A Deep Investigation on *Geometric Algebra Transformer*

Shengnan Liu

I. Introduction and motivation

In many scientific and engineering domains—from robotics and molecular chemistry to computer vision and astronomy—geometric data is fundamental. Such data, which includes positions, orientations, and forces, inherently exhibits rich symmetry properties. Traditional neural models often struggle to capture this complex spatial structure, prompting the need for innovative approaches. The Geometric Algebra Transformer (GATr) is one such approach; it represents geometric information in a unified 16-dimensional vector space using projective geometric (Clifford) algebra, which not only captures the essential characteristics of the data but also facilitates efficient, symmetry-preserving computations.

GATr is particularly compelling when considering its key advantages. First, GATr is more sample-efficient, learning faster than comparable models such as SE(3)-Transformers and Graph Neural Networks. In the meantime, its design ensures robust generalization, as it remains accurate under spatial translations—an important quality for applications where absolute positions matter. For instance, in molecular modeling, the chemical properties of a molecule remain invariant under rotations and translations. Such efficiency and generalization are crucial for applications where data may be scarce or variable, allowing for more reliable performance in diverse settings.

Furthermore, the implications for robotics are significant. Traditional robotics relies on pre-programmed control methods that perform well in fixed, highly controlled environments, such as factory floors. However, these methods inherently limit the robot's capabilities when confronted with dynamic, unstructured settings. By contrast, with a geometric framework and the integration of machine learning techniques, GATr could potentially enable robots to learn from experience and adapt to a broader range of environments. Since robotics involves inherently geometric problems—managing positions, orientations, joint movements, and forces—a model that naturally handles these complex relationships and aspects can significantly enhance performance and versatility.

To be more concrete, a central underlying principle in those applications is symmetry, meaning that one shape is identical to the other shape when it is translated, rotated, or flipped. Any model that fails to account for this symmetry may produce inconsistent or inefficient results. GATr addresses this by incorporating equivariance with respect to $E(3)$, where $E(3)$ comprises all translations, rotations, reflections, and arbitrary finite combinations of them in *3D* Euclidean space. Mathematically, equivariance is defined as $f(R(x)) = R(f(x))$, where R represents a translation, rotation, or reflection matrix. This property guarantees that applying a transformation before or after the network layer yields the same result, ensuring geometric consistency in GATr's outputs.

Recent works have explored various ways to harness geometric algebra for computing and modeling, yet they differ in the specific algebras used, how they handle equivariance, and the target application scenarios. GATr distinguishes itself by integrating a faithful E(3)-equivariant framework within the Transformer architecture, combining the scalability and flexibility of attention mechanisms with the powerful inductive biases provided by geometric algebra.

In this report, I will delve deeply into the Geometric Algebra Transformer published by Brehmer et al. I will detail its architectural innovations and theoretical foundations, and offer a brief discussion based on my understanding of GATr. By translating complex real-world geometric challenges into a coherent, theoretically grounded framework, GATr paves the way for more adaptive, efficient, and scalable solutions across diverse fields.

II. Problem Statement

The Geometric Algebra Transformer (GATr) represents inputs, outputs, and hidden states in the projective geometric (or Clifford) algebra, providing an efficient 16-dimensional vector-space representation for common geometric objects (e.g., points, planes, and lines) as well as the operators acting on them. This construction offers a unified framework to handle the inherent geometry of the data.

Moreover, GATr is designed to be equivariant, $f(R(x)) = R(f(x))$, with respect to $E(3)$ which comprises all translations, rotations, reflections, and arbitrary finite combinations of them in *3D* Euclidean space. In contrast to $O(3)$ which only accounts for rotations and reflections around the origin, $E(3)$ includes translations as well, preserving distances and angles everywhere rather than solely at the origin. *Table 1* highlights the key differences between these two groups:

	Transformations	Physical Interpretation	Matrix Representation	Determinant
O(3) Orthogonal Group in 3D	Rotations Reflections	Preserves distances and angles around the origin	Orthogonal matrices	+1 if rotation -1 if reflection
E(3) Euclidean Group in 3D	Rotations Reflections **Translations**	Preserves distances and angles everywhere	Combination of rotation matrix and translation vector	N/A

Table 1

By adopting the richer symmetry group $E(3)$, GATr can handle absolute positions via translations alongside rotations and reflections. This property is crucial for tasks that require consistent transformations of data under real-world operations, such as robotics, molecular modeling, and other geometric domains where preserving spatial relationships is essential.

