is 2.20×10^{-11}, and the reference-refinement error is 1.66×10^{-4} of the Euler-128 terminal error, well below the prespecified 1% tolerance.

Log–log fits on $L \in \{16, 32, 64, 128\}$ give empirical orders 0.94399, 0.97292, and 0.95068 for terminal state, full path, and centered logits, with $R^2 = 0.999729$, 0.999942, and 0.999776. Prediction KL has an empirical order near two because the first variation of KL vanishes at matching normalized probabilities; this does not make the Euler state solver second order. Figure 4 displays the propagation of state refinement to the classifier.

The learned gate profiles in Figure 5 are nonconstant, confirming that the queried field is observably non-autonomous rather than merely written with a formal time argument. Because these profiles are learned jointly with the spatial field, they describe this checkpoint's schedule but do not establish a causal advantage over an independently retrained autonomous model. A separate random-depth checkpoint obtains 56.28% accuracy at Euler-16, compared with 55.96% here; the paired test-prediction McNemar value is $p = 0.452$. Thus fixed-16 training is not claimed to improve classification, and the depth result concerns only numerical refinement after freezing.

F.2 Analytic and ViT-6 width controls

Metric. For held-out state–field pairs $(t_i, \boldsymbol{X}_i, F_i)$ and a fixed orthogonal projection P, the real-data experiment reports the distributional proxy

$$\widehat{\delta}_P^{\mathrm{RMS}} = \left(\frac{\sum_i \|P\widehat{F}(t_i, \boldsymbol{X}_i) - PF_i\|_{\mathrm{F}}^2}{\sum_i \|PF_i\|_{\mathrm{F}}^2} \right)^{1/2}. \tag{68}$$

It is invariant to the number of repeated samples only through empirical averaging and probes the clean trajectory distribution. It is neither an unbiased estimator nor an upper bound for the uniform error in Equation equation 32 without additional coverage and concentration assumptions.

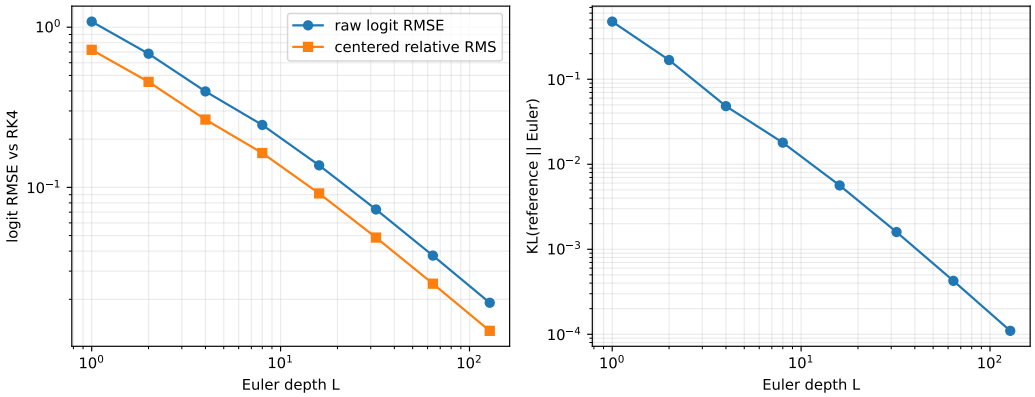

Figure 4: **Propagation of depth refinement to the classifier.** Raw- and centered-logit discrepancies are measured against the same RK4-512 reference used in Table 3. Centered-logit relative RMS has empirical order 0.951 on the four finest grids. The near-quadratic decay of prediction KL reflects its local dependence on small logit perturbations and is not a second-order state-solver claim.

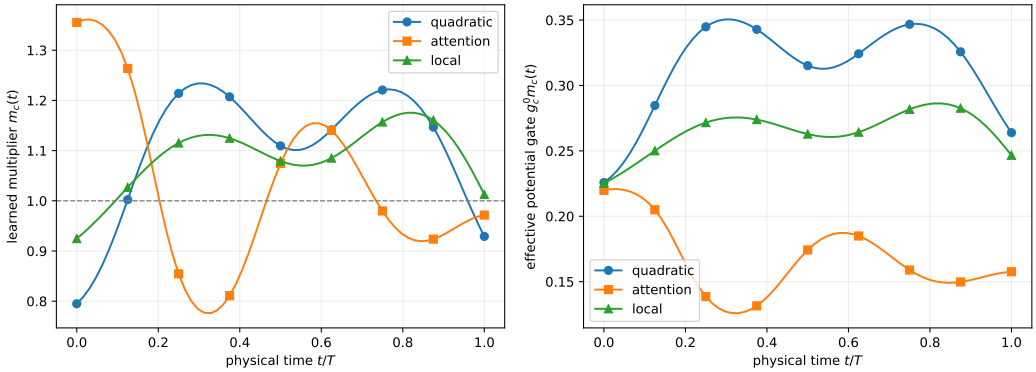

Figure 5: **Learned non-autonomous modulation.** Bounded Fourier functions modulate the quadratic, attention, and local/MLP potential components. The top panel gives dimensionless multipliers and the bottom panel the resulting effective coefficients. Their nonconstant profiles show that the learned field varies with normalized layer time; they are descriptive of one training seed and are not an ablation against a retrained autonomous field.

