INHERENT TRADE-OFFS WHEN ENFORCING CORRECT DATA SYMMETRIES IN MULTI-HEAD SELF-ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We give a complete characterization of equivariant multi-head self-attention (MHSA): if an MHSA layer is equivariant to a symmetry group G, then G can only act by permuting head-clusters, with QK and OV matrices satisfying an equivariance constraint tied to the group action. As a consequence, we prove that any fixed MHSA architecture that achieves exact equivariance by polynomially parameterizing unconstrained MHSA parameters inevitably leads to expressivity loss within the class of equivariant maps: the equivariance locus of unconstrained MHSA forms a union of extremely many Zariski-irreducible components in a reduced parameter space, and any single architecture covers at most one. For $G = D_4$ acting on C copies of the regular representation as the token feature space, we show that there are $\Omega(C^{64})$ components for eight attention heads. We illustrate this with experiments on artificial tasks and show that the predicted head-cluster structures can arise naturally when the attention layer is pressured to become equivariant.

1 INTRODUCTION

Since symmetry is present in nature, it is present in data and therefore in many machine learning tasks. Usually, given a task we even know a priori certain symmetries that it respects. For instance, the semantic segmentation of an aerial or medical image does not depend on orientation and the ground state energy of a molecule does not depend on the spatial reference frame used to encode each atom's position. The property of respecting a given symmetry is often formalized as equivariance, which a model may possess either by design at the architectural level or by learning.

It may seem like enforcing a correct symmetry at the architectural level is an open goal, but this is not necessarily the case. First of all, it may not be clear how a given equivariance can be enforced. For common symmetry groups, such as permutations and rotations of 2D and 3D space, much effort has been devoted to designing equivariant architectures. In the 1990's, researchers started building equivariant MLPs (Shawe-Taylor, 1989; Wood & Shawe-Taylor, 1996), followed by equivariant CNNs in the 2010's (Cohen & Welling, 2016; Weiler et al., 2018; Kondor & Trivedi, 2018; Bekkers et al., 2018; Cohen et al., 2020) and a shift towards equivariant attention and transformer architectures in the 2020's (Romero & Cordonnier, 2021; Xu et al., 2023; Islam et al., 2025).

Second, even when a procedure is known for making an architecture equivariant, it may not yield all possible equivariant models available to an unconstrained baseline. For instance, the standard way to make an MLP equivariant is to enforce an equivariance constraint on each layer, but for ReLU-MLPs there exist special parameter configurations that give equivariant networks without each layer being equivariant, and these special configurations can parameterize different functions than layerwise equivariant networks of the same width (Agrawal & Ostrowski, 2023). So in ReLU-MLPs, enforcing a correct symmetry via any layerwise equivariance construction misses out on equivariant functions available to an unconstrained baseline.

Finally, even if a simple procedure is available that yields all possible equivariant models, it can happen that one ends up with a forced choice among mutually exclusive options. This is the case for rank-bounded linear maps, where any equivariant architecture only covers one of many irreducible components of equivariant rank-bounded linear maps (Kohn et al., 2025).

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

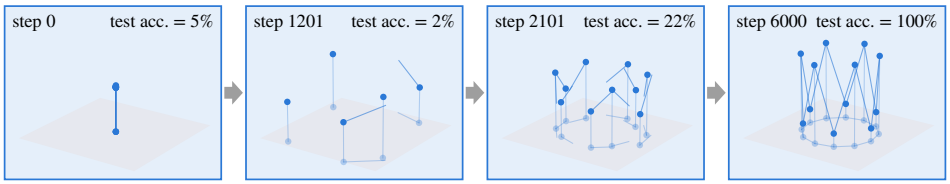

Figure 1: QK matrices form a "crown", training on modular addition under equivariance pressure.

In this paper we give a precise account of these three potential conundrums in the case of equivariant multi-head self-attention (MHSA). First, we prove a structure theorem. Assuming that a symmetry group G acts on input and output token feature spaces $\mathbb{R}^{D_{\text{in}}}, \mathbb{R}^{D_{\text{out}}}$, we give a complete characterization of maps $\mathbb{R}^{L\times D_{\text{in}}} \to \mathbb{R}^{L\times D_{\text{out}}}$ that are both G-equivariant and expressible as a multi-head self-attention. Slightly simplified, every instance of equivariant MHSA is given by a partition of the attention heads into orbits under G and a joint equivariance constraint on the heads in an orbit. Every equivariant MHSA is implementable using relatively simple weight sharing rules and equivariance constraints, so finding equivariant architectures is not an issue. In particular, there are no "unexpected" equivariant parameterizations, in contrast to the ReLU-MLP case.

Based on the structure result, we use algebraic geometry to show that the space of equivariant functions is very intricate: it is a union of extremely many "components", and any given architecture (way of constraining equivariance) can cover at most one component. Hence, an MHSA with architectural equivariance never represents all equivariant functions representable by the unconstrained MHSA. Our second main result gives a combinatorial lower bound on the number of equivariant models needed to cover all equivariant components of a single unconstrained MHSA of the same size. In general, as the feature dimension grows, the bound explodes.

Our results therefore show that MHSA is analogous to the rank-bounded linear map case. There is also a further analogy with the linear maps. All rank-k equivariant linear maps can be expressed by an equivariant linear map that is not rank constrained, which corresponds to a single "wider" model. In a similar sense, we will see that all components of equivariant MHSA with H heads can be expressed by a single model with $|G|H$ heads, via a group averaging construction. The situation is illustrated in Figure 2. The different components of the equivariance locus of a particular MHSA stem from a single equivariant component in a larger parameter space, but in the original parameter space we need to make a choice among mutually exclusive options.

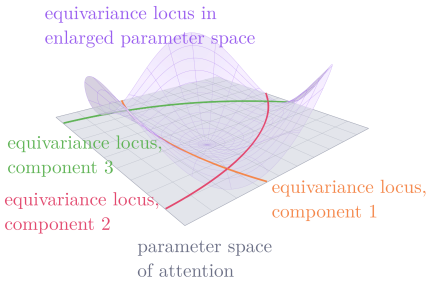

Figure 2: The equivariance dilemma.

Interestingly, our results together present simultaneously a case *for* hard-coded equivariance in attention and a case *against*:

For (corollary of Theorem 1): For an MHSA layer to be equivariant, its weights must be highly structured. There is only a finite list of constraints on the weights, each corresponding to one component, to choose from, and any one of them results in a valid equivariant attention layer that is easy to implement. However, to our knowledge, only two extreme cases in this family have been implemented so far.

Against (corollary of Theorem 2): Since only one component can be chosen, any architecture that implements equivariant attention inevitably encodes *more inductive bias than just equivariance*, relative to the unconstrained baseline with the same number of heads and the same head dimension.

To illustrate our theoretical results, we conduct two interpretability experiments: a modular addition task and a vision transformer task. We find that in these tasks approximate equivariance is encoded structurally in MHSA, consistent with our theory for exact equivariance. For instance, in Figure 1 we show the emergence during training of a "crown"-shaped orbit of size 12, consisting of QK matrices of different heads in MHSA trained on addition modulo 24.

2 PRELIMINARIES: ATTENTION AND EQUIVARIANCE

Multi-head self-attention. Let $D_{\text{in}}, D_{\text{out}}$ be the dimensions of the input and output token feature spaces, d the head dimension, and H the number of heads. Usually $D_{\text{in}} = D_{\text{out}} = dH$.

For weights $W_q^{(h)}, W_k^{(h)}, W_v^{(h)} \in \mathbb{R}^{d \times D_{\text{in}}}, W_o^{(h)} \in \mathbb{R}^{D_{\text{out}} \times d}$, where $1 \le h \le H$, the multi-head self-attention on L input tokens $x = (x_1, x_2, \ldots, x_L) \in \mathbb{R}^{L \times D_{\text{in}}}$ is given by[1]

$$\text{MHSA}(x; W_q, W_k, W_v, W_o)_i = \sum_{h=1}^{H} W_o^{(h)} \frac{\sum_{j=1}^{L} \exp\left((W_q^{(h)} x_i)^\top (W_k^{(h)} x_j)\right) W_v^{(h)} x_j}{\sum_{j=1}^{L} \exp\left((W_q^{(h)} x_i)^\top (W_k^{(h)} x_j)\right)}. \quad (1)$$

Eq. (1) depends only on $M_h := W_q^{(h)\top} W_k^{(h)}$ and $R_h := W_o^{(h)} W_v^{(h)}$, the QK and OV matrices in the terminology of Elhage et al. (2021).

Equivariance. Symmetry is commonly present in both real-world tasks as well as artificial tasks used in interpretability experiments. The mathematical formalization of "model respecting symmetry" is equivariance, which we motivate with two concrete examples in the following, both in the context of attention/transformers.

Example 1 (Modular addition (Nanda et al., 2023))**.** *Let p be a positive integer. Suppose we want to learn modular addition, i.e., the function $(a, b) \mapsto a + b \pmod p$. The inputs (a, b) are one-hot encoded as (e_a, e_b). For example, $e_2 = (0, 0, 1, 0)^\top$ for $p = 4$. The model T is supposed to output p logits, the argmax of which should be $a + b \pmod p$. The set $C_p := \{0, 1, \ldots, p-1\}$ is an abelian group under modular addition. It acts on $\mathbb{R}^p$ by permuting the basis vectors, which are just the one-hot encodings: $\rho_{\text{in}}(g) e_a = e_{g+a}$ where $g \in C_p$. For example, if $p = 3$ and $g = 1$, then $\rho_{\text{in}}(g)$ is the matrix $\left(\begin{smallmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{smallmatrix}\right)$.*

Modular addition satisfies $(g + a) + (g + b) \equiv 2g + a + b \pmod p$, so the model should satisfy the following corresponding condition:

$$T(\rho_{\text{in}}(g) e_a, \rho_{\text{in}}(g) e_b) = \rho_{\text{out}}(g) T(e_a, e_b), \quad \text{with } \rho_{\text{out}}(g) = \rho_{\text{in}}(g)^2. \quad (2)$$

Example 2 (Semantic segmentation)**.** *An input image $u \in \mathbb{R}^{C \times N \times N}$ is divided into L square patches, viewed as L tokens $(x_1, \ldots, x_L) \in \mathbb{R}^{L \times CP^2}$, where $P := N/\sqrt{L}$ is the number of pixels on one side of each patch. The dihedral group D_4 (rotations by 90° and reflections) acts on the image u, rotating each patch and permuting the patches simultaneously. For each $g \in D_4$, let $\rho_{\text{in}}(g) \in \mathbb{R}^{CP^2 \times CP^2}$ be the permutation matrix that rotates or reflects the $P \times P$ pixels without touching the channel index. The model's task is to output a pixel-level segmentation mask $m \in \mathbb{R}^{L \times DP^2}$, where $m_i \in \mathbb{R}^{DP^2}$ is the $P \times P$ array of D class logits of pixels in patch i.*

We expect "rotate image → segment" to yield the same result as "segment → rotate mask" for data with rotation symmetry. In other words, if $(m'_1, \ldots, m'_L) = T(\rho_{\text{in}}(g) x_{g^{-1}.1}, \ldots, \rho_{\text{in}}(g) x_{g^{-1}.L})$ is the output of the model on the rotated image, we expect $m'_i = \rho_{\text{out}}(g) m_{g^{-1}.i}$, where $\rho_{\text{out}}(g) \in \mathbb{R}^{DP^2 \times DP^2}$ rotates the mask at patch i. Because attention (and transformers in general) is blind to the order of the tokens, this condition is the same as

$$T(\rho_{\text{in}}(g) x_1, \ldots, \rho_{\text{in}}(g) x_L)_i = \rho_{\text{out}}(g) T(x_1, \ldots, x_L)_i. \quad (3)$$

In each of the preceding examples, we concluded that it is natural for the model T to satisfy $T(\rho_{\text{in}}(g) x_1, \ldots, \rho_{\text{in}}(g) x_L)_i = \rho_{\text{out}}(g) T(x_1, \ldots x_L)_i$, with different ρ_{in} and ρ_{out} in each case. If this holds, we say that T is *equivariant at token i*. For next-token prediction, this is the natural condition with $i = L$. For semantic segmentation, one expects equivariance at all tokens.

3 THE STRUCTURE THEOREM FOR EQUIVARIANT MHSA

In this section, we present Theorem 1, which is the first of our two main results. It completely characterizes the set of parameters, i.e., tuples $W = (W_q, W_k, W_v, W_o)$ where $W_q, W_k, W_v \in$

[1]The conventional temperature scaling in the QK matrices M_h can be absorbed by scaling W_q and W_k and does not affect our theorem statements and proofs, so we simply leave it out in the definition of attention.

$\mathbb{R}^{H\times d\times D_{\text{in}}}$ and $W_o \in \mathbb{R}^{H\times d\times D_{\text{out}}}$, for which $x \mapsto \text{MHSA}(x; W)$ is G-equivariant with respect to the token-wise G-action. That is,

$$\text{MHSA}\big(\rho_{\text{in}}(g)x_1, \ldots, \rho_{\text{in}}(g)x_L; W\big)_i = \rho_{\text{out}}(g)\text{MHSA}\big(x_1, \ldots, x_L; W\big)_i \quad \forall g \in G. \tag{4}$$

Example 3. *Inspecting Eq.* (1)*, we see that if $W_q^{(h)}, W_k^{(h)}, W_v^{(h)}, W_o^{(h)}$ all commute with G for each head h, then $\text{MHSA}(-; W_q, W_k, W_v, W_o)$ is G-equivariant. However, according to the following theorem, this is not the only possibility.*

Theorem 1 (Structure of equivariant MHSA)**.** *Let G be a group acting on $\mathbb{R}^{D_{\text{in}}}$ and $\mathbb{R}^{D_{\text{out}}}$ orthogonally*[2]*. Assume the number of tokens L is at least three. Let W_q, W_k, W_v, W_o be the weights of an H-head self-attention. Let $M_h := W_q^{(h)\top} W_k^{(h)} \in \mathbb{R}^{D_{\text{in}}\times D_{\text{in}}}$ and $R_h := W_v^{\top(h)} W_o^{(h)} \in \mathbb{R}^{D_{\text{out}}\times D_{\text{in}}}$ be the QK and OV matrices of head h. For any matrix $m \in \mathbb{R}^{D_{\text{in}}\times D_{\text{in}}}$, define $R_m := \sum_{h:M_h=m} R_h \in \mathbb{R}^{D_{\text{out}}\times D_{\text{in}}}$. This is the total OV matrix corresponding to one QK matrix. If no M_h is equal to m, then $R_m = 0$ by definition.*

Let $T := \text{MHSA}(-; W_q, W_k, W_v, W_o) : \mathbb{R}^{L\times D_{\text{in}}} \to \mathbb{R}^{L\times D_{\text{out}}}$ denote the function represented by the multi-head self-attention layer with the given weights.

The following are equivalent:

(1) T is equivariant at one output token.
(2) T is equivariant at all output tokens.
(3) $R_{\rho_{\text{in}}(g)m\rho_{\text{in}}(g)^{-1}} = \rho_{\text{out}}(g)R_m\rho_{\text{in}}(g)^{-1}$ for all $m \in \mathbb{R}^{D_{\text{in}}\times D_{\text{in}}}$ and $g \in G$.

(1)/(2) $\Rightarrow$ (3) is the non-obvious direction. The proof is found in Appendix B.

For example, suppose that all heads have distinct QK matrices, i.e., $M_h \neq M_{h'}$ for $h \neq h'$, and suppose, additionally, that no head's OV matrix vanishes, i.e., $R_h \neq 0$ for all h. Then Theorem 1 says that T is G-equivariant if and only if for each QK matrix M_h and any group element $g \in G$, there is some head h' ($h' = h$ *is* allowed) such that $M_{h'} = \rho_{\text{in}}(g)M_h\rho_{\text{in}}(g)^{-1}$ and $R_{h'} = \rho_{\text{out}}(g)R_h\rho_{\text{in}}(g)^{-1}$.