III. Preliminaries

1. Metric space:

   Definition: Let X be a set. A function d: $X ✖ X \rightarrow R$ is called a metric on X if, for all $x, y, z \in X$, it satisfies:

   a. Non-negativity: $d(x, y) \geq 0,\ and\ d(x, y)\ =\ 0\ iff\ x\ =\ y$
   b. Symmetry: $d(x, y)\ =\ d(y, x)$
   c. Triangle inequality: $d(x, z) \leq d(x, y)\ +\ d(y, z)$.

2. Field:

   Definition: A field is a set F with two operations: $+$: $F \times F \rightarrow F$; $\bullet$: $F \times F \rightarrow F$ satisfying the following axioms:

   For addition:

   a. Associativity: $\forall v, w, x \in F,\ v\ +\ (w\ +\ x)\ =\ (v\ +\ w)\ +\ x$
   b. Commutativity: $\forall v, w \in F,\ v\ +\ w\ =\ w\ +\ v$
   c. Identity: $\exists 0 \in\ F\ s.t.\ \forall v \in F,\ 0\ +\ v\ =\ v$
   d. Inverse: $\forall v \in F,\ \exists(-\ v) \in\ F\ s.t.\ v\ +\ (-\ v)\ =\ 0$

   For multiplication:

   a. Associativity: $\forall v, w, x \in F,\ v \bullet (w \bullet x)\ =\ (v \bullet w) \bullet x$
   b. Commutativity: $\forall v, w \in F,\ v \bullet w\ =\ w \bullet v$
   c. Identity: $\exists 1 \in\ F\ -\ \{0\}\ s.t.\ \forall v \in F,\ 1 \bullet v\ =\ v$
   d. Inverse: $\forall v \in F\ -\ \{0\},\ \exists v^{-1} \in\ F\ s.t.\ v \bullet v^{-1}\ =\ 1$

   For both addition and multiplication:

   a. Distributive Property: $\forall v, w, x \in F,\ v \bullet (w\ +\ x)\ =\ v \bullet w\ +\ v \bullet x$

3. Vector Space:

   Definition: Let F be a field. A vector space over F is a set V, together with two operations

   Addtion $+$: $V \times V \rightarrow V$ and Scalar Multiplication $\bullet$: $F \times V \rightarrow V$

   that satisfies the following 8 axioms:

   For addition:

   a. Associativity: $\forall v, w, x \in V,\ v\ +\ (w\ +\ x)\ =\ (v\ +\ w)\ +\ x$
   b. Commutativity: $\forall v, w \in V,\ v\ +\ w\ =\ w\ +\ v$
   c. Identity: $\exists 0 \in\ V\ s.t.\ \forall v \in V,\ 0\ +\ v\ =\ v$
   d. Inverse: $\forall v \in V,\ \exists(-\ v) \in\ V\ s.t.\ v\ +\ (-\ v)\ =\ 0$

   For scalar multiplication:

   a. Associativity: $\forall a, b \in F,\ \forall v \in V,\ a \bullet (b \bullet v)\ =\ (a \bullet b) \bullet v$
   b. Unitary: $\forall v \in V,\ 1 \bullet v\ =\ v$

   For both additivity and multiplicativity:

   a. Distributive Property 1: $\forall a, b \in F,\ \forall v \in V,\ (a\ +\ b) \bullet v\ =\ a \bullet v\ +\ b \bullet v$
   b. Distributive Property 2: $\forall a \in F,\ \forall v, w \in V,\ a \bullet (v\ +\ w)\ =\ a \bullet v\ +\ a \bullet w$

4. Bilinearity:

   (1) Let V be a vector space over a field F. A function B: V ✖ $V \rightarrow F$ is called a bilinear form if for all $u, v, w \in V$ and all scalars $a \in F$, the following two conditions holds:

   Linear in the first argument:

   a. $B(u + v, w) = B(u, w) + B(v, w)$ and $B(a \cdot u, w) = aB(u, w)$

   Linear in the second argument:

   a. $B(u, v + w) = B(u, v) + B(u, w)$ and $B(u, a \cdot w) = aB(u, w)$

   (2) A bilinear form B: V ✖ $V \rightarrow R$ on a vector space V is called non-degenerate if the only vector $v \in V$ that satisfies $B(v, w) = 0$ for all $w \in V$ is $v = 0$.