Analytic conservative control. The analytic target is a 16-dimensional non-autonomous conservative field generated by a potential containing 128 time-dependent log-cosh ridges and a fixed quadratic term. Three-layer scalar-potential networks with Softplus$(\beta = 5)$ use widths $16, 32, 64, 128, 256, 512$ and five paired optimization seeds. Mean test relative RMS values are

$$0.203961,\ 0.132054,\ 0.107554,\ 0.099210,\ 0.084789,\ 0.076808.$$

All five seeds improve from the smallest to the largest width, with mean endpoint reduction 62.34%. Four seeds decrease at every adjacent width; one has a small reversal from 64 to 128. The fitted empirical exponent is 0.25944 ± 0.01061 (one seed SD). These are finite-range empirical scalings, not a claimed approximation-theoretic optimum.

Frozen standard-ViT target. The teacher is a standard CIFAR-10 ViT-6 with six blocks, feature dimension 128, four heads, patch size 4, MLP hidden dimension 256, and 79.25% test accuracy. It is not constrained to be energy-compatible. Each block is exactly decomposed into its attention and subsequent MLP residual update, producing 12 substeps over a unit horizon. One global, unwhitened rank-64 PCA basis is fitted from 6,144 substep states obtained from 512 training images; the ambient state dimension is 8,320. Surrogate image counts are 4,000/1,024/2,048 for train/validation/test.

Table 4: Held-out projected relative RMS for the frozen ViT-6 field. Values are mean ± one training-seed SD ($n = 3$). The capacity tier is a paired parameter-budget index; the vector model's actual hidden width is given separately.

Tier q	Potential width	Vector width	Potential	Vector
32	32	24	0.6792 ± 0.0003	0.6838 ± 0.0027
64	64	53	0.6246 ± 0.0011	0.6017 ± 0.0028
128	128	115	0.5653 ± 0.0035	0.5250 ± 0.0045
256	256	242	0.5064 ± 0.0023	0.4567 ± 0.0015
512	512	497	0.4640 ± 0.0021	0.4091 ± 0.0012

At nominal capacity tiers $q \in \{32, 64, 128, 256, 512\}$, scalar-potential models use hidden widths equal to q, while unconstrained vector MLPs use widths $24, 53, 115, 242, 497$ to match parameter counts within 0.698%. Both families use three-layer Softplus($\beta = 5$) networks, 50 epochs, batch size 1,024, AdamW with learning rate 2×10^{-4} and weight decay 10^{-4}, and paired seeds 0, 1, and 2.

The potential and vector families improve by $31.68\% \pm 0.29\%$ and $40.17\% \pm 0.08\%$ from the first to last paired tier. All six seed-by-family curves decrease at all four transitions. This directional consistency supports the within-protocol width trend. It does not establish a model-family difference: with three paired seeds the smallest possible two-sided exact sign-flip value is $p = 0.25$, and all five Holm-adjusted values equal one.

The fixed PCA-64 basis captures 42.29% of state energy but only 22.78% of held-out field energy. If outputs are restricted to this subspace, the corresponding empirical lifted-error floor is $\sqrt{1 - 0.2278} = 0.8787$; this is conditional on the chosen basis and is not an irreducible floor for higher-rank or full-space models. Finally, all 30 validation-selected surrogate checkpoints occur at epoch 50, the final epoch, so optimization budget remains a confounder. The experiment consequently supports only the following bounded statement: at fixed target, data, projection, and training protocol, increasing surrogate capacity reduces the observed held-out projected field-error proxy.

G Potential-Compatibility Assessment of a Standard Transformer

Which symmetry is audited. This appendix makes precise the empirical meaning of "symmetry breaking" used above. The measured symmetry is transpose symmetry of the full real Jacobian, which is required locally by a Euclidean scalar-potential field. It is not the $U(N)$ token-isometry invariance in Axiom 2, the feature-basis covariance in Axiom 3, or the trace-self-adjointness condition on the restricted quartic Gram operator in Lemma 3. These structures are conceptually related through potential generation but are different mathematical statements.

Frozen teacher and residual secant fields. The teacher is a standard pre-LN ViT trained from scratch on CIFAR-10 using 45,000/5,000/10,000 train/validation/test images. Images are divided into 4×4 patches, giving 64 patch tokens and one classification token. The model has six independently parameterized blocks, feature dimension 128, four attention heads, MLP hidden dimension 256, and 811,146 trainable parameters. It was trained for 100 epochs with batch size 256, AdamW, peak learning rate 3×10^{-4}, weight decay 0.05, five warm-up epochs, and label smoothing 0.1. The validation-selected epoch-99 checkpoint obtains 80.84% validation accuracy, 79.25% test accuracy, and test negative log-likelihood 0.7120. All subsequent measurements freeze this checkpoint.