In general, the QK matrices of different heads $h \neq h'$ might coincide ($M_h = M_{h'}$). If a set of heads $S \subset \{1, \cdots, H\}$ share one QK matrix M_S, and $M_h \neq M_S$ for all $h \notin S$, we say S forms a *head cluster*. Thus, the set of all heads $\{1, \cdots, H\}$ is partitioned into disjoint head clusters. For each cluster S, the total OV matrix is $R_S := R_{M_S} = \sum_{h\in S} R_h$ (note that the symbol R is now triply overloaded: the subscript can be a head, a $D_{\text{in}} \times D_{\text{in}}$ matrix, or a head cluster). If $R_S = 0$, we say the cluster S is *dead*.

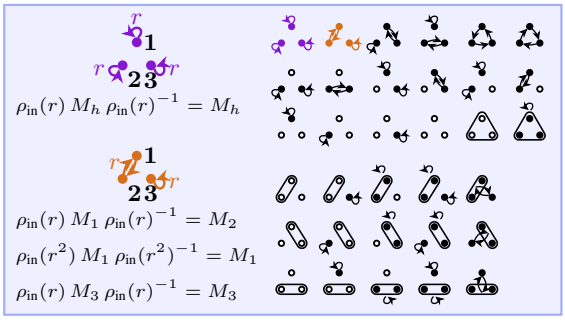

Figure 3: All 33 partial C_6-clusterings of three heads and the equivariance constraints corresponding to the first two. Empty dots are dead heads.

We can then rephrase Theorem 1 in the following way: there is a G-action σ on the non-dead head clusters such that $M_{\sigma(g)S} = \rho_{\text{in}}(g)M_S\rho_{\text{in}}(g)^{-1}$ and $R_{\sigma(g)S} = \rho_{\text{out}}(g)R_S\rho_{\text{in}}(g)^{-1}$. We will call such an arrangement (head clusters, which clusters are dead, and a G-action on the non-dead clusters) a *partial G-clustering*. We will often use X, X_0 to denote the sets of non-dead and dead clusters respectively ($X := \{C : R_C \neq 0\}, X_0 := \{C : R_C = 0\}$). Note that X and X_0 are sets of sets of heads.

Example 4. *The partial G-clustering where all clusters consist of exactly one head, no cluster is dead, and the G-action is trivial, is called the* ***trivial (G-)clustering****.*

The weaker version of Theorem 1 when the QK matrices M_h are pairwise distinct can be obtained by combining the generic identifiability result of Henry (2026) with the framework of

[2]That is, $\rho_{\text{in}}(g)^\top = \rho_{\text{in}}(g)^{-1}$ and $\rho_{\text{in}}(g)\rho_{\text{in}}(g') = \rho_{\text{in}}(gg')$. Same for ρ_{out}.

Shahverdi et al. (2026), which builds on Marchetti et al. (2024). Theorem 1 is stronger in that it is a global statement: it also describes nongeneric parameter configurations in which several heads share the same QK matrix. In this case, the individual effective OV matrices R_h are not identifiable; only their sums over the corresponding fibers are intrinsic, as noted by Cirrincione (2026), and required to respect equivariance. The global form is crucial to the proof of Theorem 2: it gives a Zariski-closed cover of the equivariance locus (see Section C), and having control of every member of the cover is necessary for proving maximality of the counted irreducible components.

Example 5. *Let $G = C_2 = \{e, r\}$ with two attention heads $H = \{1, 2\}$. If $M_1 \neq M_2$ and $X_0 = \varnothing$, then χ is bijective and G acts on H. There are two possibilities: either (a) the action is trivial or (b) $\sigma(r)(1) = 2$ and $\sigma(r)(2) = 1$. In case (a), $\rho_{\text{out}}(r)R_h\rho_{\text{in}}(r)^{-1} = R_h$ and $\rho_{\text{in}}(r)M_h\rho_{\text{in}}(r)^{-1}$ commute with $\rho(r)$ for both $h \in H$. In case (b), R_1 and R_2 are related by $\rho_{\text{out}}(r)R_1\rho_{\text{in}}(r)^{-1} = R_2$, and similarly $\rho_{\text{in}}(r)M_1\rho_{\text{in}}(r)^{-1} = M_2$.*

Example 6. *For $H = 1$, if the OV matrix is not zero, the equivariance conditions simply become that both M and R are G-equivariant, recovering Theorem 1 of Ông & Bökman (2026).*

Example 7. *Consider an arbitrary finite group G. Take $X = I \times G$, where I is some finite index set. Let G act on X by left multiplication on the second factor, i.e., $\sigma(g)(j, g') = (j, gg')$. In this case, the equivariance condition reads $gM_{j,g'}g^{-1} = M_{j,gg'}$ and similarly for the $R_{j,g}$. In particular, this means $M_{j,g} = gM_{j,e}g^{-1}$ and $R_{j,g} = gR_{j,e}g^{-1}$, with $M_{j,e}$ and $R_{j,e}$ unconstrained. This is the construction used by the Platonic Transformer (Islam et al., 2025).*

For a Lie group G, the connected component G_0 of the identity must act trivially on the non-dead head clusters. That is, we equivalently have a $\mathfrak{C}$-action, where $\mathfrak{C} := G/G_0$ is the group of connected components. Therefore, if G is connected, each head must be individually equivariant.

4 A Lower Bound on the Number of Equivariant Models

We now discuss our second main result, built on Theorem 1, which is, roughly speaking: "*An unconstrained MHSA can represent many more G-equivariant functions than a polynomially parameterized MHSA with exact G-equivariance.*" We will quantify this lack of expressivity by counting how many polynomially parameterized models are needed in order to represent all G-equivariant MHSAs. This motivates the following definition:

Definition 1. *Let $\Lambda = \mathbb{R}^P$ be some parameter space, and let $\beta : \Lambda \to \mathbb{R}^{H\times d\times D_{\text{in}}} \times \mathbb{R}^{H\times d\times D_{\text{in}}} \times \mathbb{R}^{H\times d\times D_{\text{in}}} \times \mathbb{R}^{H\times D_{\text{out}}\times d}$ be a polynomial parameterization map. We define the (β-)**polynomially parameterized MHSA** to be the parameterized map*

$$\begin{aligned} \beta\text{-MHSA} : (\mathbb{R}^{L\times D_{\text{in}}}) \times \Lambda &\to (\mathbb{R}^{L\times D_{\text{out}}}) \\ (x, \lambda) &\mapsto \text{MHSA}(x; \beta(\lambda)). \end{aligned} \tag{5}$$

*Given G-representations ρ_{in}, ρ_{out} on $\mathbb{R}^{D_{\text{in}}}$ and $\mathbb{R}^{D_{\text{out}}}$ respectively, a family $\mathcal{M} = (\beta\text{-MHSA})_{\beta\in\mathbf{B}}$ is G-**complete** if each β-MHSA $\in \mathcal{M}$ realizes only G-equivariant maps and any G-equivariant map expressible by MHSA with H d-dimensional heads can be expressed by β-MHSA for some $\beta \in \mathbf{B}$.*

For a G-representation $\rho : G \to \mathbb{R}^{D\times D}$, let $c_{G,d}(\rho)$ denote the number of maximal (with respect to containment) subrepresentations $V \subset \mathbb{R}^D$ such that $0 < \dim V \leq d$ up to equivalence. For example, if $G = C_4$ acts on $\mathbb{R}^4$ by one copy of each of A_1, A_2, and E_1, there are two maximal subrepresentations with $0 < \dim V \leq 2$: $\mathrm{A}_1 \oplus \mathrm{A}_2$ and E_1, so $c_{C_4,2}(\mathrm{A}_1 \oplus \mathrm{A}_2 \oplus \mathrm{E}_1) = 2$.

Theorem 2. *Let G be finite. Assume $D_{\text{in}} = D_{\text{out}}$ and $\rho_{\text{in}} = \rho_{\text{out}}$ (that is, the G-actions on input and output token features coincide). A G-complete family of polynomially parameterized MHSA with H heads and head dimension d necessarily contains at least $\frac{1}{H!}c_{G,d}(\rho)^{2H}$ elements (models).*

The proof (Appendix C) involves understanding the Zariski-irreducible components of the equivariance locus in a suitable parameter space (that of RMHSA, see Appendix B.1), which we consider to be part of the neuroalgebraic geometry program (Marchetti et al., 2025). These components have the same algebro-geometric origin as the ones in equivariant linear networks (Kohn et al., 2025). The lower bound is obtained by counting only components corresponding to the trivial action on the heads (trivial G-clusterings), and including nontrivial actions could only increase the bound.

Example 8. *Let G be the trivial group acting on $\mathbb{R}^D$, $C > 0$. With $d = \dim V/|H| = C/|H|$, we have $c_{G,d}(V) = 1$, and the lower bound is $1/H!$, which is vacuous. Indeed, for a trivial group, any function is equivariant, so the equivariance locus is the entire parameter space, which can be easily covered by one parameterization.*

Example 9. *Take $G = D_4$ acting on $\mathbb{R}^{8C}$ via C copies of the regular representation ρ_{reg}. With eight attention heads of dimension C , the bound is $8!^{-1}c_{D_4,C}(\rho_{\text{reg}}^{\oplus C})^{16}$. For $C = 96$ (total token dimension $= 8C = 768$), this number is about 1.6×10^{96} (derivation in Appendix C.5). That is roughly enough atoms for each ant on Earth to get its own copy of the observable universe.*

How different are these components? For $G = D_4$ acting on $\mathbb{R}^8$ via the regular representation with $H = 2$ heads and head dimension $d = 4$, we take six of these components, giving six possible D_4-equivariant multi-head self-attention layers. These are indexed by (1) a D_4-action on the two heads, and (2) for each D_4-orbit, a pair (one for OV and one for QK) of four-dimensional subrepresentations of $\mathbb{R}^8$ restricted to the stabilizer. More intrinsically, each is labeled by a pair of D_4-equivariant vector bundles over the same base space, see Appendix D. We consider here two D_4 actions with three subrepresentation choices for each, as shown in Fig. 4. We randomly initialize each of these six equivariant models as the teacher and train all six models as well as the unconstrained MHSA against it with MSE loss. The results are shown in Fig. 4's heatmap: the unconstrained MHSA can learn to be equivariant in all six ways, but each equivariant model is only able to learn anything when the teacher is in the same component, save for two exceptions when the student has all irreps the teacher has: for example, with D_4/C_4, the $2B \oplus E_1$ model has a relatively low loss against the $2E_1$ teacher.

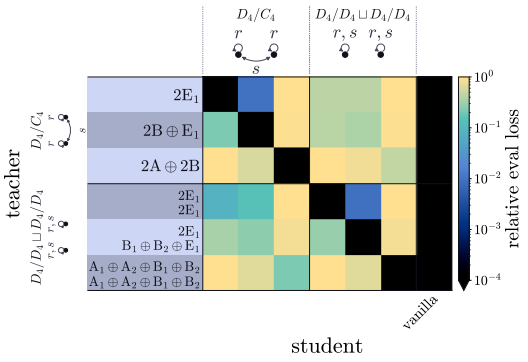

Figure 4: Teacher-student experiments among six possible D_4-equivariant architectures. The last column is an unconstrained attention student. See Appendix E.1 for more details.

Theorem 2 shows that it is generally impossible to match the equivariant expressivity of an unconstrained MHSA using a single equivariant model of the same size. However, given H, d, it is possible to construct larger equivariant "covering models", which can express all equivariant MHSA with H heads and head dimension d. For finite G, we can use group averaging to construct a covering model with $H' = H|G|$ heads and head dimension $d' = d$. The group average of $F : \mathbb{R}^{D_{\text{in}}} \to \mathbb{R}^{D_{\text{out}}}$ is $\bar{F}(x) = \frac{1}{|G|}\sum_{g\in G} \rho_{\text{out}}(g)F(\rho_{\text{in}}(g)^{-1}x)$, which is always equivariant and satisfies $\bar{F} = F$ for equivariant F. Given unconstrained weight matrices $W_q^{(h)}, W_k^{(h)}, W_v^{(h)}, W_o^{(h)}$, we obtain the group average of the unconstrained MHSA by setting

$$W_t^{(h,g)} = W_t^{(h)}\rho_{\text{in}}(g)^{-1}, \text{ for } t = q, k, v, \qquad W_o^{(h,g)} = \tfrac{1}{|G|}\rho_{\text{out}}(g)W_o^{(h)}.$$

In particular, this covering model has precisely as many parameters as the unconstrained MHSA. Thus we get an upper bound on the number of heads needed for a covering model as $H|G|$. It would be interesting to find minimal covering models for practically useful cases such as Example 9.

5 Experiments

5.1 Modular Addition

It has been shown (Nanda et al., 2023) that a single transformer block trained on modular addition can learn to perform the discrete Fourier transform over C_p, the cyclic group of order p. Here, we train a single unconstrained 32-head self-attention layer with head dimension $d = 8$ to learn addition of three tokens $(a, b, c) \mapsto (b + c, c + a, a + b)$ modulo $p = 24$ or $p = 15$. In addition to the task loss, we also include an explicit equivariance loss l_{eq}. See Appendix E.2 for details.

We briefly summarize the representation theory of C_p [3], which is used to visualize the QK and OV matrices $M_h, R_h \in \mathbb{R}^{p\times p} = \mathbb{R}^p \otimes \mathbb{R}^p$: the irreps of C_p are U_0 (trivial, 1D, frequency 0), U_ω (2D, frequency ω) for $1 \le \omega < p/2$, and $U_{p/2}$ (sign, 1D, frequency $p/2$, only if p is even). The tensor product of two irreps corresponds to taking sum and difference of the two frequencies. For example, $U_2 \otimes U_2 \cong 2U_0 \oplus U_4$ (if $p \ge 9$). The regular representation decomposes as $\mathbb{R}^p \cong \bigoplus_{\omega=0}^{\lfloor \frac{p}{2} \rfloor} U_\omega$.

For a pair (query frequency ω_q, key frequency ω_k), pick a plane in $U_{\omega_q} \otimes U_{\omega_k} \subset \mathbb{R}^p \otimes \mathbb{R}^p$ isomorphic to U_ω, where $\omega = \omega_q + \omega_k$ or $\omega_q - \omega_k$ (note that the frequencies are defined modulo p and signs). We project each M_h onto this plane for visualization. For the OV matrices, the representation ρ_{out} on the output is doubled. For odd p ($= 15$), we still label the irrep subspaces in the output using the original frequency, so "frequency ω" in the output space refers to the subspace isomorphic to $U_{2\omega}$. Thus, for a pair (input frequency ω_i, output frequency ω_o), we pick an irrep plane in $U_{\omega_q} \otimes U_{2\omega_o}$ to project the OV matrices on.

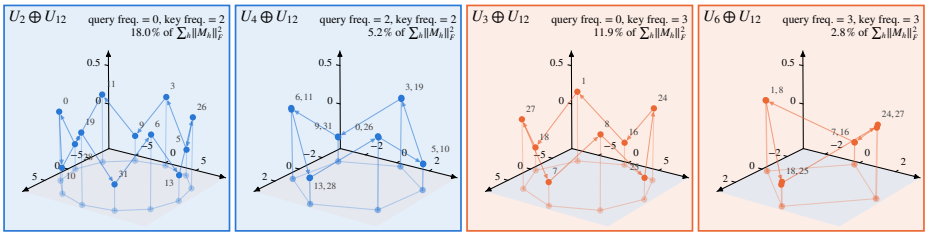

Figure 5: Learned QK matrices form "crowns" in 3-dimensional subspaces of $\mathbb{R}^{p\times p}$ isomorphic to $U_\omega \oplus U_{12}$, where ω is the frequency of a two-dimensional irrep. We omit heads that have small components in the U_ω plane. The numbers next to the points are the head indices. Blue panels: C_{24}/C_2-orbit formed by heads $0, 10, 19, 31, 6, 13, 26, 5, 3, 9, 11, 28$. Orange panels: C_{24}/C_3-orbit formed by heads $7, 8, 25, 24, 16, 1, 18, 27$. The arrow attached to a dot points to where it is carried by the action of the generator of C_{24}.