5. In n-dimensional inner product space $(V, < \cdot, \cdot >)$, given an orthonormal basis of $\{e_1, ..., e_n\}$, we have :

$$< e_i, e_j >= \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases}$$

   The inner product $a \cdot b$ is symmetric ($a \cdot b = b \cdot a$).

6. Wedge product:

   (1) $P = a \wedge b$. P, a bivector, signifies a directed plane area in the same manner that vectors a and b signify directed line segments. Under this setting, the orientation of P is not determined by right-hand-rule. Instead, it is determined by the direction of a sweeping toward b. (The first vector sweeping toward the second vector)

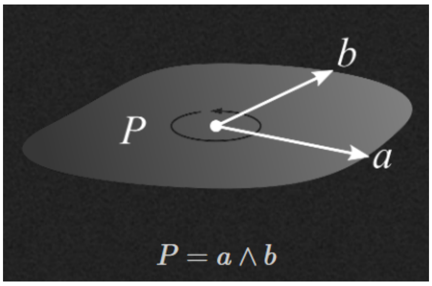

Figure 1

   The wedge product is an alternating bilinear map.

   $\forall a, a \wedge a = 0$. This can be immediately seen by substituting vector b by a in *Figure 1*.

   (2) The wedge product is *anti-symmetric*.

   Proof: Consider two arbitrary vectors a and b.

   Then, $(a + b) \wedge (a + b) = 0 = a \wedge a + a \wedge b + b \wedge a + b \wedge b$

   $\Rightarrow 0 + a \wedge b + b \wedge a + 0 = 0$

   $\Rightarrow a \wedge b + b \wedge a = 0$

   $\Rightarrow a \wedge b =- b \wedge a$

Or visually, consider swapping the position of a and b in the expression $a \wedge b$ and think about how *Figure 1* would change.

(3) An n-dimensional wedge product space is written as $(\Lambda^n V)$

If $u \in \Lambda^r V$ and $v \in \Lambda^s V$, then $u \wedge v \in \Lambda^{r+s} V$.

7. Geometric/Clifford algebra:

(1) Definition: Given a finite-dimensional vector space V over a field F with a symmetric bilinear form (e.g., Euclidean metric) g: $V \times V \rightarrow F$, a unital associative algebra $Cl(V, g)$ with a nondegenerate symmetric bilinear form g: $V \times V \rightarrow F$ is the Clifford algebra of the quadratic space (V, g) if:

a. it contains F and V as distinct subspaces
b. $a^2 = g(a, a)1\ for\ a \in V$
c. V generates $Cl(V, g)$ as an algebra
d. $Cl(V, g)$ is not generated by any proper subspace of V

(2) Arithmetic: Consider the real case $F = R$.

Geometric product: $ab = a \cdot b + a \wedge b$

- $a \cdot b$ is the inner product (symmetric) and $a \wedge b$ is the wedge product (anti-symmetric)

(3) Question: Given $a \cdot b$ a scalar and $a \wedge b$ a bivector, how do we understand "+" in ab?

Answer: Here, the summation means the grouping of the scalar and the bivector into a mixed-grade product called a multivector. They need not be added together into a single data type. If this does not make sense, consider the complex number analogy. For any complex number $z = x + iy$, it consists of the real and imaginary parts, but they can never be added together.

(4) Given $G(p, q)$, we have the following properties of the geometric product:

a. Closed under multiplication: $ab \in G(p, q)$
b. Existence of identity element: $1a = a1 = a$, where 1 is the identity element
c. Associativity: $(ab)c = a(bc) = abc$
d. Distributivity: $a(b + c) = ab + ac$ and $(b + c)a = ba + ca$
e. $a^2 = g(a, a)1$ for $a \in V$

8. Given an orthogonal set of basis vectors $\{e_0, ..., e_{n-1}\}$,

$for\ i = j \neq 0,\ e_i e_j = e_i \cdot e_j = 1$ $\qquad$ $for\ i \neq j,\ e_i e_j = e_i \wedge e_j = - e_j \wedge e_i = - e_j e_i$

9. Projective Geometric Algebra $G(3, 0, 1)$:

(1) $G(3, 0, 1)$ refers to a specific geometric (or Clifford) algebra that models 3-dimensional Euclidean space in its projective or homogeneous form.

3: **3** basis vectors with positive square norms (representing the usual Euclidean directions),
0: **0** basis vector with negative square norms,
1: **1** basis vector with zero square norms (degenerate).