Let B_ℓ denote block ℓ and let $\boldsymbol{X}_\ell \in \mathbb{R}^{65 \times 128}$ be its input state. Then

$$\boldsymbol{X}_{\ell+1} = B_\ell(\boldsymbol{X}_\ell), \qquad F_\ell(\boldsymbol{X}) = \frac{B_\ell(\boldsymbol{X}) - \boldsymbol{X}}{h} = 6\big[B_\ell(\boldsymbol{X}) - \boldsymbol{X}\big], \quad h = \frac{1}{6}. \tag{69}$$

The scaling interprets one block displacement as a velocity over a unit depth horizon. It does not turn the finite block map into a known infinitesimal generator. Moreover, $B_0, \ldots, B_5$ do not share parameters, so Equation 69 defines six discrete non-autonomous residual secant fields rather than one autonomous field evaluated at six times.

Jacobian symmetry as a local potential condition. Flattening the token state gives $d = 65 \times 128 = 8320$ real coordinates. If, on a neighborhood, a field has the Euclidean energy form $F = -\nabla E$ with $E \in C^2$, then

$$J_F = DF = -\nabla^2 E = J_F^\top.$$

For each block and sampled state, decompose

$$J_\ell = S_\ell + A_\ell, \qquad S_\ell = \frac{J_\ell + J_\ell^\top}{2}, \qquad A_\ell = \frac{J_\ell - J_\ell^\top}{2}. \tag{70}$$

The two terms are Frobenius-orthogonal, so $\|J_\ell\|_F^2 = \|S_\ell\|_F^2 + \|A_\ell\|_F^2$. A nonzero A_ℓ is therefore a local certificate that the measured field is not an exact Euclidean gradient field in the current token coordinates. The converse is only local and requires the usual smoothness and domain conditions; the diagnostic is not a global Helmholtz or Hodge decomposition.

Matrix-free JVP/VJP estimator. An explicit Jacobian has about 6.92×10^7 entries per image and block. Instead, automatic differentiation computes a Jacobian–vector product $J_\ell v$ and a vector–Jacobian product $J_\ell^\top v$. Their sum and difference give $(J_\ell + J_\ell^\top)v$ and $(J_\ell - J_\ell^\top)v$ without materializing J_ℓ. For isotropic Rademacher vectors, $\mathbb{E}\|Mv\|_2^2 = \|M\|_F^2$, motivating

$$\rho_{\mathrm{skew},\ell} = \left(\frac{\sum_v \|(J_\ell - J_\ell^\top)v\|_2^2}{\sum_v \left[\|(J_\ell - J_\ell^\top)v\|_2^2 + \|(J_\ell + J_\ell^\top)v\|_2^2\right]} \right)^{1/2}. \tag{71}$$

Under the ideal isotropic-probe identity, $\rho_{\mathrm{skew},\ell} \approx \|A_\ell\|_F / \|J_\ell\|_F$, so $\rho_{\mathrm{skew},\ell}^2$ and $1 - \rho_{\mathrm{skew},\ell}^2$ estimate the skew and symmetric Frobenius-energy fractions, respectively.

Probe directions, uncertainty, and controls. The ambient audit uses a fixed random subset of 128 test images and eight float64 Rademacher probes for every image and block. We also evaluate data-chord directions $v = \boldsymbol{X}_\ell^{(a)} - \boldsymbol{X}_\ell^{(b)}$ formed from pairs of real hidden states. Ambient probes estimate a full-coordinate Frobenius ratio, whereas data chords provide only a directional diagnostic along data-related displacements. We resample images as independent clusters for 2,000 bootstrap replicates and report percentile 95% intervals. As implementation controls, an exactly symmetric linear map yields $\rho_{\mathrm{skew}} = 1.53 \times 10^{-16}$, while an exactly skew-symmetric linear map yields $\rho_{\mathrm{skew}} = 1$.

Layerwise results. Table 5 gives the exact layerwise values. Ambient skew ratios lie between 0.6328 and 0.6863, corresponding to 40.04%–47.10% skew Jacobian energy. Symmetric energy remains slightly larger in every block and averages 56.02%, but no block approaches the exact-potential condition $\rho_{\mathrm{skew}} = 0$. The layerwise values are not monotone with depth.

Interpretation and performance boundary. The result supports a bounded statement: in the current Euclidean token coordinates, this frozen seed-0 standard ViT-6 contains both a substantial Jacobian-symmetric sector satisfying the local necessary condition for potential compatibility and a substantial skew sector that violates it. Its 79.25% test accuracy shows that task-effective representations can coexist with this local violation. It does not show that the skew component improves

Table 5: Matrix-free Jacobian audit of the frozen standard ViT-6. Ambient intervals are 95% image-cluster bootstrap intervals. Skew and symmetric energy fractions are derived from the ambient ratio. Data-chord ratios are directional diagnostics and should not be interpreted as ambient Frobenius fractions.