For $p = 24$, we additionally project each M_h onto $U_0 \otimes U_{12} \cong U_{12}$. The generator of C_{24} acts on the three-dimensional space $U_\omega \oplus U_{12}$ by rotating the U_ω plane by $2\pi \cdot \frac{\omega}{24}$ and simultaneously flipping the U_{12} axis. Hence, if ω divides 24, Theorem 1 predicts that the QK matrices organize themselves in this 3D subspace into regular $\frac{24}{\omega}$-gons in the U_ω plane with alternating components in U_{12}. This is indeed what we find in the experiment (Fig. 5)

The $p = 15$ run exhibits a behavior not present in the $p = 24$ run. In Figure 6, we project the QK and OV matrices onto irrep planes with various frequencies. In the top-left plot (U_2), we see 26 heads forming a regular 15-gon orbit. This is fine because three of them clustered together, and another 18 of them paired into nine two-head clusters. This is then reflected in the OV plot (bottom left, also U_2): the individual R_h (blue dots) themselves do not form a size-15 orbit of C_{15}, but the cluster-summed values (purple dots) do. The same discussion carries over to the four other blue panels, where the same 26 heads are arranged differently due to different frequencies.

That still leaves us with $32 - 26 = 6$ heads. Heads 11 and 14 are singleton clusters, i.e., they are individually equivariant and they only every appear in the center of each plot in Fig. 6. The remaining four form a "pseudo-orbit" (orange panels in Fig. 6). Since 4 does not divide 15, it is impossible for these four heads to form a C_{15}-orbit. Instead, they form a square in the chosen U_1 planes, an arrangement that breaks equiviariance, but not too much since their total squared Frobenius norm is only about 0.48% of the total of all heads (the final equivariance loss is $\sim 10^{-4}$).

We would like to note a caveat: Since the input and output are one-hot encodings, solving the task does not necessarily imply equivariance on the entire $\mathbb{R}^p$, so an equivariance loss is used to is explicitly encouraged to become equivariant. In our case, using solely l_{task} resulted in the model learning the task without becoming equivariant. On the other hand, l_{eq} alone often results in collapse of the OV matrices. So both losses are necessary for the structure. Moreover, both models reach 100% training accuracy, but the $p = 24$ model generalizes perfectly (test accuracy 100%), while the $p = 15$ model does not (test accuracy 0%), suggesting that the intricate structures come entirely from equivariance, and not the task. This is consistent with Theorem 1.

[3] which is just discrete Fourier transform.

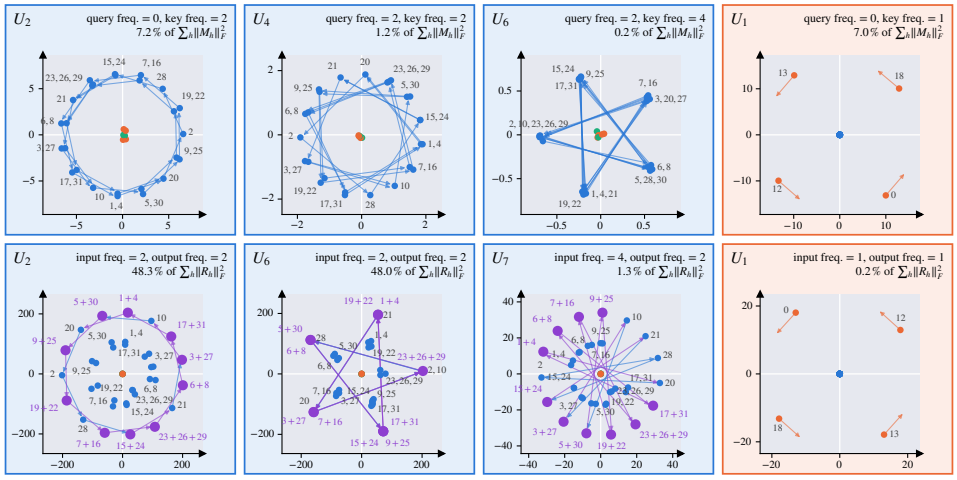

Figure 6: Learned QK and OV matrices in the $p = 15$ run. Top row: QK matrices projected onto 2D irrep planes. Bottom row: OV matrices. Blue panels: 26 heads form an orbit of size 15, with one 3-head cluster sharing one QK matrix, and 9 two-head clusters. In the plots of OV matrices, we sum the R_h within clusters, and the results are shown as purple dots. In the each of the two U_2 planes, the M_h and the cluster-summed R_h form regular 15-gons. This is because the rotation generator is mapped to rotation by $2/15 \cdot 2\pi$ in U_2 (as indicated by the arrows), and 2 is coprime to 15. The same applies to frequencies 4 and 7. In frequency 6, however, the generator rotates the plane by $6/15 \cdot 2\pi = 2/5 \cdot 2\pi$, so the 15-gons collapse into pentagrams.

5.2 Almost Equivariant Vision Transformers

We train a vision transformer to process an input vector field represented by *raster arrows* (see Fig. 7(a)). The model's task is to output the vector field $\mathbf{v}$ itself and all four components of the derivative $\nabla\mathbf{v}$ (see Appendix E.3 for details). In total, the target has $D_{\text{out}} = 6$ components. In this experiment, we do *not* use an explicit equivariance loss.

We use a D_4-equivariant vision transformer (Nordström et al., 2026; Ông & Bökman, 2026) with $D_{\text{in}} = 288$ channels (36 copies of the regular representation) and six heads, with one twist: the final block is one unconstrained MHSA layer, directly followed by a tokenwise linear layer with D_{out} output channels (Fig. 7(a)). That is, the last block does not have an MLP, residual connections, or any normalization. We also do not perform any invariantization, since the task is only equivariant and not invariant by design.

As before, the trained model found the structure predicted by Theorem 1. Specifically, we compute $\|\rho_{\text{in}}(g) M_h \rho_{\text{in}}(g)^{-1} - M_{h'}\|$ for each pair of heads (h, h') and group element $g \in G$ (and similarly for the OV matrices) and find that heads $0, 1, 2, 4$ organized themselves into a D_4-orbit, and heads $3, 5$ are each invariant (Fig. 7(b)–7(e)).

To visualize the geometric arrangement of the learned QK matrices M_h, for each irrep ρ of D_4, we pick the subspace of $\mathbb{R}^{288\times 288}$ isomorphic to ρ that has the largest overlap with the M_h, and project them onto the three-dimensional subspace spanned by the chosen E_1 plane and a 1D irrep. The results are shown in Fig. 7(f). In $\mathrm{E}_1 \oplus \mathrm{A}_2$ for example, the reflection generator s flips the sign of the A_2 component. But head 0 is invariant under s, so its QK matrix M_0 is pinned at zero in the A_2 axis. The rotation generator then makes this into a flat square. The same happens in $\mathrm{E}_1 \oplus \mathrm{B}_2$.

6 Summary, Limitations & Future Work

We proved an exact structure theorem for equivariant MHSA (Theorem 1), which was then used to provide a lower bound on the number of equivariant models required to match the equivariant expressivity of a single unconstrained MHSA (Theorem 2). Finally, we illustrated in two experiments

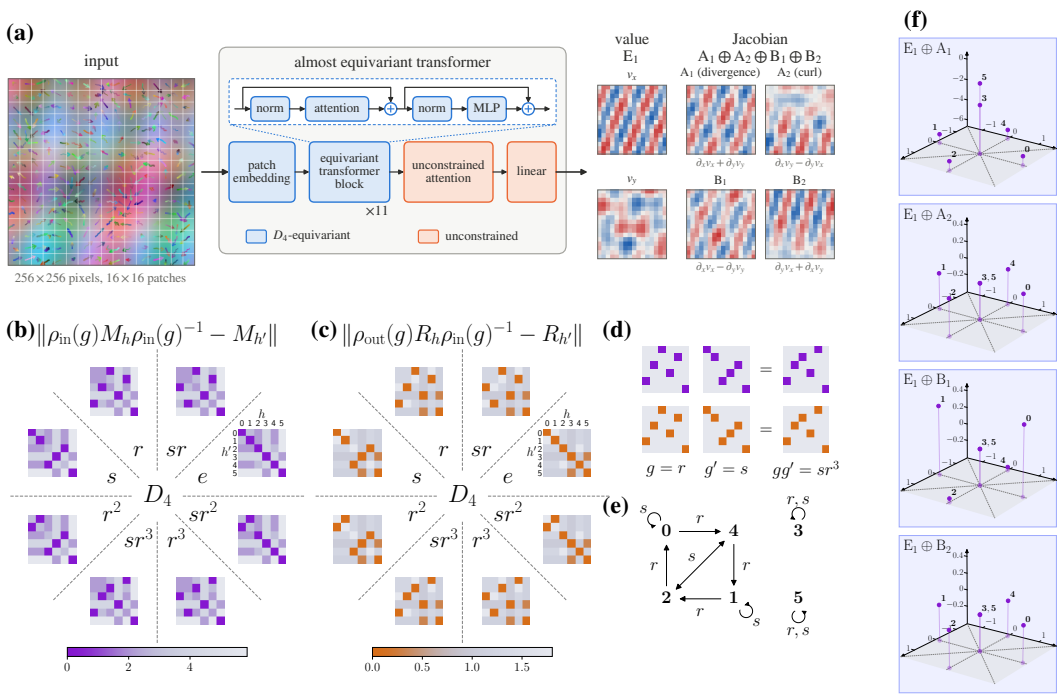

Figure 7: **(a)** Almost equivariant vision transformer experiment. Data: a random vector field drawn as one arrow per 16×16 patch, with six equivariant targets per patch. Model: a D_4-equivariant vision transformer backbone followed by an unconstrained attention layer. **(b)** Distances $\|\rho_{\text{in}}(g) M_h \rho_{\text{in}}(g)^{-1} - M_{h'}\|$ between the QK matrices and their conjugates, one heatmap per $g \in D_4$, placed in the chamber of g. Theorem 1 predicts that the map $g \mapsto (h \mapsto \text{argmin}_{h'} \|\rho_{\text{in}}(g) M_h \rho_{\text{in}}(g)^{-1} - M_{h'}\|)$ is a homomorphism, and it holds empirically here. **(c)** The same for the OV matrices R_h. **(d)** Products of the resulting permutation matrices, where purple/orange $= 1$ and gray $= 0$. **(e)** The inferred action of D_4 on the heads: heads $0, 1, 2, 4$ form one orbit, heads 3 and 5 form singleton orbits. **(f)** The six QK matrices M_h projected onto four D_4-stable three-dimensional subspaces of $\mathbb{R}^{288\times 288}$. We pick the subspaces where the M_h have most overlap.

that the intricate structure predicted by Theorem 1 can be found "in the wild" in networks trained on equivariant tasks.

Other trade-offs. The equivariance-expressivity trade-off, formalized by Theorem 2 as well as the group-averaging argument in Sec. 1, is in terms of parameter count and sets of expressible equivariant functions. Other trade-offs, such as compute, are not accounted for. For example, achieving equivariance at the same parameter count without losing expressivity with $|G|$ times more heads still results in $|G|$ times more operations in each forward, on top of the conjugation operations needed for weight sharing. In the opposite direction, if one is willing to compromise expressivity, equivariance constraints at fixed width result in fewer operations in the Fourier domain due to Schur's lemma.

Going deeper. Theorem 1 covers a single layer of MHSA. The same characterization question remains open for an entire transformer, or even just for a single transformer block. As a piece of empirical evidence, Nanda et al. (2023) found that one transformer block learns Fourier/irrep structures when trained on a modular addition task, which is equivariant, suggesting a plausible extension of Theorem 1.

Approximate equivariance. The algebro-geometric approach in the proof of Theorem 2 worked because exact equivariance is translated to clean algebraic constraints on the MHSA parameter space via Theorem 1. In particular, Theorem 1 says nothing about approximately equivariant architectures, or about approximating equivariant maps. It would be nice to have a statement of the form T approximately equivariant $\Rightarrow (M, R)$ close to the equivariance locus, in the form of bounds. This could explain the pseudo-orbits we see in the modular addition experiments.

AI USE STATEMENT

We used LLMs (Anthropic's Claude models, accessed through the Claude desktop app, Claude Code and the Claude VS Code extension, as well as Open AI's Codex via its VS Code extension from June-September 2026) for:

- Generating synthetic data and code for the experiments in Section. 5.
  - Source code was read and the synthetic data were inspected by one of the authors.
  - LLM-generated unit tests passed and were inspected by one of the authors.
- Checking for typos.
- Identifying proof gaps
  - Every gap identified by the LLM was closed by one of the authors with their own human-written proof.
- Checking the proofs of Theorem 1 and Theorem 2 by having the LLM translate the authors' natural-language statements into Lean formalizations. The resulting Lean code compiled and one of the authors checked that the Lean formalizations of the statements correctly represented the natural-language statements.
- Identifying relevant literature (checked by one of the authors).
- Generating the TikZ and Matplotlib code for the figures.

We did not use LLMs for:

- Helping develop theoretical models or conceptual frameworks
- Formulating any mathematical claim.
- Suggesting presentation structure or style.
- Designing or providing feedback on research methodology or experiments.
- Proposing or refining hypotheses.
- Assisting in natural-language translation.
- Interpreting results.
- Generating proofs.
- Drafting any part of the paper or writing anything in general apart from correcting isolated typos.
- Formatting references.

The following tasks are not applicable to this work:

- Supporting qualitative and thematic data analysis.
- Cleaning and reformatting dataset

We have reviewed all LLM-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Devanshu Agrawal and James Ostrowski. A Classification of G-invariant Shallow Neural Networks, January 2023. URL http://arxiv.org/abs/2205.09219. arXiv:2205.09219 [cs.LG].

Erik J. Bekkers, Maxime W. Lafarge, Mitko Veta, Koen AJ Eppenhof, Josien PW Pluim, and Remco Duits. Roto-Translation Covariant Convolutional Networks for Medical Image Analysis, April 2018. URL https://arxiv.org/abs/1804.03393v3.

Georg Bökman, David Nordström, and Fredrik Kahl. Flopping for FLOPs: Leveraging equivariance for computational efficiency, June 2025. URL http://arxiv.org/abs/2502.05169. arXiv:2502.05169 [cs].

Nick Cammarata, Shan Carter, Gabriel Goh, Chris Olah, Michael Petrov, Ludwig Schubert, Chelsea Voss, Ben Egan, and Swee Kiat Lim. Thread: Circuits. *Distill*, 2020. doi: 10.23915/distill.00024. https://distill.pub/2020/circuits.

Giansalvo Cirrincione. Rank, Head-Channel Non-Identifiability, and Symmetry Breaking: A Precise Analysis of Representational Collapse in Transformers, April 2026. URL http://arxiv.org/abs/2604.23681. arXiv:2604.23681 [cs.LG].

Taco Cohen, Mario Geiger, and Maurice Weiler. A General Theory of Equivariant CNNs on Homogeneous Spaces, January 2020. URL http://arxiv.org/abs/1811.02017. arXiv:1811.02017 [cs].

Taco S. Cohen and Max Welling. Group Equivariant Convolutional Networks, February 2016. URL https://arxiv.org/abs/1602.07576v3.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing Convolutional Neural Networks for Equivariance to Lie Groups on Arbitrary Continuous Data, September 2020. URL http://arxiv.org/abs/2002.12880. arXiv:2002.12880 [cs, stat].

Jiahong Fu, Qi Xie, Deyu Meng, and Zongben Xu. Vanilla Group Equivariant Vision Transformer: Simple and Effective, February 2026. URL http://arxiv.org/abs/2602.08047. arXiv:2602.08047 [cs].

Fabian B. Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks, November 2020. URL http://arxiv.org/abs/2006.10503. arXiv:2006.10503 [cs].

Nathan W. Henry. Generic Fibers and Functional Dimension of Multi-Head Attention. In *ICML 2026 Workshop on Weight-Space Symmetries: from Foundations to Practical Applications*, 2026. URL https://openreview.net/forum?id=7vPmnHdAPl.

Nathan W. Henry, Giovanni Luca Marchetti, and Kathlén Kohn. Geometry of Lightning Self-Attention: Identifiability and Dimension, June 2026. URL http://arxiv.org/abs/2408.17221. arXiv:2408.17221 [cs.LG].

Michael Hutchinson, Charline Le Lan, Sheheryar Zaidi, Emilien Dupont, Yee Whye Teh, and Hyunjik Kim. LieTransformer: Equivariant self-attention for Lie Groups, June 2021. URL http://arxiv.org/abs/2012.10885. arXiv:2012.10885 [cs].