(2) The basis and metric of $G(3, 0, 1)$:

Scalar	Vector				Bivector						Trivector				I (Identity)
Grade 0	Grade 1				Grade 2						Grade 3				Grade 4
1	e_0	e_1	e_2	e_3	e_{01}	e_{02}	e_{03}	e_{12}	e_{31}	e_{23}	e_{021}	e_{013}	e_{032}	e_{123}	e_{0123}
+1	0	+1	+1	+1	0	0	0	-1	-1	-1	0	0	0	-1	0
	Plane p				Line l						Point P				Pseudoscalar

Table 2

The values from the third row are the squared norms of each basis vector.

(3) The multiplication table of $G(3, 0, 1)$:

1	e_0	e_1	e_2	e_3	e_{01}	e_{02}	e_{03}	e_{12}	e_{31}	e_{23}	e_{021}	e_{013}	e_{032}	e_{123}	e_{0123}
e_0	0	e_{01}	e_{02}	e_{03}	0	0	0	$-e_{021}$	$-e_{013}$	$-e_{032}$	0	0	0	e_{0123}	0
e_1	$-e_{01}$	1	e_{12}	$-e_{31}$	$-e_0$	e_{021}	$-e_{013}$	e_2	$-e_3$	e_{123}	e_{02}	$-e_{03}$	e_{0123}	e_{23}	e_{032}
e_2	$-e_{02}$	$-e_{12}$	1	e_{23}	$-e_{021}$	$-e_0$	e_{032}	$-e_1$	e_{123}	e_3	$-e_{01}$	e_{0123}	e_{03}	e_{31}	e_{013}
e_3	$-e_{03}$	e_{31}	$-e_{23}$	1	e_{013}	$-e_{032}$	$-e_0$	e_{123}	e_1	$-e_2$	e_{0123}	e_{01}	$-e_{02}$	e_{12}	e_{021}
e_{01}	0	e_0	$-e_{021}$	e_{013}	0	0	0	e_{02}	$-e_{03}$	e_{0123}	0	0	0	$-e_{032}$	0
e_{02}	0	e_{021}	e_0	$-e_{032}$	0	0	0	$-e_{01}$	e_{0123}	e_{03}	0	0	0	$-e_{013}$	0
e_{03}	0	$-e_{013}$	e_{032}	e_0	0	0	0	e_{0123}	e_{01}	$-e_{02}$	0	0	0	$-e_{021}$	0
e_{12}	$-e_{021}$	$-e_2$	e_1	e_{123}	$-e_{02}$	e_{01}	e_{0123}	-1	e_{23}	$-e_{31}$	e_0	e_{032}	$-e_{013}$	$-e_3$	$-e_{03}$
e_{31}	$-e_{013}$	e_3	e_{123}	$-e_1$	e_{03}	e_{0123}	$-e_{01}$	$-e_{23}$	-1	e_{12}	$-e_{032}$	e_0	e_{021}	$-e_2$	$-e_{02}$
e_{23}	$-e_{032}$	e_{123}	$-e_3$	e_2	e_{0123}	$-e_{03}$	e_{02}	e_{31}	$-e_{12}$	-1	e_{013}	$-e_{021}$	e_0	$-e_1$	$-e_{01}$
e_{021}	0	e_{02}	$-e_{01}$	e_{0123}	0	0	0	e_0	e_{032}	$-e_{013}$	0	0	0	e_{03}	0
e_{013}	0	$-e_{03}$	e_{0123}	e_{01}	0	0	0	$-e_{032}$	e_0	e_{021}	0	0	0	e_{02}	0
e_{032}	0	e_{0123}	e_{03}	$-e_{02}$	0	0	0	e_{013}	$-e_{021}$	e_0	0	0	0	e_{01}	0

e_{123}	e_{0123}	e_{23}	e_{31}	e_{12}	e_{032}	e_{013}	e_{021}	$-e_3$	$-e_2$	$-e_1$	$-e_{03}$	$-e_{02}$	$-e_{01}$	-1	e_0
e_{0123}	0	$-e_{032}$	$-e_{013}$	$-e_{021}$	0	0	0	e_{03}	e_{02}	e_{01}	0	0	0	$-e_0$	0

Table 3

This is a locally symmetric or antisymmetric matrix.