Block	Ambient ρ_{skew}	95% CI	Skew (%)	Symmetric (%)	Data chord ρ
0	0.6608	[0.6603, 0.6613]	43.67	56.33	0.6166
1	0.6509	[0.6500, 0.6517]	42.37	57.63	0.6292
2	0.6863	[0.6858, 0.6868]	47.10	52.90	0.6865
3	0.6726	[0.6718, 0.6734]	45.24	54.76	0.6788
4	0.6745	[0.6738, 0.6752]	45.49	54.51	0.6671
5	0.6328	[0.6318, 0.6337]	40.04	59.96	0.6486
Mean	0.6630	–	43.98	56.02	–

accuracy. Establishing such a causal benefit would require a matched intervention that suppresses Jacobian skew while controlling architecture, parameter count, optimization, and random seed. Nor can 56.02% be called "the percentage of the Transformer explained by a potential": the statistic concerns local derivatives of finite residual secant fields, not a global field-value projection. Finally, the audit neither rules out a different coordinate system or state-dependent Riemannian metric nor proves that the finite block maps share a smooth continuous generator.

H Learning Three-Dimensional Particle Dynamics from Trajectories

We tested whether reciprocal parameter tying improves trajectory-only identification of a three-dimensional particle field. We compared a symmetry-tied pair-interaction network (Tied model), an endpoint-untied pair-interaction network (Untied model), and a vector-output multilayer perceptron (MLP), using identical position observations under zero noise and coordinate noise of standard deviation 10^{-3}. An integral ridge estimator in the tied model's function class provided an additional non-neural reference. The experiment concerned learning an unknown field, not accelerating the integration of a known equation.

Trajectory-based inference of distance-dependent interactions has been studied nonparametrically (Lu et al., 2019), and potential-based particle configurations arise in distance-based formation control (Krick et al., 2009). Transformers have also been used to learn stochastic dynamics from active-matter trajectories (Casert et al., 2024). Here we considered a deterministic, structure-matched gradient flow to examine the effect of reciprocal tying.

H.1 System and observation protocol

The system contained $M = 128$ labelled particles with configuration $X = [x_1, \dots, x_M] \in \mathbb{R}^{3\times M}$. For each candidate pair $e = (i, j)$ in $\mathcal{E} = \{(i, j) : i < j\}$, define

$$c_e = \mathbf{e}_i - \mathbf{e}_j, \quad P_e = c_e c_e^\top, \quad r_e(X) = Xc_e, \quad s_e(X) = \|r_e(X)\|_2^2. \tag{72}$$

The teacher was a fixed overdamped nonlinear elastic network,

$$\dot{X} = F_*(X) = -\nabla_X E_*(X), \qquad E_*(X) = \frac{1}{4M}\sum_{e\in\mathcal{E}} \kappa_e \big(s_e(X) - \ell_e^2\big)^2. \tag{73}$$

Rest lengths ℓ_e came from a unit reference circle embedded in $\mathbb{R}^3$. Each particle was connected to its 21 nearest neighbours on either side in the cyclic ordering, giving 2,688 active pairs. Active stiffnesses were sampled once from $[0.6, 1.4)$; inactive pairs had $\kappa_e = 0$. The pair models considered all $K = 8,128$ candidate pairs without access to the active mask, stiffnesses or rest lengths.

We used $N = 128$ training trajectories, 32 validation trajectories, and 64 trajectories in each of the in-distribution (ID) and out-of-distribution (OOD) test sets. Initial states were generated by deforming the reference ring, adding independent three-coordinate perturbations, and applying random three-dimensional rotations and translations. Thus the initial clouds were non-planar even though the reference geometry was planar. OOD states used larger deformations, perturbations and translations; the governing equation was unchanged. These were synthetic relaxation trajectories, not observations of a calibrated three-dimensional material.

Training and validation observations covered $[0, 2]$ at spacing $\Delta t = 0.04$. We compared $Y(t_k) = X^*(t_k) + \sigma Z_k$, with independent standard Gaussian entries in Z_k and $\sigma \in \{0, 10^{-3}\}$. Noise was added to both training and validation positions, not to the dynamics. Its units were the reference radius, rather than a fraction of particle displacement. Test initial conditions and reference trajectories remained clean. The two noise conditions shared the teacher and clean trajectory splits. At $\sigma = 10^{-3}$, the expected coordinate RMS of endpoint noise differences was 1.243 times the clean training-window displacement RMS.