Mohammad Mohaiminul Islam, Rishabh Anand, David R. Wessels, Friso de Kruiff, Thijs P. Kuipers, Rex Ying, Clara I. Sánchez, Sharvaree Vadgama, Georg Bökman, and Erik J. Bekkers. Platonic Transformers: A Solid Choice For Equivariance, October 2025. URL http://arxiv.org/abs/2510.03511. arXiv:2510.03511 [cs].

Kathlén Kohn, Anna-Laura Sattelberger, and Vahid Shahverdi. Geometry of Linear Neural Networks: Equivariance and Invariance Under Permutation Groups. *SIAM J. Matrix Anal. Appl.*, 46(2):1378–1415, June 2025. ISSN 0895-4798, 1095-7162. doi: 10.1137/24M166053X. URL https://epubs.siam.org/doi/10.1137/24M166053X.

Risi Kondor and Shubhendu Trivedi. On the Generalization of Equivariance and Convolution in Neural Networks to the Action of Compact Groups, November 2018. URL http://arxiv.org/abs/1802.03690. arXiv:1802.03690 [stat].

Giovanni Luca Marchetti, Christopher Hillar, Danica Kragic, and Sophia Sanborn. Harmonics of Learning: Universal Fourier Features Emerge in Invariant Networks, June 2024. URL http://arxiv.org/abs/2312.08550. arXiv:2312.08550 [cs.LG].

Giovanni Luca Marchetti, Vahid Shahverdi, Stefano Mereta, Matthew Trager, and Kathlén Kohn. Algebra Unveils Deep Learning – An Invitation to Neuroalgebraic Geometry, May 2025. URL http://arxiv.org/abs/2501.18915. arXiv:2501.18915 [cs.LG].

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability, January 2023. URL https://arxiv.org/abs/2301.05217v3.

David Nordström, Johan Edstedt, Fredrik Kahl, and Georg Bökman. Quick ViTs: Speeding up Vision Transformers through Equivariance, July 2026. URL http://arxiv.org/abs/2505.15441. arXiv:2505.15441 [cs.CV].

Marco Pacini, Mircea Petrache, Bruno Lepri, Shubhendu Trivedi, and Robin Walters. On Universality of Deep Equivariant Networks, October 2025. URL http://arxiv.org/abs/2510.15814. arXiv:2510.15814 [stat].

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets, January 2022. URL http://arxiv.org/abs/2201.02177. arXiv:2201.02177 [cs].

Siamak Ravanbakhsh. Universal Equivariant Multilayer Perceptrons. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 7996–8006. PMLR, November 2020. URL https://proceedings.mlr.press/v119/ravanbakhsh20a.html.

David W. Romero and Jean-Baptiste Cordonnier. Group Equivariant Stand-Alone Self-Attention For Vision, March 2021. URL http://arxiv.org/abs/2010.00977. arXiv:2010.00977 [cs].

Vahid Shahverdi, Giovanni Luca Marchetti, Georg Bökman, and Kathlén Kohn. Identifiable Equivariant Networks are Layerwise Equivariant, May 2026. URL http://arxiv.org/abs/2601.21645. arXiv:2601.21645 [cs.LG].

J. Shawe-Taylor. Building symmetries into feedforward networks. In *1989 First IEE International Conference on Artificial Neural Networks, (Conf. Publ. No. 313)*, pp. 158–162, October 1989. URL https://ieeexplore.ieee.org/document/51951.

Sihui Wei, Gavin McCracken, Gabriela Moisescu-Pareja, Harley Wiltzer, Doina Precup, Irina Rish, and Jonathan Love. Deep neural networks divide and conquer dihedral multiplication. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=hFmWsd2m6w.

Maurice Weiler and Gabriele Cesa. General E(2)-Equivariant Steerable CNNs. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/45d6637b718d0f24a237069fe41b0db4-Abstract.html.

Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco Cohen. 3D Steerable CNNs: Learning Rotationally Equivariant Features in Volumetric Data, October 2018. URL http://arxiv.org/abs/1807.02547. arXiv:1807.02547 [cs].

Maurice Weiler, Patrick Forré, Erik Verlinde, and Max Welling. Coordinate Independent Convolutional Networks – Isometry and Gauge Equivariant Convolutions on Riemannian Manifolds, June 2021. URL http://arxiv.org/abs/2106.06020. arXiv:2106.06020 [cs.LG].

Maurice Weiler, Patrick Forré, Erik Verlinde, and Max Welling. *Equivariant and Coordinate Independent Convolutional Networks: A Gauge Field Theory of Neural Networks*. Progress in Data Science. World Scientific, 2026. ISBN 978-981-98-0662-1. doi: 10.1142/14143.

Jeffrey Wood and John Shawe-Taylor. Representation theory and invariant neural networks. *Discrete Applied Mathematics*, 69(1):33–60, 1996. ISSN 0166-218X. doi: https://doi.org/10.1016/0166-218X(95)00075-3. URL https://www.sciencedirect.com/science/article/pii/0166218X95000753.

Renjun Xu, Kaifan Yang, Ke Liu, and Fengxiang He. $E(2)$-Equivariant Vision Transformer, July 2023. URL http://arxiv.org/abs/2306.06722. arXiv:2306.06722 [cs].

Dmitry Yarotsky. Universal approximations of invariant maps by neural networks, April 2018. URL http://arxiv.org/abs/1804.10306. arXiv:1804.10306 [cs.NE].

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=S5wmbQc1We.

Tīkun Ông and Georg Bökman. A Unified Framework for Vision Transformers Equivariant to Discrete Subgroups of $\mathrm{O}(2)$, June 2026. URL http://arxiv.org/abs/2606.27864. arXiv:2606.27864 [cs.CV].

A Related Work

Equivariant architectures. Most existing equivariant neural network architectures are equivariant convolutional neural networks (Cohen & Welling, 2016; Kondor & Trivedi, 2018; Bekkers et al., 2018; Weiler & Cesa, 2019; Weiler et al., 2018; Cohen et al., 2020; Finzi et al., 2020; Weiler et al., 2021; 2026). This paper says nothing about these. Equivariant transformers, on the other hand, rely on equivariant attention, which is the topic of this paper and has been achieved either by lifting features to functions on the group (Hutchinson et al., 2021; Romero & Cordonnier, 2021; Xu et al., 2023; Fu et al., 2026) or working directly in the Fourier space (Fuchs et al., 2020; Bökman et al., 2025; Nordström et al., 2026; Ông & Bökman, 2026).

Universality. Universality of equivariant networks is usually studied in the approximate sense, i.e., uniform convergence on compact sets, for shallow or deep MLPs (Yarotsky, 2018; Ravanbakhsh, 2020; Pacini et al., 2025). Our work concerns the structure of the space of equivariant functions expressible by MHSA. We do not prove anything about how good the different found components are in the sense of, say, approximating different classes of equivariant functions that are not expressible by a given MHSA.

Identifiability and neuroalgebraic geometry. Identifiability refers to the ability to recover the parameters of a model, up to certain predefined "gauge symmetries", from the function it represents. Shahverdi et al. (2026) observe that identifiability together with the so-called *adjunction property* implies layerwise equivariance for equivariant functions. Henry et al. (2026) establish identifiability for lightning ("linear", without softmax) self-attention. More recently, Henry (2026) characterizes the generic fiber of ordinary MHSA. The problem of understanding the image and fibers of the realization map of a neural network can sometimes be studied using tools from algebraic geometry. This is the neuroalgebraic geometry program (Marchetti et al., 2025). For example, Kohn et al. (2025) study the equivariance locus of a two-layer linear network for cyclic and permutation groups. The irreducible components they find have the same origin as the ones for MHSA described in Section C.

Interpretability. Interpretability (or mechanistic interpretability) concerns trying to reverse engineer the underlying algorithms implemented by a neural network trained to solve a given task, as well as understanding how these algorithms emerge during training (Cammarata et al., 2020; Power et al., 2022). Often, simple model tasks are chosen in the literature, such as modular addition or multiplication in a small group Nanda et al. (2023); Zhong et al. (2023); Wei et al. (2026). These tasks, and real world tasks alike, often contain symmetries. Interpretability of how the symmetries are encoded in any fully equivariant MHSA is given by our Theorem 1.

B Proof of Theorem 1

B.1 Notational conventions in the appendix

In Appendices B, C, D, we will use slightly different notation from the main text:

- We use $\langle x, y\rangle$ instead of $x^\top y$ to denote the inner product of x and y.
- The feature spaces are now $V_{\text{in}}, V_{\text{out}}$ (or V) instead of $\mathbb{R}^{D_{\text{in}}}$ or $\mathbb{R}^{D_{\text{out}}}$.
- We write $\mathrm{Hom}(V_{\text{in}}, V_{\text{out}})$ instead of $\mathbb{R}^{D_{\text{out}}\times D_{\text{in}}}$, and $\mathrm{End}(V_{\text{in}})$ instead of $\mathbb{R}^{D_{\text{in}}\times D_{\text{in}}}$.
- The adjoint of a lienar map with respect to an inner product is ϕ^* instead of $\phi^\top$.
- H is now the set of heads rather than the number of heads.
- We will simply write gx instead of $\rho_{\text{in}}(g)x$ or $\rho_{\text{out}}(g)x$.

We also introduce the notion of reduced MHSA, which is just attention parameterized directly by the QK and OV matrices instead of query, key, value, and output projection maps.

Definition 2. *For $(M_h)_{h=1}^H, (R_h)_{h=1}^H$ with $M_h \in \mathrm{End}(V_{\text{in}})$, $R_h \in \mathrm{Hom}(V_{\text{in}}, V_{\text{out}})$, we define **reduced multi-head self-attention (RMHSA)** to be*

$$\mathrm{RMHSA}(x; (M_h)_{h\in H}, (R_h)_{h\in H})_i := \sum_{h=1}^{H} \frac{\sum_{j=1}^{L} \exp(\langle x_i, M_h x_j\rangle) R_h x_j}{\sum_{j=1}^{L} \exp(\langle x_i, M_h x_j\rangle)}. \tag{6}$$

Note that "reduced" refers to the reduction in parameterization redundancy, not in expressive power. Indeed, a priori, RMHSA could have more expressive power than MHSA if the ranks of the M_h and R_h are unconstrained.

B.2 THE PROOF

We will show (3) $\Rightarrow$ (2) $\Rightarrow$ (1) $\Rightarrow$ (3). The implication (2) $\Rightarrow$ (1) is trivial. (3) $\Rightarrow$ (2) is direct computation:

$$\begin{aligned}\sum_{h\in H}\frac{\sum_{j=1}^{L}\exp(\langle gx_i, M_h gx_j\rangle)R_h gx_j}{\sum_{j=1}^{L}\exp(\langle gx_i, M_h gx_j\rangle)} &= \sum_{p\in X}\frac{\sum_{j=1}^{L}\exp(\langle gx_i, M_p gx_j\rangle)R_p gx_j}{\sum_{j=1}^{L}\exp(\langle gx_i, M_p gx_j\rangle)}\\ = \sum_{p\in X}\frac{\sum_{j=1}^{L}\exp(\langle x_i, M_{g^{-1}p}x_j\rangle)gR_{g^{-1}p}x_j}{\sum_{j=1}^{L}\exp(\langle x_i, M_{g^{-1}p}x_j\rangle)} &= g\sum_{p\in X}\frac{\sum_{j=1}^{L}\exp(\langle x_i, M_p x_j\rangle)R_p x_j}{\sum_{j=1}^{L}\exp(\langle x_i, M_p x_j\rangle)}\\ = g\sum_{h\in H}\frac{\sum_{j=1}^{L}\exp(\langle x_i, M_h x_j\rangle)R_h x_j}{\sum_{j=1}^{L}\exp(\langle x_i, M_h x_j\rangle)}.\end{aligned} \tag{7}$$

We now prove (1) $\Rightarrow$ (3).

The equivariance condition for L tokens $(x_1,\dots,x_L)\in\mathbb{R}^L\otimes V$ and $|H|$ heads $V=\bigoplus_{h\in H}V_h$ at the i-th output token reads

$$\sum_{h\in H}\frac{\sum_{j=1}^{L}\exp(\langle gx_i, M_h gx_j\rangle)R_h gx_j}{\sum_{j=1}^{L}\exp(\langle gx_i, M_h gx_j\rangle)} = g\sum_{h\in H}\frac{\sum_{j=1}^{L}\exp(\langle x_i, M_h x_j\rangle)R_h x_j}{\sum_{j=1}^{L}\exp(\langle x_i, M_h x_j\rangle)} \tag{8}$$

for all $g\in G$. Hence,

$$\sum_{h\in H}\frac{\sum_{j=1}^{L}\exp(\langle x_i, M_h^g x_j\rangle)R_h^g x_j}{\sum_{j=1}^{L}\exp(\langle x_i, M_h^g x_j\rangle)} = \sum_{h\in H}\frac{\sum_{j=1}^{L}\exp(\langle x_i, M_h x_j\rangle)R_h x_j}{\sum_{j=1}^{L}\exp(\langle x_i, M_h x_j\rangle)}, \tag{9}$$

where $M_h^g := gM_hg^{-1}$ and $R_h^g := gR_hg^{-1}$. Setting all tokens to the same $x\in V$ yields $g\sum_h R_hg^{-1}x = \sum_h R_h x$, or $\sum_h R_h\in \mathrm{End}_G(V)$. Now, set $x_{j\geq 2}=x$, $x_1=x+y$, and $i=2$ in Eq. (9) to get

$$\sum_{h\in H}\frac{(L-1)R_h^g x + e^{\langle x, M_h^g y\rangle}R_h^g(x+y)}{L-1+e^{\langle x, M_h^g y\rangle}} = \sum_{h\in H}\frac{(L-1)R_h x + e^{\langle x, M_h y\rangle}R_h(x+y)}{L-1+e^{\langle x, M_h y\rangle}}. \tag{10}$$

Cancelling out $\sum_h R_h^g x = \sum_h R_h x$ yields

$$\sum_{h\in H}\frac{e^{\langle x, M_h^g y\rangle}R_h^g y}{L-1+e^{\langle x, M_h^g y\rangle}} = \sum_{h\in H}\frac{e^{\langle x, M_h y\rangle}R_h y}{L-1+e^{\langle x, M_h y\rangle}}. \tag{11}$$

Or, writing $\sigma_\alpha(s) := e^s/(\alpha+e^s)$,

$$\sum_{h\in H}\sigma_{L-1}(\langle x, M_h^g y\rangle)R_h^g y = \sum_{h\in H}\sigma_{L-1}(\langle x, M_h y\rangle)R_h y. \tag{12}$$

Differentiating Eq. (12) n times with respect to x at $x=0$ yields

$$\sigma_{L-1}^{(n)}(0)\sum_{h\in H}(M_h^g y)^{\otimes n}\otimes R_h^g y = \sigma_{L-1}^{(n)}(0)\sum_{h\in H}(M_h y)^{\otimes n}\otimes R_h y. \tag{13}$$

Here, the superscript (n) denotes the n-th derivative of a function.

Lemma 1. *If α is an integer larger than one, then $\sigma_\alpha^{(n)}(0)\neq 0$ for all $n\in\mathbb{N}_{\geq 0}$.*

Proof. Let $\lambda := 1/\alpha < 1$. We can rewrite the sigmoid-like function as

$$\sigma_\alpha(x) = \lambda e^x\frac{1}{1+\lambda e^x} = \lambda e^x\sum_{b=0}^{\infty}(-1)^b\lambda^b e^{bx} = \sum_{b=1}^{\infty}(-1)^{b+1}\lambda^b e^{bx}. \tag{14}$$

Thus

$$\sigma_\alpha^{(n)}(0) = \sum_{b=1}^{\infty} (-1)^{b+1} \lambda^b b^n, \tag{15}$$

which can be related to the Eulerian polynomials A_n for $n \geq 1$ (the claim is clearly true for $n = 0$):

$$\sigma_\alpha^{(n)}(0) = \frac{\lambda A_n(-\lambda)}{(1+\lambda)^{n+1}}. \tag{16}$$

So $\sigma_\alpha^{(n)}(0) = 0 \Leftrightarrow A_n(-\lambda) = 0$. But A_n is a monic polynomial with constant term 1, so by the rational root theorem, the only possible rational roots are ± 1. □

Thus, if the number of tokens is at least three, then for all $n \geq 0$ we have

$$\sum_{h \in H} (M_h^g y)^{\otimes n} \otimes R_h^g y = \sum_{h \in H} (M_h y)^{\otimes n} \otimes R_h y, \tag{17}$$

or

$$\sum_{h \in H} (M_h y)^{\otimes n} \otimes R_h y + \sum_{h \in H} (M_h^g y)^{\otimes n} \otimes (-R_h^g y) = 0. \tag{18}$$

Lemma 2. *Let V be a finite dimensional real vector space and let $(R_i)_{i \in I}, (M_i)_{i \in I}$ be finite indexed families in $\mathrm{End}(V)$. Suppose*

$$\forall n \geq 0, \forall y \in V : \sum_{i \in I} (M_i y)^{\otimes n} \otimes R_i y = 0. \tag{19}$$

Partition $I = \bigsqcup_{p \in Z} C_p$ such that $M_i = M_j \Leftrightarrow \exists p : i, j \in C_p$. Then $\sum_{i \in C_p} R_i = 0$ for all $p \in Z$.