(4) The squared norms of entrees from *Table 2*:

1	e_0	e_1	e_2	e_3	e_{01}	e_{02}	e_{03}	e_{12}	e_{31}	e_{23}	e_{021}	e_{013}	e_{032}	e_{123}	e_{0123}
e_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
e_1	0	1	-1	-1	0	0	0	1	1	-1	0	0	0	-1	0
e_2	0	-1	1	-1	0	0	0	1	-1	1	0	0	0	-1	0
e_3	0	-1	-1	1	0	0	0	-1	1	1	0	0	0	-1	0
e_{01}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
e_{02}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
e_{03}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
e_{12}	0	1	1	-1	0	0	0	1	-1	-1	0	0	0	1	0
e_{31}	0	1	-1	1	0	0	0	-1	1	-1	0	0	0	1	0
e_{23}	0	-1	1	1	0	0	0	-1	-1	1	0	0	0	1	0
e_{021}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
e_{013}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
e_{032}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
e_{123}	0	-1	-1	-1	0	0	0	1	1	1	0	0	0	1	0
e_{0123}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 4

This is a symmetric matrix: symmetric along the main diagonal.

(5) Examples of handling multiplications in detail:

$$e_{01} = e_0 e_1 = e_0 \cdot e_1 + e_0 \wedge e_1 = e_0 \wedge e_1 = - e_1 \wedge e_0 = - e_1 \wedge e_0 + e_1 \cdot e_0 = - e_1 e_0 = - e_{10}$$

$$e_{12} e_{12} = (e_1 e_2)(e_1 e_2) = (e_1 e_2)(- e_2 e_1) = - e_1 (e_2 e_2) e_1 = - e_1 * 1 * e_1 = - e_1 e_1 = - 1$$

(6) Given $G(3,0,1)$ with basis vectors $\{e_0, e_1, e_2, e_3\}$, a general multivector x takes the form

$x = a + b_0e_0 + b_1e_1 + b_2e_2 + b_3e_3 + c_{01}e_{01} + c_{02}e_{02} + c_{03}e_{03} + c_{12}e_{12} + c_{31}e_{31} + c_{23}e_{23} + d_{021}e_{021} + d_{013}e_{013} + d_{032}e_{032} + d_{123}e_{123} + f_{0123}e_{0123}$,

or, after simplification,

$x = a + b_0e_0 + b_1e_1 + b_2e_2 + b_3e_3 +$
$c_{01}(e_0 \wedge e_1) + c_{02}(e_0 \wedge e_2) + c_{03}(e_0 \wedge e_3) + c_{12}(e_1 \wedge e_2) + c_{31}(e_3 \wedge e_1) + c_{23}(e_2 \wedge e_3) +$
$d_{021}(e_0 \wedge e_2 \wedge e_1) + d_{013}(e_0 \wedge e_1 \wedge e_3) + d_{032}(e_0 \wedge e_3 \wedge e_2) + d_{123}(e_1 \wedge e_2 \wedge e_3) +$
$f_{0123}(e_0 \wedge e_1 \wedge e_2 \wedge e_3)$

with real coefficients $(a, b_0, b_1, \ldots, f_{0123}) \in R^{16}$.

By adding a fourth homogeneous coordinate b_0e_0 and dimensions with respect to e_0, it creates $2^4 = 16 - dimensions$.

10. A k-blade is a grade k element that can be expressed as the wedge product of k 1-vectors.

11. Grade involution:

(1) Definition: A linear involutive bijection that flips the sign of odd-grade elements.

(2) In $G(3,0,1)$, vectors (Grade 1) and trivectors (Grade 3) are flipped.

Mathematically, given a general multivector

$x = a + b_0e_0 + b_1e_1 + b_2e_2 + b_3e_3 + c_{01}e_{01} + c_{02}e_{02} + c_{03}e_{03} + c_{12}e_{12} + c_{31}e_{31} + c_{23}e_{23} + d_{021}e_{021} + d_{013}e_{013} + d_{032}e_{032} + d_{123}e_{123} + f_{0123}e_{0123}$,

the grade involution of x is:

$$\widehat{x} = a - b_0e_0 - b_1e_1 - b_2e_2 - b_3e_3 + c_{01}e_{01} + c_{02}e_{02} + c_{03}e_{03} + c_{12}e_{12} + c_{31}e_{31} + c_{23}e_{23} - d_{021}e_{021} - d_{013}e_{013} - d_{032}e_{032} - d_{123}e_{123} + f_{0123}e_{0123}$$

(3) Grade involution is an algebra automorphism $\widehat{xy} = \widehat{x}\widehat{y}$ and $\wedge$-algebra automorphism.