H.2 THREE EXPLICIT VECTOR-FIELD ARCHITECTURES

Tied model. The tied network learned two coefficients (a_e, β_e) per pair:

$$F_\theta^{\text{tied}}(X) = -\frac{1}{M}\sum_e (a_e + \beta_e s_e) r_e c_e^\top \tag{74}$$

$$= -\frac{1}{M}\sum_e \left(a_e X P_e + \beta_e X P_e X^\top X P_e\right). \tag{75}$$

The same coefficient acted with opposite signs on the two endpoints, so $F_\theta^{\text{tied}} = -\nabla_X E_\theta$ for

$$E_\theta(X) = \frac{1}{M}\sum_e \left(\frac{a_e}{2} s_e + \frac{\beta_e}{4} s_e^2\right). \tag{76}$$

This construction enforces $O(3)$ equivariance, translation invariance and zero total velocity, $F_\theta^{\text{tied}}(X)\mathbf{1} = 0$. The teacher belongs to this class through $a_e^* = -\kappa_e \ell_e^2$ and $\beta_e^* = \kappa_e$.

Its connection to the paper's linearized Transformer is algebraically explicit. For a no-softmax head $\mathcal{H}(X; A, B) = XAX^\top XB$, where $A = W_Q W_K^\top$ and $B = W_V W_O^\top$, set $A_e = P_e$ and $B_e = -\beta_e P_e/(2M)$. Then

$$\begin{aligned} F_\theta^{\text{tied}}(X) &= XL_\theta + \sum_e \left[\mathcal{H}(X; A_e, B_e) + \mathcal{H}(X; B_e, A_e)\right], \\ L_\theta &= -\frac{1}{M}\sum_e a_e P_e. \end{aligned} \tag{77}$$

Each pair admits two width-one heads with exchanged query–key and value–output products. The implementation evaluates the equivalent pair formula directly; it does not materialize dense attention heads or differentiate a generic scalar-output MLP. Pair directions are fixed and only a_e, β_e are trained. In this matrix convention the particle axis is the feature axis, not a conventional particle-token sequence. The linear term is the linear feed-forward branch; there is no additional nonlinear feed-forward block, softmax or LayerNorm.

Untied model. The untied network instead learned independent coefficients $(a_e^{(i)}, \beta_e^{(i)})$ and $(a_e^{(j)}, \beta_e^{(j)})$ at the two endpoints of each pair, and evaluated

$$\begin{aligned} F_\theta^{\text{untied}}(X) = \frac{1}{M}\sum_{e=(i,j)} r_e \big[&- (a_e^{(i)} + \beta_e^{(i)} s_e)\mathbf{e}_i^\top \\ &+ (a_e^{(j)} + \beta_e^{(j)} s_e)\mathbf{e}_j^\top\big]. \end{aligned} \tag{78}$$

Geometric equivariance and translation invariance are retained, but reciprocal interactions and a potential-gradient field are not enforced. The tied network is recovered by equating the two endpoint coefficients. This is an endpoint-sharing ablation within a restricted pair-polynomial class, not a standard untied softmax Transformer.

MLP. With $z = \text{flat}(X) \in \mathbb{R}^{384}$, the MLP was

$$\begin{aligned} F_\theta^{\text{MLP}}(X) &= \text{unflat}\left[W_3\phi(W_2\phi(W_1 z + b_1) + b_2) + b_3\right], \\ \phi(u) &= \frac{u}{1 + e^{-u}}. \end{aligned} \tag{79}$$

It used the architecture $384 \to 16 \to 16 \to 384$: three affine layers, two SiLU activations and a linear vector output. No centering, equivariant layers, reciprocal interactions or force-balance projection were imposed. Table 6 gives the model sizes; the comparison is not parameter matched.

H.3 TRAJECTORY FITTING AND EVALUATION

Each trajectory supplied 46 overlapping windows, each containing an initial observation and $H = 5$ future observations. Starting at the first observation, we integrated the learned autonomous field with

Table 6: Model classes for 128 particles in three dimensions. Counts exclude fixed geometry buffers. Integral ridge uses the same pair-polynomial class as the Tied model, with a different estimator.

Model	Parameters	Structural constraint
Tied model	16,256	$O(3)$ equivariance and reciprocal tying
Untied model	32,512	$O(3)$ equivariance; independent endpoints
MLP	12,960	None of the above
Integral ridge	16,256	Same field class as Tied model

classical fourth-order Runge–Kutta (RK4), without resetting intermediate states:

$$\mathcal{L}(\theta) = \frac{1}{S_{\text{train}} BH(3M)} \sum_{q=1}^{B} \sum_{j=1}^{H} \|\widehat{X}_{q,j}(\theta) - Y_{q,j}\|_F^2. \tag{80}$$

Here $B = 128$ windows were sampled with replacement, and S_{train} was the mean squared observed displacement from window initial states, pooled over all windows and target coordinates and floored at 10^{-10}. RK4 used two substeps per observation interval, with gradients through all 40 shared-parameter field calls per window. An RK4 step is therefore not a single Transformer residual layer.

All neural models received 1,200 Adam updates, gradient-norm clipping at 10, and learning rates of 0.02 for the pair models and 0.003 for the MLP. Pair coefficients were initialized at $a = -0.02$, $\beta = 0.02$, identically at both untied endpoints. Checkpoints were selected by observed validation rollout loss after the first update and every 100 updates. No true field or teacher coefficient entered fitting or selection. Seeds 0, 1 and 2 gave 18 neural fits across the two noise conditions.