Proof. Roughly speaking, the proof idea is to first prove the claim pointwise and then employ a density argument.

Let

$$\begin{aligned} S :=& \{y \in V : \forall i, j : M_i y = M_j y \Leftrightarrow M_i = M_j\} \\ & \cap \{y \in V : \forall i : M_i \neq 0 \Rightarrow M_i y \neq 0\} \end{aligned} \tag{20}$$

Then S is nonempty and open.

Fix $y \in S$. Find $\alpha \in V^*$ that separates the $M_i y$, i.e. $\alpha(M_i y) = \alpha(M_j y) \Leftrightarrow M_i y = M_j y$. Furthermore, we may assume $\alpha(M_i y) = 0 \Leftrightarrow M_i y = 0$ ($\Leftrightarrow M_i = 0$ because $y \in S$). Define $c_p := \alpha(M_i y)$ for any $i \in C_p$. Then, $\alpha^{\otimes n} \otimes \mathrm{Id}$ evaluated on Eq. (19) reads

$$\forall n \geq 0 : \sum_{p \in Z} c_p^n \sum_{i \in C_p} R_i y = 0. \tag{21}$$

First, suppose $c_p \neq 0$ for all $p \in Z$. Then the c_p are nonzero and pairwise distinct. Thus, the matrix with entries $\mathbf{V}_{np} = c_p^n$, $n = 0, 1, \ldots, |Z| - 1$ is invertible, so $\sum_{i \in C_p} R_i y = 0$ for all $p \in Z$ and $y \in S$.

If $c_{p_0} = 0$ for some $p_0 \in Z$ ($\Leftrightarrow M_i = 0$ for some i), then the same argument still implies $\sum_{i \in C_p} R_i y = 0$ for $p \in Z \setminus \{p_0\}$ and $y \in S$, which in turn implies $\sum_{i \in C_{p_0}} R_i y = 0$ since $\sum_{i \in I} R_i y = 0$ (Eq. (19) at $n = 0$).

Hence, $\sum_{i \in C_p} R_i y = 0$ for all $p \in Z$ and $y \in S$. Since S is nonempty and open, $\sum_{i \in C_p} R_i = 0$. □

Applied to Eq. (18), the summation index set is $I = H \sqcup H = \bigsqcup_{p \in Z} C_p$, partitioned by the values of $(M_h)_{h \in H}$ and $(M_h^g)_{h \in H}$. Partition $H = \bigsqcup_{p \in Y} B_p$ by $(M_h)_{h \in H}$, and write $M_p := M_h$ for any $h \in B_p$ and $R_p := \sum_{h \in B_p} R_h$. Clearly, we have $M_h^g = M_{h'}^g \Leftrightarrow h, h'$ are in some B_p. That is, the partition of $H \sqcup H$ must be coarser than $\left(\bigsqcup_{p \in Y} B_p\right) \sqcup \left(\bigsqcup_{p \in Y} B_p\right)$. Define

$$X_0 := \{p \in Y : R_p = 0\} \qquad X := Y \setminus X_0. \tag{22}$$

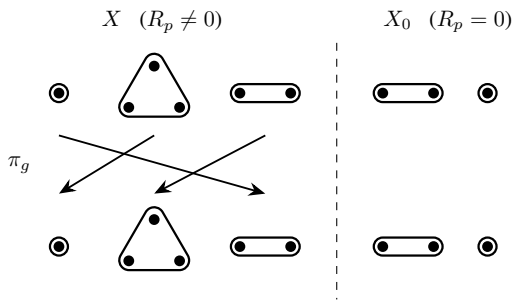

Figure 8: Illustration of the argument that applies Lemma 2 to Eq. (18) to get a partial G-clustering.

Define a map $\pi_g : X \to X$ as follows: Take $p \in X$. Then $R_p \neq 0$ by construction. So, by Lemma 2, there must be exactly one $p' \in Y$ such that $M_p = M^g_{p'}$ and $R_p = R^g_{p'}$. But then $R_{p'} \neq 0$, so p' must belong to X. Put $\pi_g(p) = p'$. If $\pi_g(p_1) = \pi_g(p_2)$, then $M_{p_1} = M_{p_2}$, so π_g is bijective.

Next, we show that $g \mapsto \pi_g$ is an antihomomorphism. Take $g_1, g_2 \in G$ and $p \in X$. Let $p' = \pi_{g_1}(p)$, $p'' = \pi_{g_2} \circ \pi_{g_1}(p) = \pi_{g_2}(p')$. Then $M_p = M^{g_1}_{p'}$ and $M_{p'} = M^{g_2}_{p''}$, so $M_p = M^{g_1 g_2}_{p''}$, i.e., $p'' = \pi_{g_1 g_2}(p)$. That is, $\pi_{g_1 g_2} = \pi_{g_2} \circ \pi_{g_1}$.

Let G act on X by $g \cdot p = \pi_g^{-1}(p)$. Then $M_{g \cdot p} = M_{\pi_g^{-1}(p)} = M^g_{\pi_g \circ \pi_g^{-1}(p)} = M^g_p$, and similarly $R_{gp} = R^g_p$. So $p \mapsto M_p$ and $p \mapsto R_p$ are G-equivariant on X. This concludes the proof of Theorem 1, where the partial G-clustering is given by the natural surjection $\chi : H \to Y$ giving the partition. This last argument is illustrated in Fig. 8.

Finally, we would like to remark that the same technique can be used to prove the following identifiability result, which slightly generalizes Henry (2026):

Theorem 3. *Suppose $M, R \in \mathrm{End}(V)^H$ and $\bar{M}, \bar{R} \in \mathrm{End}(V)^{\bar{H}}$ satisfy*

$$\mathrm{RMHSA}(-; M, R) = \mathrm{RMHSA}(-; \bar{M}, \bar{R}) : \mathbb{R}^L \otimes V \to \mathbb{R}^L \otimes V \tag{23}$$

for some $L \geq 3$. Then

$$\sum_{h : M_h = m} R_h = \sum_{\bar{h} : \bar{M}_{\bar{h}} = m} \bar{R}_{\bar{h}} \tag{24}$$

for any $m \in \mathrm{End}(V)$.

C Geometry of the Equivariance Locus

The proof is split into four sections. Section C.1 merely motivates the machinery we use; the actual proof is completely contained in Sections C.2, C.3, and C.4.

C.1 Motivation: Equivariant Vector Bundles

In the proof, we will make heavy use of *equivariant vector bundles*, the definition and properties of which are reviewed in Appendix D. This is but a convenient way to pack a lot of information into one single object, motivated in three steps as follows:

(1) Let V, W be G-representations. Suppose we want to study the family $\mathcal{F}$ of G-equivariant linear maps $M \in \mathrm{Hom}_G(V, W)$ with $\mathrm{rk}(M) \leq r$. It is true that any such M factorizes through some G-representation U of dimension $\leq r$. That is, $M = \psi_1 \psi_2$ with $\psi_2 \in \mathrm{Hom}_G(V, U)$ and $\psi_1 \in \mathrm{Hom}_G(U, W)$. Thus, $\mathcal{F} = \bigcup_U \mathcal{D}(V, W; U)$, where $\mathcal{D}(V, W; U) := \mathrm{Hom}_G(U, W)\mathrm{Hom}_G(V, U)$, and U ranges over all G-representations of dimension at most r up to isomorphism.

(2) Fix a tuple of integers $(r_h)_{h \in H}$, one for each head. Suppose we now want to study the family $\mathcal{F}$ of tuples $M \in \mathrm{Hom}_G(V, W)^H$ such that each M_h has rank at most r_h. The same argument in (1) applies headwise, so, for each $M \in \mathcal{F}$, we can find $(U_h)_{h \in H}$, where

each U_h is a G-representation of dimension at most r_h, such that $M_h \in \mathcal{D}(V, W; U_h)$. Now, pack the U_h into $E := (U_h)_{h\in H}$, and define $\Gamma(E) := \bigoplus_{h\in H} U_h$, which is a G-representation. Let $\mathcal{D}(V, W; E)$ be the set of tuples $M \in \mathrm{Hom}_G(V, W)^H$ such that there exists $\phi_2 \in \mathrm{Hom}_G(V, \Gamma(E)), \phi_1 \in \mathrm{Hom}_G(\Gamma(E), W)$ satisfying $M_h = \phi_1 \mathrm{ev}_h \phi_2$, where $\mathrm{ev}_h : \Gamma(E) \to \Gamma(E)$ projects onto the h-th summand. Hence, we have $\mathcal{F} = \bigcup_E \mathcal{D}(V, W; E)$. Nothing fancy is done: we simply repackaged the U_h into one object, E, which is an equivariant bundle with trivial action on the base space H. This could seem somewhat ad-hoc, since each fiber U_h is really just an independent G-representation. In particular $\mathcal{D}(V, W; E) = \prod_{h\in H} \mathcal{D}(V, W; U_h)$, and $\mathcal{F}$ is also a product.

(3) Finally, let $H = G/K$ be a homogeneous space, and let $[e]$ $(= K)$ denote the identity coset of K. Suppose we want to study the family $\mathcal{F}$ of tuples $M \in \mathrm{Hom}(V, W)^H$ such that $h \mapsto M_h$ is G-equivariant and $\mathrm{rk}(M_h) \le r_h$ for each h. That is, $M_{gh} = gM_h g^{-1}$, which is exactly the equivariance condition in Theorem 1. The family $\mathcal{F}$ is not a product over H anymore because $M_{[e]}$ determines all the other M_h for $h \in G/K$. However, (1) still applies to $M_{[e]}$ for the subgroup K: because $M_{[e]} = M_{[ke]} = kM_{[e]}k^{-1}$, we have $M_{[e]} \in \mathrm{Hom}_K(\mathrm{Res}^G_K(V), \mathrm{Res}^G_K(W))$, so there is a K-representation of dimension at most $r_{[e]}$ such that

$$M_{[e]} \in \mathrm{Hom}_K(U, \mathrm{Res}^G_K(W))\mathrm{Hom}_K(\mathrm{Res}^G_K(V), U). \tag{25}$$

By Frobenius reciprocity,

$$\begin{aligned} \mathrm{Hom}_K(U, \mathrm{Res}^G_K(W)) &\cong \mathrm{Hom}_G(\mathrm{Ind}^G_K(U), W) \\ \mathrm{Hom}_K(\mathrm{Res}^G_K(V), U) &\cong \mathrm{Hom}_G(V, \mathrm{Ind}^G_K(U)) \end{aligned} \tag{26}$$

canonically, so any decomposition $M_{[e]} = \psi_1\psi_2$ with respect to Eq. (25) gives two G-equivariant maps: $\phi_1 \in \mathrm{Hom}_G(\mathrm{Ind}^G_K(U), W)$ and $\phi_2 \in \mathrm{Hom}_G(V, \mathrm{Ind}^G_K(U))$. Now, $\mathrm{Ind}^G_K(U)$ is nothing but the space of sections $\Gamma(E)$ of the associated bundle $E = G \times_K U$ (where we view $G \to G/K$ as a principal K-bundle), and it is easily shown (see Appendix D) that the entire tuple $(M_h)_h$ is given by $M_h = \phi_1 \mathrm{ev}_h \phi_2$, where $\mathrm{ev}_h : \Gamma(E) \to \Gamma(E)$ again projects onto the h-th fiber. The upshot is that $\mathcal{F} = \bigcup_E \mathcal{D}(V, W; E)$ still holds with the same definition of $\mathcal{D}(V, W; E)$, but now the union is taken over all G-equivariant vector bundles over $H = G/K$, which are in bijection with K-representations via $U \mapsto G \times_K U$. In this case, the rank constraint is equivalent to $\mathrm{rk}_h(E) := \dim(E_h) = \dim U \le r_h$.

In the context of Theorem 1, $V = W$ and we are given a partial G-clustering $\chi : H \to X \sqcup X_0$, and the objective is to study the family of tuples $(M_p)_{p\in X} \in \mathrm{End}(V)^X$ satisfying $M_{gp} = gM_p g^{-1}$ and $\mathrm{rank}(M_p) \le \dim V_h$ for all $h \in \chi^{-1}(p)$ (and similarly for the R_p with slightly different rank constraints). If $\chi = \mathrm{Id}$ (which means a trivial action on H with $X_0 = \varnothing$, see Example 4), then we are in case (2). In the general case, decompose X into G-orbits, and apply (3) to each orbit $O \subset X$ to get an equivariant bundle E_O, which can then be assembled into an equivariant bundle E over X, where the local ranks can depend on the orbit.

C.2 Step 0: Setup and Notations

We will start by fixing some notations. Recall the setting: the token feature space V is a finite-dimensional orthogonal G-representation with some given orthogonal decomposition $V = \bigoplus_{h\in H} V_h$. Assume $L \ge 3$.

For a G-equivariant vector bundle $E \to X$, define

$$\mathcal{D}(V; E) := \left\{(\phi_1 \mathrm{ev}_p \phi_2)_{p\in X} : \phi_1 \in \mathrm{Hom}_G(\Gamma(E), V), \phi_2 \in \mathrm{Hom}_G(V, \Gamma(E))\right\} \subset \mathrm{End}(V)^X. \tag{27}$$

This is the set of all elements M in $\mathrm{End}(V)^X$ that satisfy the equivariance condition $M_{gp} = gM_p g^{-1}$ and factor through the bundle E in the sense of (3) of Section C.1.

For $r : X \to \mathbb{N}_{\ge 0}$, define

$$\mathcal{D}_r(V; X) := \left\{M \in \mathrm{End}(V)^X : \mathrm{rank}(M_p) \le r(p), M_{gp} = gM_p g^{-1}\right\}, \tag{28}$$

which is the family of elements in $\mathrm{End}(V)^X$ that satisfy the same equivariance condition and are rank constrained by r, i.e., what we called $\mathcal{F}$ in Section C.1.

Let $\mathrm{Sub}^r(E)$ be the set of equivalence classes of maximal r-rank-bounded subbundles of E (Definition 3) up to bundle isomorphisms, where maximality is defined with respect to inclusions. Then, the argument in Section C.1 is formalized as

$$\mathcal{D}_r(V;X) = \bigcup_{E \in \mathrm{Sub}^r(X \times V)} \mathcal{D}(V;E). \tag{29}$$

See Lemma 4 in Appendix D for a proof.

Define

$$\begin{aligned} \mathcal{Z} &:= \{(\phi_2 \pi_h \phi_1)_{h \in H} : \phi_1, \phi_2 \in \mathrm{End}(V)\} = \{B \in \mathrm{End}(V)^H : \mathrm{rank}(B_h) \le \dim V_h\} \\ &\subset \mathrm{End}(V)^H \\ \mathcal{E} &:= \Big\{(M,R) \in \mathcal{Z} \times \mathcal{Z} : \mathrm{RMHSA}(-;M,R) : \mathbb{R}^L \otimes V \to \mathbb{R}^L \otimes V \text{ is } G\text{-equivariant}\Big\} \\ &\subset \mathrm{End}(V)^H \oplus \mathrm{End}(V)^H, \end{aligned} \tag{30}$$

That is, $\mathcal{E}$ is the subset of the RMHSA equivariance locus satisfying the rank constraints defining $\mathcal{Z}$.