Integral ridge fitted the same pair-polynomial class using trapezoidal features of observed positions. Each window spanned five observation intervals, and windows did not overlap. The regularization candidates were 0, 10^{-6}, 10^{-4}, 10^{-2} and 1, selected by the same observed validation rollout criterion. Fits used column-scaled, matrix-free conjugate gradients in float64, with 2,000 iterations and normal-residual tolerance 10^{-7}. These six additional fits used no ground-truth field labels either.

Evaluation used the fixed first 32 ID and 32 OOD trajectories. Models were integrated from clean initial states to $t = 6$ using eight RK4 substeps per observation interval. Training used float32; teacher references and evaluation used float64. Reference trajectories used 16 substeps, checked against 32 on two ID and two OOD trajectories. We measured the relative field error on 512 true ID states and the displacement-normalized trajectory error:

$$\epsilon_F = \left(\frac{\sum_q \|F_\theta(X_q) - F_*(X_q)\|_F^2}{\sum_q \|F_*(X_q)\|_F^2} \right)^{1/2}, \tag{81}$$

$$\epsilon_{\text{disp}} = \left(\frac{\sum_{q,k} \|\widehat{X}_q(t_k) - X_q^*(t_k)\|_F^2}{\sum_{q,k} \|X_q^*(t_k) - X_q^*(0)\|_F^2} \right)^{1/2}. \tag{82}$$

Both long-horizon ID and OOD trajectory errors covered $[0, 6]$; only $[0, 2]$ was observed during training. Means and sample standard deviations describe three fitting seeds on one fixed teacher, not three independent physical systems. At $N = 128$, seeds share all clean training paths; they vary row order, optimization and, when present, measurement noise. Test paths are shared throughout.

H.4 Clean and noisy prediction results

The Tied model had lower mean field and trajectory errors than the Untied model and MLP in both noise conditions (Table 7). With clean observations, its long-horizon ID error was $6.67\% \pm 0.18\%$, compared with $10.08\% \pm 0.22\%$ for the Untied model and $107.12\% \pm 1.94\%$ for the MLP. At $\sigma = 10^{-3}$, the corresponding errors were $17.49\% \pm 0.74\%$, $25.96\% \pm 0.72\%$, and $107.15\% \pm 1.64\%$. The Tied model's OOD error changed from $4.51\% \pm 0.16\%$ to $8.43\% \pm 0.15\%$; the smaller normalized OOD error does not imply smaller absolute errors because OOD displacements have a different scale.

Table 7: Prediction with $M = N = 128$ and three spatial coordinates. Entries are mean $\pm$ sample SD across three seeds, in percent; lower is better. Both trajectory columns use true-displacement normalization on $[0, 6]$. The dagger marks a selected iterative ridge solve that did not reach the prescribed residual tolerance.

σ	Model	Field $\downarrow$	ID long $\downarrow$	OOD $\downarrow$
0	Tied model	8.009 ± 0.229	6.667 ± 0.178	4.514 ± 0.158
0	Untied model	12.314 ± 0.242	10.078 ± 0.224	6.569 ± 0.071
0	MLP	108.737 ± 2.820	107.123 ± 1.936	99.919 ± 0.442
0	Integral ridge†	0.0104 ± 0.0040	0.0091 ± 0.0035	0.0053 ± 0.0021
10^{-3}	Tied model	21.112 ± 0.588	17.495 ± 0.743	8.432 ± 0.152
10^{-3}	Untied model	30.247 ± 0.514	25.960 ± 0.721	10.241 ± 1.104
10^{-3}	MLP	107.923 ± 1.849	107.145 ± 1.643	100.322 ± 0.334
10^{-3}	Integral ridge†	6.220 ± 0.042	5.354 ± 0.047	3.635 ± 0.046

The fixed examples in Figures 6 and 7 show the same ordering: both pair models tracked the displacement patterns, whereas the MLP predicted much less motion than the reference. Figure 8 shows error accumulation through the training-duration interval and beyond it. All panels used seed 0 and the preselected first ID/OOD trajectory, retaining all 128 particles and 151 recorded times. The plotted single-example errors are not the three-seed means in Table 7. Full three-coordinate errors were computed before projection; eight-versus-sixteen-substep refinement changed these example paths by at most 1.74×10^{-13} relative to the true displacement norm.

Integral ridge achieved still lower errors at this training size. Its mean field error was 0.0104% without noise and 6.220% with noise, and its long-horizon ID errors were 0.0091% and 5.354%, respectively. All 24 evaluations in the present $N = 128$ comparison were finite. However, all six selected ridge solves exhausted the iteration budget without meeting the residual tolerance; these are the measured predictions of budget-limited iterative estimates, not certified solutions of the ridge objective. Thus the experiment does not establish superiority of trajectory-trained networks over regression.