For each partial G-clustering $\chi : H \to X_0 \sqcup X$, define the *attention width* a^χ and *value width* u^χ as

$$\begin{aligned} a^\chi : X &\to \mathbb{N}_{\ge 0} & \qquad u^\chi : X &\to \mathbb{N}_{\ge 0} \\ p &\mapsto \min_{h \in \chi^{-1}(p)} \dim V_h & \qquad p &\mapsto \sum_{h \in \chi^{-1}(p)} \dim V_h. \end{aligned} \tag{31}$$

Now, set

$$\mathcal{E}^\chi := \Big[\mathcal{P}_\chi^*\Big(\mathrm{End}(V)^{X_0} \times \mathcal{D}_{a^\chi}(V;X)\Big) \cap \mathcal{Z}\Big] \times \Big[\mathcal{P}_\chi^{-1}\Big(\{0\}^{X_0} \times \mathcal{D}_{u^\chi}(V;X)\Big) \cap \mathcal{Z}\Big], \tag{32}$$

where $\mathcal{D}_{a^\chi}(V;X)$ and $\mathcal{D}_{u^\chi}(V;X)$ are defined in Eq. (28). The set $\mathcal{E}^\chi$ is the subset of $\mathrm{End}(V)^H \oplus \mathrm{End}(V)^H$ consisting of the RMHSA parameters corresponding to the partial G-clustering χ, rank-constrained by the attention and value widths. Clearly, $\mathcal{E}^\chi = \mathcal{E}^{\chi'}$ if $\chi \cong \chi'$. By Theorem **??**, we have $\mathcal{E} = \bigcup_\chi \mathcal{E}^\chi$, where the union is taken over all partial G-clusterings $\chi : H \to X_0 \sqcup X$ up to isomorphism.

For $E \in \mathrm{Sub}^{a^\chi}(X \times V)$ and $E' \in \mathrm{Sub}^{u^\chi}(X \times V)$, define

$$\mathcal{E}^{\chi,E,E'} = \Big[\mathcal{P}_\chi^*\Big(\mathrm{End}(V)^{X_0} \times \mathcal{D}(V;E)\Big) \cap \mathcal{Z}\Big] \times \Big[\mathcal{P}_\chi^{-1}\Big(\{0\}^{X_0} \times \mathcal{D}(V;E')\Big) \cap \mathcal{Z}\Big]. \tag{33}$$

Proposition 1. $\mathcal{E}^\chi = \bigcup_{E \in \mathrm{Sub}^{a^\chi}(X \times V)} \bigcup_{E' \in \mathrm{Sub}^{u^\chi}(X \times V)} \mathcal{E}^{\chi,E,E'}$.

Proof. This is a straightforward consequence of Lemma 4, more precisely of Eq. (29). □

C.3 Step 1: Geometry of the Equivariance Locus in the RMHSA Parameter Space

In the following (especially in Section C.4), we will let $\mathfrak{R} : \mathrm{End}(V)^H \oplus \mathrm{End}(V)^H \to \{\mathbb{R}^L \otimes V \to \mathbb{R}^L \otimes V\}$ denote the realization map of RMHSA. Ultimately, we are interested in the class of functions $\mathfrak{R}(\mathcal{E})$. However, we will first work in $\mathrm{End}(V)^H \oplus \mathrm{End}(V)^H$, where algebraic geometry is more easily applicable.

Proposition 2.

(1) $\mathcal{E}^{\chi,E,E'} \subset \mathrm{End}(V)^H \oplus \mathrm{End}(V)^H$ is Zariski-closed. If χ is injective (hence bijective), then $\mathcal{E}^{\chi,E,E'}$ is irreducible.

(2) For a fixed χ, if (E_1, E_1') and (E_2, E_2') are not isomorphic, then neither $\mathcal{E}^{\chi,E_1,E_1'}$ nor $\mathcal{E}^{\chi,E_2,E_2'}$ contains the other.

(3) Consider the trivial G-clustering $\mathrm{Id} : H \to H = X$ (Example 4). For any $E \in \mathrm{Sub}^{a^{\mathrm{Id}}}(X \times V)$ and $E' \in \mathrm{Sub}^{u^{\mathrm{Id}}}(X \times V)$ without zero fibers (i.e. $\dim E_p, \dim E'_p > 0$ for all $p \in X = H$), the set $\mathcal{E}^{\mathrm{Id},E,E'}$ is an irreducible component (i.e., maximal irreducible subset) of $\mathcal{E}$.

Proof.
(1)
Since $\mathcal{P}^*_\chi$ is an injective linear map, the set $\mathcal{E}^{\chi,E,E'}$, as defined by Eq. (33), is closed by Lemma 5.

If χ is injective, then without loss of generality $\chi = \mathrm{Id}_H$, in which case $\mathcal{P}_\chi = \mathrm{Id}_H$. So Eq. (33) reduces to

$$\mathcal{E}^{\chi,E,E'} = \Big[\big(\mathrm{End}(V)^{X_0} \times \mathcal{D}(V;E)\big) \cap \mathcal{Z}\Big] \times \Big[\big(\{0\}^{X_0} \times \mathcal{D}(V;E')\big) \cap \mathcal{Z}\Big]. \tag{34}$$

This is irreducible because it is a Cartesian product of irreducible sets. More precisely, $\mathcal{Z} \subset \mathrm{End}(V)^{X_0 \sqcup X}$ (defined in Eq. (30)) is rectangular with respect to $\mathrm{End}(V)^{X_0} \times \mathrm{End}(V)^X$, and both $\mathcal{D}(V;E)$ and $\mathcal{D}(V;E')$ are irreducible by Lemma 5.

(2)
Suppose $\mathcal{E}^{\chi,E_1,E'_1} \subset \mathcal{E}^{\chi,E_2,E'_2}$. Then

$$\mathcal{P}^*_\chi\Big(\mathrm{End}(V)^{X_0} \times \mathcal{D}(V;E_1)\Big) \cap \mathcal{Z} \subset \mathcal{P}^*_\chi\Big(\mathrm{End}(V)^{X_0} \times \mathcal{D}(V;E_2)\Big) \cap \mathcal{Z}. \tag{35}$$

Apply $\mathcal{P}_\chi(\cdot)$ to both sides to get [4] $\mathcal{D}(V;E_1) \subset \mathcal{D}(V;E_2)$. Similarly, we have

$$\mathcal{P}^{-1}_\chi\Big(\{0\}^{X_0} \times \mathcal{D}(V;E'_1)\Big) \cap \mathcal{Z} \subset \mathcal{P}^{-1}_\chi\Big(\{0\}^{X_0} \times \mathcal{D}(V;E'_2)\Big) \cap \mathcal{Z}. \tag{36}$$

Again, apply $\mathcal{P}_\chi(\cdot)$ to get [5] $\mathcal{D}(V;E'_1) \subset \mathcal{D}(V;E'_2)$.

By Lemma 5, E_1 is isomorphic to a subbundle of E_2 and E'_1 is isomorphic to a subbundle of E'_2.

By maximality, we must have $E_1 \cong E_2$ and $E'_1 \cong E'_2$.

(3)
By (1), $\mathcal{E}^{\mathrm{Id},E,E'}$ is irreducible. It remains to show maximality. Suppose $\mathcal{E}^{\mathrm{Id},E,E'} \subset \mathcal{E}^{\tilde\chi,\tilde E,\tilde E'}$ for some partial G-clustering $\tilde\chi : H \to \tilde X_0 \sqcup \tilde X$. If $\mathrm{Id} \cong \tilde\chi$, that is, $\tilde X_0 = \varnothing$, $|\tilde X| = |H|$, and G acts on $\tilde X$ trivially, then we must have $(E,E') \cong (\tilde E, \tilde E')$ by (2), assuming $\tilde\chi = \mathrm{Id}$ without loss of generality. Thus, suppose $\mathrm{Id} \ncong \tilde\chi$. There are three (not mutually exclusive) possibilities:

(a) $\tilde X_0 \neq \varnothing$

(b) $\tilde\chi$ has a nontrivial fiber

(c) $\tilde X \neq \varnothing$ and the G-action on $\tilde X$ is not trivial.

In each of these three cases, we can show $\mathcal{E}^{\mathrm{Id},E,E'} \setminus \mathcal{E}^{\tilde\chi,\tilde E,\tilde E'}$ is not empty:

In case (a), we can find $(M,R) \in \mathcal{E}^{\mathrm{Id},E,E'}$ such that $\sum_{h \in \tilde\chi^{-1}(\tilde X_0)} R_h \neq 0$.
In case (b), suppose $\tilde\chi(h) = \tilde\chi(h')$ with $h \neq h'$. Find $(M,R) \in \mathcal{E}^{\mathrm{Id},E,E'}$ such that $M_h \neq M_{h'}$.
In case (c), we may assume $\tilde X_0 = \varnothing$ and $\tilde\chi$ is injective, and suppose $g \cdot p = p'$ with $p \neq p' \in \tilde X$. Then simply take $(M,R) \in \mathcal{E}^{\mathrm{Id},E,E'}$ such that $gM_{\tilde\chi^{-1}(p)}g^{-1} \neq M_{\tilde\chi^{-1}(p')}$. □

[4] This needs some care. $\mathcal{Z} = \mathcal{Z}_0 \times \mathcal{Z}_1 \subset \mathrm{End}(V)^{X_0} \times \mathrm{End}(V)^X$ is rectangular, and the projection can be also split accordingly into $\mathcal{P}_0 = \mathcal{P}_\chi|_{\chi^{-1}(X_0)}, \mathcal{P}_1 = \mathcal{P}_\chi|_{\chi^{-1}(X)}$ (slight abuse of notation). Then use $\mathcal{D}(V;E_1) \cap \mathcal{Z}_1 = \mathcal{D}(V;E_1)$ and the injectivity of $\mathcal{P}_1\mathcal{P}_1^*$.

[5] Similarly, the left hand side is $(\cdots) \times [\mathcal{P}_1^{-1}(\mathcal{D}(V;E'_1)) \cap \mathcal{Z}_1]$, and $\mathcal{P}_1(\cdot)$ on the second factor recovers $\mathcal{D}(V;E'_1)$.

Corollary 1. *Let $\beta : \Lambda \to \mathrm{End}(V)^4$ be a polynomial map, where Λ is a finite-dimensional real vector space. Define*

$$\mathcal{E}^\beta := \left\{((\phi_q^*\pi_h\phi_k)_{h\in H}, (\phi_o\pi_h\phi_v)_{h\in H}) : (\phi_q,\phi_k,\phi_v,\phi_o) \in \beta(\Lambda)\right\} \subset \mathrm{End}(V)^H \oplus \mathrm{End}(V)^H. \tag{37}$$

If $\mathcal{E}^\beta \subset \mathcal{E}$, then there is a triple (χ, E, E') such that $\mathcal{E}^\beta \subset \mathcal{E}^{\chi,E,E'}$.

Proof. $\mathcal{E}^\beta \subset \mathcal{E}$ implies $\mathcal{E}^\beta = \bigcup_{(\chi,E,E')} \mathcal{E}^{\chi,E,E'} \cap \mathcal{E}^\beta$. Since the $\mathcal{E}^{\chi,E,E'}$ are Zariski-closed, each $\mathcal{E}^\beta \cap \mathcal{E}^{\chi,E,E'}$ is closed in $\mathcal{E}^\beta$. But $\mathcal{E}^\beta$ is irreducible: it is the image of an irreducible set under a polynomial map. Thus, $\mathcal{E}^\beta \subset \mathcal{E}^{\chi,E,E'}$ for some (χ, E, E'). □

Corollary 2. *The set $\mathcal{E}$ has at least $|\mathrm{Sub}_+^{(\dim V_h)_{h\in H}}(H \times V)|^2$ irreducible components, where the subscript "+" means bundles with zero fibers are excluded.*

Note that $\mathrm{Sub}_+^{(\dim V_h)_{h\in H}}(H \times V) \cong \prod_{h\in H} \mathrm{Sub}_+^{\dim V_h}(V)$, where we think of V as the G-equivariant vector bundle over one point. Equivalently, $\mathrm{Sub}_+^r(V)$ consists of equivalence classes of maximal G-subrepresentations of dimension at most r. In the main text, $|\mathrm{Sub}_+^r(V)|$ is denoted by $c_{G,r}(V)$.

Example 10.

(1) Let $G = C_3$ act on $\mathbb{R}^2$ by $r \mapsto R(2\pi/3)$, which is irreducible (and equivalent to E_1). Then $\mathrm{Sub}_+^1(\mathbb{R}^2) = \varnothing$ because there are no one-dimensional subrepresentations.

(2) For the regular representation $\mathbb{R}[D_4]$ of the dihedral group D_4, which decomposes as $\mathrm{A}_1 \oplus \mathrm{A}_2 \oplus \mathrm{B}_1 \oplus \mathrm{B}_2 \oplus 2\mathrm{E}_1$, we have

$$\begin{aligned}&\mathrm{Sub}^2(\mathbb{R}[D_4]) = \mathrm{Sub}_+^2(\mathbb{R}[D_4])\\ &= \{\mathrm{A}_1\oplus\mathrm{A}_2, \mathrm{A}_1\oplus\mathrm{B}_1, \mathrm{A}_1\oplus\mathrm{B}_2, \mathrm{A}_2\oplus\mathrm{B}_1, \mathrm{A}_2\oplus\mathrm{B}_2, \mathrm{B}_1\oplus\mathrm{B}_2, \mathrm{E}_1\}.\end{aligned} \tag{38}$$

C.4 Step 2: Passing to the Function Space

Corollary 2 is a statement made in the parameter space of RMHSA. Transferring this to the function space requires modding out permutations:

Proposition 3. *Suppose $\mathfrak{R}(\mathcal{E}^{\mathrm{Id},E,E'}) \subset \bigcup_{k=1}^N \mathfrak{R}(\mathcal{E}^{\chi^k,\tilde{E}^k,\tilde{E}'^k})$, with $E, E' \in \mathrm{Sub}_+^{(\dim V_h)_{h\in H}}(H \times V)$ and $\tilde{E}^k \in \mathrm{Sub}^{a^{\chi^k}}(X^k \times V), \tilde{E}'^k \in \mathrm{Sub}^{u^{\chi^k}}(X^k \times V)$. Then there is a $k \in \{1,\dots,N\}$ and a permutation $\nu \in \mathrm{Aut}(H)$ such that $\chi^k \cong \mathrm{Id}$ and $E \hookrightarrow \tilde{E}^{k,\nu}$ and $E' \hookrightarrow \tilde{E}'^{k,\nu}$, where $\tilde{E}^{k,\nu}, \tilde{E}'^{k,\nu}$ denote the permuted bundles.*

If all heads have the same dimension $d := \dim V/|H| = \dim V_h$, then the same statement holds with $E \cong \tilde{E}^{k,\nu}$ and $E' \cong \tilde{E}'^{k,\nu}$.

(Note: we have implicitly identified χ^k with Id*, which is harmless because there is only one isomorphism.)*

Proof. Let

$$U := (\mathcal{D}(V;E) \setminus \mathcal{Q}) \times (\mathcal{D}(V;E') \setminus \mathcal{Q}'), \tag{39}$$

where

$$\begin{aligned}\mathcal{Q} &:= \{M : \exists h \neq h'\ M_h = M_{h'}\} \cap \mathcal{D}(V;E)\\ \mathcal{Q}' &:= \{R : \exists h : R_h = 0\} \cap \mathcal{D}(V;E').\end{aligned} \tag{40}$$

Then U is dense in $\mathcal{D}(V;E) \times \mathcal{D}(V;E') = \mathcal{E}^{\mathrm{Id},E,E'}$, since E and E' do not have zero fibers.

Take $(M,R) \in U$. There is some k for which $\mathfrak{R}(M,R) \in \mathfrak{R}(\mathcal{E}^{\chi^k,\tilde{E}^k,\tilde{E}'^k})$. By Theorem 3, for any $\tilde{M}, \tilde{R} \in \mathrm{End}(V)^H$, a necessary condition for $\mathfrak{R}(M,R) = \mathfrak{R}(\tilde{M},\tilde{R})$ is $|\{\tilde{M}_h : \tilde{R}_h \neq 0\}| = |H|$.

Take any $(\tilde{M},\tilde{R}) \in \mathcal{E}^{\chi^k,\tilde{E}^k,\tilde{E}'^k}$. If χ^k is not injective, then $|\{\tilde{M}_h : \tilde{R}_h \neq 0\}| \leq |X^k| < |H|$. Thus, suppose χ^k is injective. If $X_0^k \neq \varnothing$, then $|\{\tilde{M}_h : \tilde{R}_h \neq 0\}| = |X^k| < |H|$. Finally, assume χ^k is

injective and $X_0^k = \varnothing$, but $gp = p'$ for distinct points $p, p' \in X^k$ and $g \in G$. By Theorem 3, there must exist an $h \in H$ such that $\tilde{M}_{(\chi^k)^{-1}(p)} = M_h$. But then $\tilde{M}_{(\chi^k)^{-1}(p')} = g\tilde{M}_{(\chi^k)^{-1}(p)}g^{-1} = M_h = \tilde{M}_{(\chi^k)^{-1}(p)}$, so again $|\{\tilde{M}_h : \tilde{R}_h \neq 0\}| < |H|$. In other words, we must have $\chi^k \cong \mathrm{Id}$. Again by Theorem 3, there is a permutation $\nu \in \mathrm{Aut}(H)$ such that $(M, R) \in \mathcal{E}^{\mathrm{Id}, \tilde{E}^{k,\nu}, \tilde{E}'^{k,\nu}}$, where $\tilde{E}^{k,\nu}$ denotes the permuted bundle.

Hence, we have shown

$$U \subset \bigcup_{k:\chi^k \cong \mathrm{Id}}^{N} \bigcup_{\nu \in \mathrm{Aut}(H)} \mathcal{E}^{\chi^k, \tilde{E}^{k,\nu}, \tilde{E}'^{k,\nu}}. \tag{41}$$

Since $\mathcal{D}(V; E), \mathcal{D}(V; E')$ are irreducible (Lemma 5(1)), taking the closure of both sides yields

$$\mathcal{E}^{\mathrm{Id}, E, E'} \subset \bigcup_{k:\chi^k \cong \mathrm{Id}}^{N} \bigcup_{\nu \in \mathrm{Aut}(H)} \mathcal{E}^{\chi^k, \tilde{E}^{k,\nu}, \tilde{E}'^{k,\nu}}. \tag{42}$$

Again by irreducibility, $\mathcal{E}^{\mathrm{Id}, E, E'} \subset \mathcal{E}^{\chi^k, \tilde{E}^{k,\nu}, \tilde{E}'^{k,\nu}}$ for some k and ν, which is equivalent to $\mathcal{D}(V; E) \subset \mathcal{D}(V; \tilde{E}^{k,\nu})$ and $\mathcal{D}(V; E') \subset \mathcal{D}(V; \tilde{E}'^{k,\nu})$. By Lemma 5(2), $E \hookrightarrow \tilde{E}^{k,\nu}$ and $E' \hookrightarrow \tilde{E}'^{k,\nu}$.

If the head dimensions are equal, then $E \cong \tilde{E}^{k,\nu}$ and $E' \cong \tilde{E}'^{k,\nu}$ by maximality of E and E' among d-rank-bounded subbundles of $H \times V$. □

Finally, we prove Theorem 2. Let $d := \dim V/|H|$ be the uniform head dimension. Recall that the (rank-bounded) equivariance locus $\mathcal{E}$ in the RMHSA parameter space is the union of the $\mathcal{E}^{\chi,E,E'}$, which are Zariski-closed. Express this as $\mathcal{E} = \bigcup_{a \in A} \mathcal{E}_a$. Proposition 3 identifies a subset $A_0 \subset A$, namely those with $\chi \cong \mathrm{Id}$ and $E, E' \in \mathrm{Sub}_+^d(H \times V)$, with the following property: there is an $\mathrm{Aut}(H)$-action on A_0 such that if $\mathfrak{R}(\mathcal{E}_a) \subset \bigcup_{c \in C} \mathfrak{R}(\mathcal{E}_c)$ with $a \in A_0$ and finite $C \subset A$, then we can find a $c \in A_0 \cap C$ such that a, c are in the same $\mathrm{Aut}(H)$-orbit.

Now, let $(\beta\text{-MHSA})_{\beta \in \mathbf{B}}$ be a finite (otherwise there is nothing to prove) G-complete family of polynomially parameterized MHSA. By Corollary 1, there is a map $J : \mathbf{B} \to A$ such that $\mathcal{E}^\beta \subset \mathcal{E}_{J(\beta)}$. By G-completeness, we have $\bigcup_{\beta \in \mathbf{B}} \mathfrak{R}(\mathcal{E}_{J(\beta)}) = \bigcup_{a \in A} \mathfrak{R}(\mathcal{E}_a)$. That is, for any $a \in A_0$, we have $\mathfrak{R}(\mathcal{E}_a) \subset \bigcup_{\beta \in \mathbf{B}} \mathfrak{R}(\mathcal{E}_{J(\beta)})$. Proposition 3 then implies there is a β such that $J(\beta) \in A_0$ and $a, J(\beta)$ are in the same $\mathrm{Aut}(H)$-orbit. That is,

$$A_0 \subset \mathrm{Aut}(H) \cdot (J(\mathbf{B}) \cap A_0), \tag{43}$$

or,

$$|A_0| \leq |H|! \cdot |J(\mathbf{B}) \cap A_0| \leq |H|! \cdot |J(\mathbf{B})| \leq |H|! \cdot |\mathbf{B}|. \tag{44}$$

That is,

$$|\mathbf{B}| \geq \frac{1}{|H|!} |\mathrm{Sub}_+^d(H \times V)|^2 = \frac{1}{|H|!} c_{G,d}(V)^{2|H|}. \tag{45}$$

C.5 Derivation of the numerical bound for $G = D_4$

For $G = D_4$ acting on $\mathbb{R}^C \otimes \mathbb{R}[D_4]$ and eight C-dimension attention heads, the bound is $8!^{-1} c_{D_4,C}(\rho_{\mathrm{reg}}^{\oplus C})^{16}$.

$c_{D_4,C}(\rho_{\mathrm{reg}}^{\oplus C})$ is the number of C-dimensional subrepresentations up to isomorphism. So

$$c_{D_4,C}(\rho_{\mathrm{reg}}^{\oplus C}) = \binom{C+3}{3} + \binom{C+1}{3} + \cdots = \Theta(C^4). \tag{46}$$

That is, the bound is asymptotically $8!^{-1}\Theta(C^4)^{16} = \Theta(C^{64})$. For even C, this lower bound is

$$\frac{1}{8!}\left[\frac{(C+2)(C+4)(C^2+6C+6)}{48}\right]^{16}.$$

For $C = 96$ (total token dimension $= 8C = 768$), it is about 1.6×10^{96}.

D EQUIVARIANT VECTOR BUNDLES

Here, we review the notion of equivariant vector bundles, which are used in Appendix C to decompose the equivariance locus in the parameter space of RMHSA. For our purpose, it is enough to consider finite G-sets as base spaces.

While we worked with an arbitrary group G in Appendix B and Theorem 1, we will assume that G is finite in this section.

Definition 3. *A G-**equivariant vector bundle** is a G-equivariant surjective map $\pi : E \to X$, where E, X are G-sets, X is finite, $E_p := \pi^{-1}(p)$ is a finite-dimensional vector space for each $p \in X$, and $g : E_p \to E_{gp}$ is linear.*

1. *A **section** is a map $s : X \to E$ such that $\pi \circ s = \mathrm{Id}_X$. The space of sections is denoted by $\Gamma(E)$ and is isomorphic to $\bigoplus_{p\in X} E_p$. It is a G-representation through the action $(g \cdot s)(p) = g \cdot s(g^{-1}p)$. For $p \in X$, define the **evaluation map** as $\mathrm{ev}_p : \Gamma(E) \to E_p, s \mapsto s(p)$. Identifying E_p with the corresponding subspace of $\Gamma(E)$, we may think of ev_p as belonging to $\mathrm{End}(\Gamma(E))$.*
2. *The **local rank** is $\mathrm{rk}_p(E) := \dim(E_p)$, and depends only on the orbit of $p \in X$.*
3. *If E' is another G-equivariant vector bundle over X, a **bundle homomorphism** is a G-equivariant map $\varphi : E \to E'$ such that $\varphi(E_p) \subset E'_p$ and $\varphi|_{E_p}$ is linear.*
4. *A **subbundle** is a G-stable subset $D \subset E$ such that $D_p := D \cap E_p$ is a linear subspace of E_p for each $p \in X$.*
5. *Given a prescribed function $r : X \to \mathbb{N}_{\geq 0}$, a bundle $\pi : E \to X$ is said to be r-**rank-bounded** if $\mathrm{rk}_p(E) \leq r(p)$ for all $p \in X$.*

We will always assume that each fiber E_p is equipped with an inner product such that the induced inner product on $\Gamma(E)$ is G-invariant.

Recall that $\mathrm{Sub}^r(E)$ is the set of equivalence classes of maximal r-rank-bounded subbundles of E up to bundle isomorphisms (see Section C.2).

Example 11. *For any G-set X and any G-representation V, we can construct the product bundle $\pi_1 : X \times V \to X$, where G acts on both factors in $X \times V$.*

Example 12. *Let $K \subset G$ be a subgroup and let W be a K-representation. Let $E := G \times_K W = \{[g, w] : g \in G, w \in W\}$ where $[g, w] = [gk, k^{-1}w]$ for $k \in K$. Then $\pi : E \to G/K, [g, w] \mapsto gK$ is a G-equivariant vector bundle, and $\Gamma(E) \cong \mathrm{Ind}_K^G(W)$. For any G-representation V, there is a canonical isomorphism $\mathrm{Hom}_K(V, W) \cong \mathrm{Hom}_G(V, \Gamma(E))$ given as follows: if $f \in \mathrm{Hom}_K(V, W)$, define $\tilde{f} \in \mathrm{Hom}(V, \Gamma(E))$ by setting $\tilde{f}(v)(gK) = gf(g^{-1}v) \in E_{gK}$. This is well-defined: if $g' = gk$, then $g'fg'^{-1} = gkfk^{-1}g^{-1} = gfg^{-1}$ because f is K-equivariant. It is easy to check that the map $\tilde{f}$ is G-equivariant. Similarly, we also have adjunction in the other direction $\mathrm{Hom}_K(W, V) \cong \mathrm{Hom}_G(\Gamma(E), V)$.*

Lemma 3. *The map $X \to \mathrm{End}(\Gamma(E)), p \mapsto \mathrm{ev}_p$ is G-equivariant.*

Proof. Take a section $s \in \Gamma(E)$. Then

$$\begin{aligned} \mathrm{ev}_{gp}(s) &= s(gp) = gg^{-1}s(gp) = g(g^{-1} \cdot s)(p) \\ &= g \cdot \mathrm{ev}_p(g^{-1} \cdot s) = [g\mathrm{ev}_p g^{-1}](s). \end{aligned} \tag{47}$$

Since s was arbitrary, we have $\mathrm{ev}_{gp} = g\mathrm{ev}_p g^{-1}$. □

The motivation to consider equivariant vector bundles is the following lemma:

Lemma 4. *Let $M : X \to \mathrm{End}(V), p \mapsto M_p$ and $r : X \to \mathbb{N}_{\geq 0}$ an arbitrary function. The following are equivalent:*

(1) The map $p \mapsto M_p$ is G-equivariant, and $\mathrm{rank}(M_p) \leq r(p)$ for all $p \in X$.

(2) There exists a maximal r-rank-bounded subbundle $E \in \mathrm{Sub}^r(X \times V)$ and equivariant linear maps $\psi_1 \in \mathrm{Hom}_G(V, \Gamma(E)), \psi_2 \in \mathrm{Hom}_G(\Gamma(E), V)$ such that $M_p = \psi_2 \mathrm{ev}_p \psi_1$ for all $p \in X$.

More concisely (this is Eq. (29)*),*

$$\mathcal{D}_r(V; X) = \bigcup_{E \in \mathrm{Sub}^r(X \times V)} \mathcal{D}(V; E). \tag{48}$$

Proof. We may assume $X = G/K$ is a homogeneous space. The general statement is recovered by considering each orbit separately. In this case, r is just a nonnegative integer.
(1) $\Rightarrow$ (2)
Let $p_0 \in X = G/K$ be the coset of the identity of G (i.e. $p_0 = K$ but we use the notation p_0 to emphasize that we think of it as a point). Then $M_{p_0} \in \mathrm{End}_K(V)$. Let $W \subset V$ be a maximal K-subrepresentation of dimension at most r containing $\mathrm{im}(M_{p_0})$. We can then find $f_1 \in \mathrm{Hom}_K(V, W), f_2 \in \mathrm{Hom}_K(W, V)$ such that $f_2 f_1 = M_{p_0}$. Let $E := G \times_K W$ as in Example 12. Set $\psi_2 = \tilde{f}_2 \in \mathrm{Hom}_G(\Gamma(E), V)$ and $\psi_1 = \tilde{f}_1 \in \mathrm{Hom}_G(V, \Gamma(E))$ (see Example 12), for which $M_p = \psi_2 \mathrm{ev}_p \psi_1$. Moreover, $\mathrm{rk}(E) = \dim(W) \leq r$, so $E \in \mathrm{Sub}^r(X \times V)$.

(2) $\Rightarrow$ (1)
By Lemma 3, we have $\psi_2 \mathrm{ev}_{gp} \psi_1 = \psi_2 g \mathrm{ev}_p g^{-1} \psi_1 = g \psi_2 \mathrm{ev}_p \psi_1 g^{-1}$. □

Lemma 5.

(1) $\mathcal{D}(V; E)$ is Zariski-closed in $\mathrm{End}(V)^X$ and irreducible.

(2) If E, E' are subbundles of $X \times V$ and $\mathcal{D}(V; E') \subset \mathcal{D}(V; E)$, then E' is isomorphic to a subbundle of E.

Proof.
(1)
$\mathcal{D}$ is irreducible because it is the image of an affine space under a polynomial map.

We may assume $X = G/K$ is a homogeneous space; the general case follows from considering each orbit separately and the fact that the Cartesian product of two Zariski-closed sets is closed.

Let $W := E_K$ be the fiber above the coset of the identity. Consider the map

$$\begin{aligned} \delta : \mathrm{End}_K(V) &\to \mathrm{End}(V)^X \\ M &\mapsto (gMg^{-1})_{gK \in X}. \end{aligned} \tag{49}$$

The map δ is a biregular isomorphism from $\mathrm{End}_K(V)$ onto $\delta(\mathrm{End}_K(V))$, and $\delta(\mathrm{End}_K(V)) \subset \mathrm{End}(V)^X$ is closed. Moreover, we have $\delta(D) = \mathcal{D}$, where $D := \mathrm{Hom}_K(W, V)\mathrm{Hom}_K(V, W) \subset \mathrm{End}_K(V)$. It is easy to check that D is closed (decompose V, W into irreps of K, then D consists of exactly those equivariant endomorphisms with per-irrep rank constraints). Therefore, also $\mathcal{D} = \delta(D)$ is closed in $\mathrm{End}(V)^X$.

(2)
Again, we may assume $X = G/K$. Let W, W' be the fibers over K of E and E' respectively. Since $E, E' \subset X \times V$, the fibers W, W' are subrepresentations of $\mathrm{Res}^G_K(V)$. Then $\mathcal{D}(V; E') \subset \mathcal{D}(V; E)$ is equivalent to $\mathrm{Hom}_K(W', V)\mathrm{Hom}_K(V, W') \subset \mathrm{Hom}_K(W, V)\mathrm{Hom}_K(V, W)$. Clearly, this can only happen if W' embeds K-equivariantly into W, from which it follows that E' is isomorphic to a subbundle of E. □

E Experiment Details

E.1 Teacher-Student

For the teacher-student experiment involving pairs of D_4-equivariant attention layers as well as the unconstrained (vanilla) attention shown in Fig. 4, the hyperparameters are as follows:

- feature space $\mathbb{R}^{D_{\text{in}}} = \mathbb{R}^{D_{\text{out}}} = \mathbb{R}^8$ with the regular representation ($\mathbb{R}[D_4]$)
- number of heads $H = 2$
- head dimension $d = 4$
- token count $L = 10$
- total trainin steps $= 4000$
- learning raite: Cosine Annealing LR with $\eta_{\text{max}} = 10^{-2}$ and $\eta_{\text{min}} = 10^{-4}$
- training bach size $B_{\text{train}} = 64$

In total, we implement six 2-head equivariant attention architectures with four-dimensional heads. According to Theorem 1, these are indexed by (1) a G-action on the heads, and (2) for each orbit $O \subset \{1, 2\}$, a maximal subreprsentation of $\text{Res}^{D_4}_{\text{Stab}(p_O)}(\mathbb{R}[D_4])$ of dimension at most four, where p_O is a chosen base point and $\text{Stab}(p_O)$ is the stabilizer subgroup. In our case, the maximal subrepresentations all happen to saturate the dimension constraint (they are all four-dimensional). We consider two actions:

- $D_4/D_4 \sqcup D_4/D_4$: the trivial action (two orbits each consisting of one point).
- D_4/C_4: for this action, s (reflection) swaps the two heads.