Overall, reciprocal tying improved the tested neural models under clean and noisy observations, within a family that already contains the teacher. The benefit is conditional on the stated architecture and optimization budget: the models are not parameter matched, and 1,200 updates do not establish optimizer convergence. Unconstrained learned coefficients also mean that a potential-gradient structure alone does not guarantee globally bounded trajectories or numerical stability. These results support a structured field-identification use case, not a comparison against general softmax Transformers, a depth/rank convergence experiment, or a claim of better ODE integration.

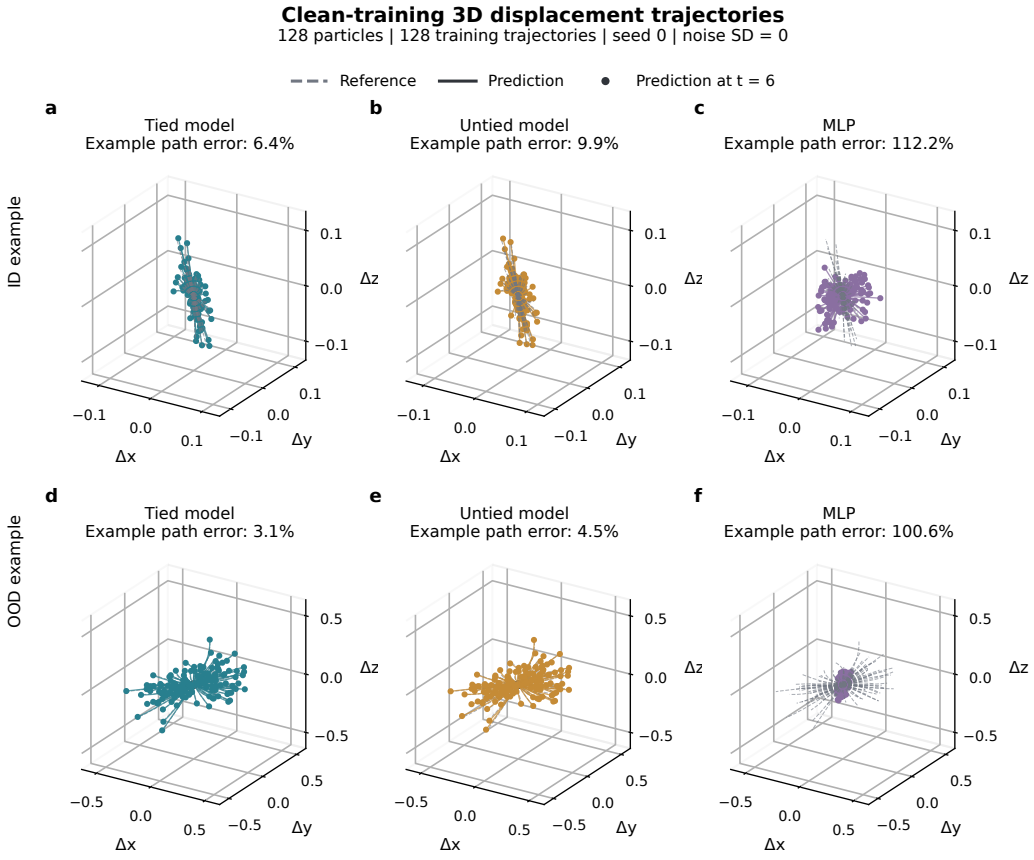

Figure 6: **Trajectories learned from clean observations.** Columns compare Tied model, Untied model and MLP; panels a–c show ID and d–f show OOD. Curves are individual displacements $x_i(t) - x_i(0)$ for all 128 particles over $[0, 6]$, not their spatial arrangement. Gray dashed curves are references, coloured curves are predictions, and dots mark predicted endpoints. Axis scales and camera are shared across models and noise conditions within each split. Annotations report displacement-normalized errors for these fixed seed-0, index-0 examples, not aggregate errors.