The architectures for the trivial action are then indexed by a pair of four-dimensional subrepresentations of $\mathbb{R}[D_4]$. For D_4/C_4, they are indexed by four-dimensional subrepresentations of $\text{Res}^{D_4}_{C_4}[\mathbb{R}[D_4]] = \mathbb{R}^2 \otimes \mathbb{R}[C_4] \cong 2\text{A} \oplus 2\text{B} \oplus 2\text{E}_1$.

The numbers reported in Fig. 4's heatmap are relative MSE loss, defined as

$$l(t, s) := \frac{1}{E_t} \frac{1}{B_{\text{eval}} LC} \sum_{b,l,c} [t(x)_{b,l,c} - s(x)_{b,l,c}]^2, \tag{50}$$

where t is the teacher, s is the student, and $E_t := \frac{1}{B_{\text{eval}} LC} \sum_{b,c,l} t(x)^2_{b,l,c}$. Here, the x is a batch of $B_{\text{eval}} = 4096$ random iid Gaussian token sequences, fixed across all runs.

We run five seeds for each teacher-student pair and report the the smallest $l(s, t)$ across the five runs.

E.2 Modular Addition

We train a single layer of unconstrained (non-equivariant) attention to learn the map

$$(a, b, c) \mapsto (b + c, c + a, a + b), \tag{51}$$

where addition is defined modulo p with $p = 24, 15, 17$ (with the results for $p = 17$ shown below). For example, $(20, 4, 13) \mapsto (17, 9, 0)$ for $p = 24$.

The input is one-hot encoded in $\mathbb{R}^p$, and the layer's output is centered. That is, the component along the direction $(1, \ldots, 1)^\top \in \mathbb{R}^p$ is projected out. The group C_{24} acts on $\mathbb{R}^{D_{\text{in}}} = \mathbb{R}^{D_{\text{out}}} = \mathbb{R}^p$ by cyclically permuting the entries. The action on the output space is "doubled", in the sense that $\rho_{\text{out}}(g) = \rho_{\text{in}}(g)^2$.

For $p = 24$, there are 5,760 training examples, each consisting of a pair of triples $((a, b, c), (b+c, c+a, a+b))$. Each test item is a triple $((a, b, c), j, (b + c, c + a, a + b)_j)$, for example $((20, 4, 3), 0, 7)$, where the model is asked to predict the j-th pairwise sum of the input. There are 3,456 test items in total. For $p = 15$ the number of training tuples and the number of test items are 1,395 and 1,080. For $p = 17$, this numbers are 2,261/1,496.

The attention layer has $H = 32$ heads, each with head dimension $d = 128$. It is trained by AdamW with $\beta = (0.9, 0.98)$ and weight decay 1.0, with each step consisting of one single batch with all examples. The $p = 24$ model is trained for 6,000 steps, while the $p = 15$ and $p = 17$ models are trained for 40,000 steps.

The learning rate linearly increases to 10^{-3} within the first 10 epochs, then stays constant afterwards. The total loss is $l_{\text{task}} + l_{\text{eq}}$, where l_{task} is cross-entropy for 24 classes, and the equivariance loss is

calculated as follows: in each training step, $B = 64$ triples of i.i.d. Gaussian tokens are drawn from $\mathcal{N}(0, \mathbb{1})$. I.e., a random $x \in \mathbb{R}^{B\times 3\times p}$ with each entry i.i.d. from $\mathcal{N}(0, 1)$. For each $b \in \{0, \dots, B-1\}$, randomly draw $S = 24$ group elements $g_b^{(s)} \in C_p$, $s = \{0, \dots, S-1\}$. Then set

$$l_{\text{eq}} := \frac{\frac{1}{B(S-1)} \sum_{b,s} \|z_{bl}^{(s)} - \bar{z}_{bl}\|^2}{\frac{1}{BS-1} \sum_{b,s} \|z_{bl}^{(s)} - \bar{z}_l\|^2}, \tag{52}$$

where

$$z_{bl}^{(s)} := \rho_{\text{out}}(g_b^{(s)})^{-1} F_j(\rho_{\text{in}}(g_b^{(s)}) x_b), \qquad \bar{z}_{bl} := \frac{1}{S} \sum_s z_{bl}^{(s)}, \quad \bar{z}_l := \frac{1}{BS} \sum_{b,s} z_{bl}^{(s)}. \tag{53}$$

For $p = 24$, test accuracy reaches 100% at all three output token positions at step $6{,}000$. While for $p = 15$ the test accuracy is still 0% at step $40{,}000$.

For $p = 17$, the final test accuracy is 95%. In Fig. 9, we carry out the same analysis as we did for $p = 24$, this time for both M and R but projecting only on irrep planes. The irreps of C_{17} are W_0 (trivial) and $W_\omega, 1 \le \omega \le 8$ (2D), and $\mathbb{R}^{17}$ again decomposes as $W_0 \oplus W_1 \oplus \cdots \oplus W_8$. For the R_h, the output frequencies are doubled, but we still label with the original frequencies. So, for the R_h, the subspace $W_{\omega_i} \otimes W_{\omega_o}$ contains of $W_{\omega_i - 2\omega_o}$ and $W_{\omega_i + 2\omega_o}$.

This time, the 32 heads organized themselves into one C_{17}-orbit of size 17, one "pseudo-orbit" of size 8, and two singleton orbits.

The size-17 orbit actually consists of 22 heads, 10 of each paired up to form 5 two-head clusters. If we sum the respective pairs $(R_h, R_{h'})$, we see in the bottom-left panels of Fig. 9 that they do land on a size-17 orbit. However, this description is not exact: none of the head-pairs share the same QK matrix. In fact, the distance between each pair of heads in the W_2 plane of Fig. 9 seems to be constant across all five pairs.

Also note that the family of eight heads (orange panels) do *not* form an actual orbit of C_{17} (any orbit must necessarily have 1 or 17 elements). Nonetheless, they still form regular polygons in the projected irrep planes, similarly to the square pseudoorbit in the $p = 15$ run (Fig. 6. We believe that this can be explained as follows: even though the octagon is not stable under rotation by $2\pi \cdot 7/17$ (which would be necessary for equivariance by Theorem 1), it is approximately a circle, which *is* stable under any rotation. So, the octagonal arrangement might be a local minimum for the equivariance loss, and once the eight heads formed such structure, finding a "better one" might require the model to sacrifice a lot of equivariance and//or task accuracy first. This explanation is not as convincing for the square, but the W_3 where this happens has very little overlap (Frobenius) with the R_h.

E.3 Vector Fields

E.3.1 Data

We generate vector fields $\mathbb{R}^2 \to \mathbb{R}^2$ of the form

$$\mathbf{v}(\mathbf{x}) = \sum_{a=1}^{M} \mathbf{u}_a \cos(\mathbf{k}_a \cdot \mathbf{x} + \phi_a) \tag{54}$$

where:

- ϕ_a is a random phase.
- $\mathbf{u}_a$ is a random amplitude vector with each component drawn from $\mathcal{N}(0, \sigma^2 = 0.06)$.
- each wave vector $\mathbf{k}_a$ is defined to be $2\pi\lambda_a^{-1}\mathbf{s}_a$, where λ_a is a wavelength uniformly sampled from $[64, 160]$ and $\mathbf{s}_a$ is a random unit vector.

In our experiments, we take $M = 6$ and generate $40{,}000$ training, $1{,}000$ validation, and $1{,}000$ testing vector fields. Each generated vector field $\mathbf{v}$ is represented by a 256×256 image with one arrow drawn on each 16×16 patch representing the value of the vector field at the center of the patch.

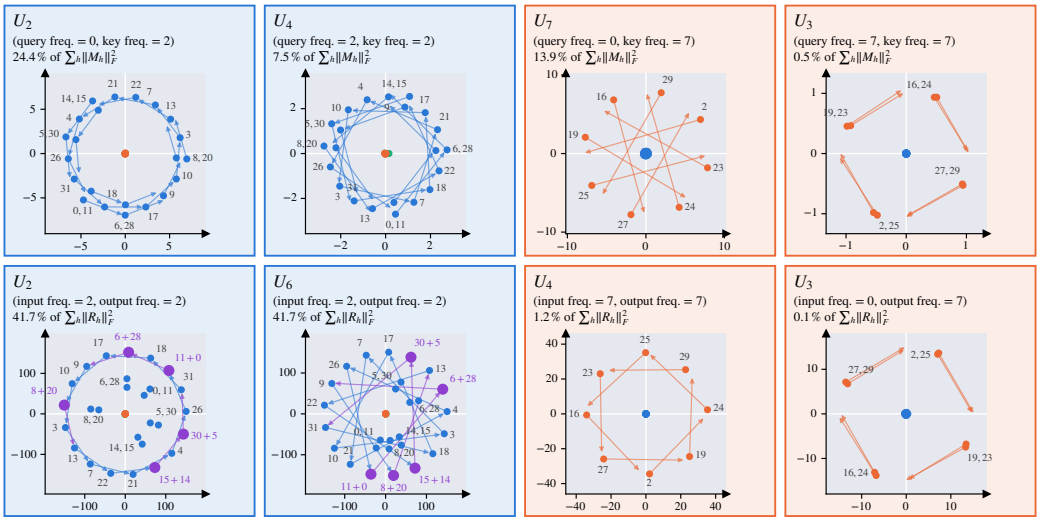

Figure 9: M_h (top row) and R_h (bottom row) for the $p = 17$ run. In the first two R plots, we sum $R_h + R_{h'}$ for two heads if the respective $M_h, M_{h'}$ are close. The sum is then shown as one large purple dot.

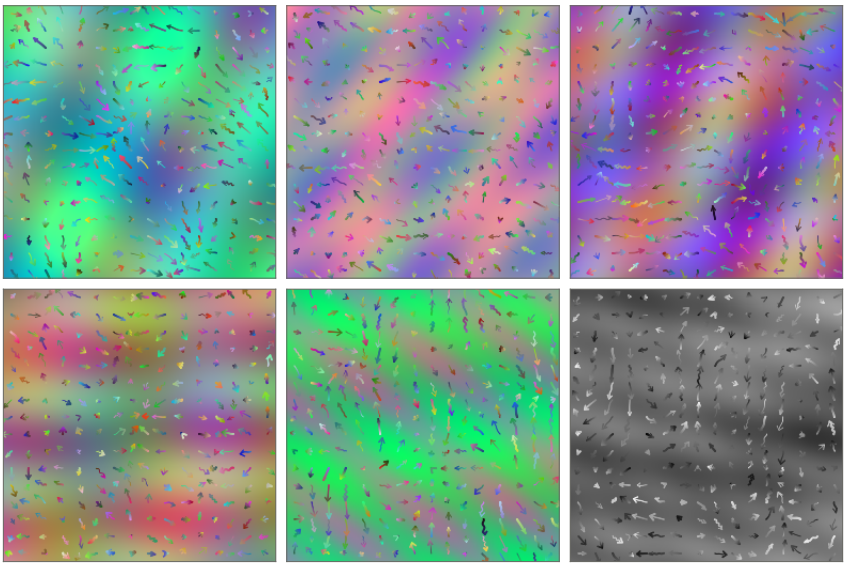

Figure 10: Six training images for the vector field task.

We allow the arrows to extend to nearby patches, just to make the task more difficult. Random color perturbations, both in the drawn arrows and in the background, are added to each image (see Fig. 10).

The target features to be learned by the model are the two components of the vector field (v_x, v_y) and the four components of the derivative $\partial_\mu v_\nu$, where $\mu, \nu \in \{x, y\}$, evaluated at the center of each 16×16 patch, i.e., at $\{-120, -104, -88, \ldots, 88, 104, 120\}^2 \subset \mathbb{R}^2$. In our dataset, we already decompose these four components into the four 1D irreps of D_4, each of which has a clear physical meaning (see Table 1). However, whether the features are decomposed into irreps does not matter mathematically since mapping $\nabla \mathbf{v}$ to the four irrep components is linear, which can be learned by the attention layer. The task of mapping an input image to $(\mathbf{v}, \nabla \mathbf{v})$ is D_4-equivariant.

target	formula	irrep
vector field	(v_x, v_y)	E_1
divergence	$\partial_x v_x + \partial_y v_y$	A_1
curl	$\partial_x v_y - \partial_y v_x$	A_2
axial shear	$\partial_x v_x - \partial_y v_y$	B_1
diagonal shear	$\partial_x v_y + \partial_y v_x$	B_2

Table 1: The target features.

E.3.2 Model and Training

We partition our model into two parts:

- A *backbone*, which is a 11-layer D_4-equivariant transformer backbone with 288 channels, consisting of 36 copies of the regular representation. The MLP ratio is 4, with the hidden layer containing 144 copies of the regular representation. Each attention layer is an `EquivariantCoupledAttention` (Ông & Bökman, 2026) with six heads.
- A vanilla (unconstrained) multi-head attention with 288 channels and six heads, followed by a linear $288 \to 6$ readout. We define the OV matrices to include this readout, so each R_h is a 6×288 matrix.

The model is trained for 200 epochs at batch size 64 with MSE loss of the six target entries. The optimizer is AdamW with $\beta = (0.9, 0.95)$, weight decay 0.05, and gradient clipping at norm 1.0. The learning rate linearly increases over the first 5% of the steps to 5×10^{-4}, then cosine-decays to 10^{-5} at epoch 150, then stays constant until epoch 200.

We perform random per-image D_4-augmentation and apply random color jitter, random color channel permutation, as well as Gaussian noise.

E.3.3 Norms

The norms used in the analysis presented in Section 5.2 are not the standard Euclidean norms. This is because the final attention layer only ever sees what the backbone outputs, and it has no reason to be equivariant in directions where the outputs have no component.

The norms are defined as follows. Suppose the backbone (blue part of the architecture in Fig. 7) outputs $Y \in \mathbb{R}^{B_{\mathrm{eval}} L \times C}$ on test data. Let $\mu_c := \frac{1}{B_{\mathrm{eval}} L} \sum_{bl} Y_{bl,c}$ denote the mean token feature, and let $\Sigma := \frac{1}{B_{\mathrm{eval}} L} (Y - \mu)^\top (Y - \mu) \in \mathbb{R}^{C \times C}$ be the covariance. Define a positive semidefinite bilinear form on $\mathrm{Mat}_{C \times C}(\mathbb{R})$ by

$$\langle M, M' \rangle_Y := \mathrm{tr}\left(\sqrt{\Sigma} M^\top \Sigma M' \sqrt{\Sigma}\right) + \mu^\top M^\top \Sigma M' \mu. \tag{55}$$

This is chosen so that

$$\|M\|_Y^2 = \mathbb{E}_{x \sim \mathcal{D}} \mathrm{Var}_{y \sim \mathcal{D}} x^\top M y, \tag{56}$$

where $\|M\|_Y^2 := \langle M, M \rangle_Y$, and $\mathcal{D}$ denotes the empirical single-token distribution. This measures how much the raw attention scores varies for a given query token, averaged over all query token.

Similarly, on $\mathbb{R}^{N\times C}$, define

$$\langle R, R' \rangle_Y := \mathrm{tr}\left(R^\top R' \Sigma\right) + \mu^\top R^\top S \mu, \tag{57}$$

which satisfies

$$\|R\|_Y^2 = \mathbb{E}_{x\sim\mathcal{D}} \|Rx\|^2. \tag{58}$$

The orthogonal projection used in panel (f) of Fig. 7 is with respect to $\langle , \rangle_D$.