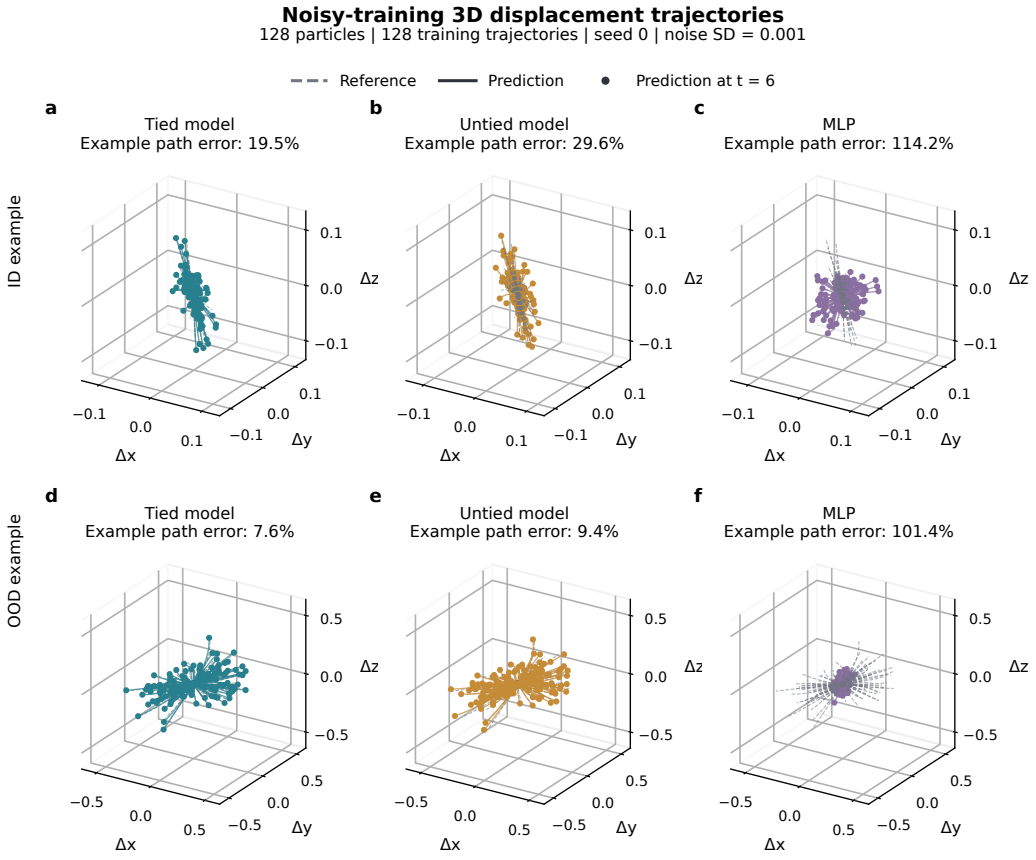

Figure 7: **Three-dimensional trajectories learned with coordinate noise $\sigma = 10^{-3}$.** Layout, initial states, reference paths, scaling and fixed-example selection match Figure 6. Noise affects training and validation observations only. Panels a–c show ID and d–f show OOD; all coordinates and times are retained without smoothing or displacement magnification. Both pair models follow the reference motion more closely than the tested MLP in these examples.

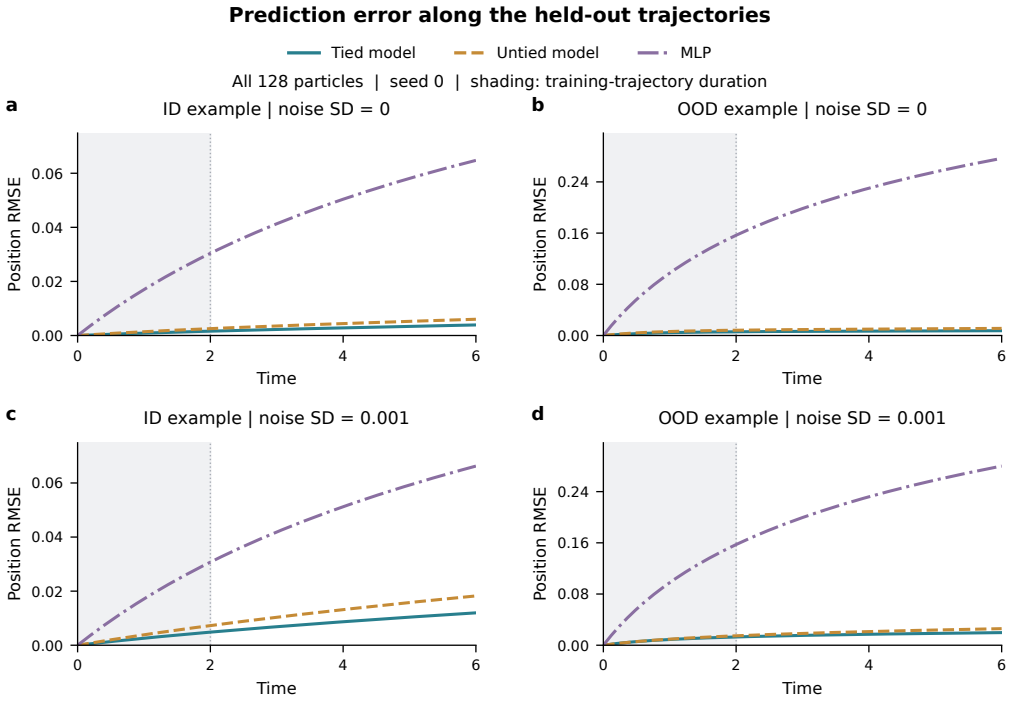

Figure 8: **Time-resolved errors of the fixed held-out examples.** Panels a,b use clean training data and c,d use $\sigma = 10^{-3}$; columns show ID and OOD. Curves report $[M^{-1} \sum_i \|\widehat{x}_i(t) - x_i^*(t)\|_2^2]^{1/2}$ in coordinate units, using all three coordinates. Shading marks the duration of training trajectories, $[0, 2]$; the illustrated paths remain held out throughout. These are single-seed, single-trajectory examples, so no across-seed uncertainty bands are implied. Line styles distinguish models independently of colour.