(4) For a k-blade $x = x_1 \wedge x_2 \wedge \ldots \wedge x_k$, the reversal in a linear involutive bijection will send x to $\widetilde{x} = x_k \wedge x_{k-1} \wedge \ldots \wedge x_1 = \pm x$, with $+\ x$ if $k \in \{0, 1, 4, 5, \ldots, 8, 9, \ldots\}$ and $-\ x$ otherwise.

This reversal operation is anti-automorphism: $\widetilde{xy} = \widetilde{y}\widetilde{x}$.

12. Dual and join in Euclidean algebra:

(1) The dual of a multivector x is a bijective, grade-reversing map $x^*: G(n,0,0) \to G(n,0,0)$ that sends each k-vectors to its complementary $(n-k)$-vectors.

Mathmatically, we write $x^* = xI^{-1}$.

This dual mapping is bijective and involutive up to a sign: $(y^*)^* = yI^{-1}I^{-1} = \pm y$ such that:

- If $n \in \{1, 4, 5, 8, 9, ...\}$, then $(y^*)^* = + y$
- If $n \in \{2, 3, 6, 7, ...\}$, then $(y^*)^* = - y$

Note: the paper chose I^{-1} instead of I in defining the dual because, given n vectors $x_1, x_2, ..., x_n$, the dual of the multivector $x = x_1 \wedge ... \wedge x_n$ can be given by the scalar of the oriented volume spanned by the vector.

The inverse of the dual is written as $x^{-*} = xI$.

In practice (Expressed in a basis), the dual swaps basis elements and may induce a sign depending on actual arithmetic.

For example, $n = 3$ ($i.e.$ $G(3, 0, 0)$) with $I = e_{123}$: $e_1^* = - e_{23}$, $e_2^* = e_{13}$, $e_{12}^* = e_3$.

(2) The bilinear join operation is defined based on the dual: for multivector x, y,

$$x \vee y := (x^* \wedge y^*)^{-*} = ((xI^{-1}) \wedge (yI^{-1}))I.$$

13. General Grade Projection:

Mathematically, any multivector x can be written as $x = \langle x \rangle_0 + \langle x \rangle_1 + ... + \langle x \rangle_n$, where $\langle x \rangle_i$ is the part of x that is homogeneous of grade i.

Then, the grade projection $\langle x \rangle_k$ outputs the grade-k portion of x.

For example, if x is an l-blade, then, by definition, $x = \langle x \rangle_l$ and $\forall\, k \neq l, \langle x \rangle_k = 0$.

14. Left Contraction:

(1) Let $l \geq k$. Given a k-vector a and l-vector b, the left contraction is defined as

$$a \rfloor b = \langle ab \rangle_{l-k}, \text{ which is a } l - k \text{ vector.}$$

For $k = 1$ and b a blade $b = b_1 \wedge b_2 \wedge ... \wedge b_l$, $a \rfloor b$ is the projection of a to the space spanned by the vectors b_i.

(2) If a and b are orthogonal, $a \rfloor b = 0 \iff \forall i, \langle a, b_i \rangle = 0$.

(3) A vector a is tangential to blade b if $a \wedge b = 0$.

Note: The term “tangential” here doesn’t mean “perpendicular” in the common-known Euclidean sense but instead means “contained within” the subspace represented by the blade.

In geometric algebra, a blade (like a 2-vector representing a plane) defines a subspace. When we say a vector a is tangential to a blade b if $a \wedge b = 0$, we mean that a does not add any new direction beyond what b already spans—that is, a lies entirely within the subspace of b.

For example, if b is a 2-vector representing a plane, any vector a that lies in that same plane will satisfy $a \wedge b = 0$ because combining a with b via the wedge product doesn’t “expand” the subspace.

This is different from the notion of orthogonality. Actually, $a\rfloor b = 0$ (the left contraction being zero) is used to define orthogonality, which corresponds to the usual Euclidean idea of being perpendicular.

(4) In the projective algebra, a blade b is defined to be ideal if it can be written as $b = e_0 \wedge z$ for another blade z.

15. Orthogonal Transformation:
Let V be a vector space. Let $(V, < \cdot, \cdot >)$ be a real inner product space.
A linear map $T\colon V \to V$ is called an orthogonal transformation if for all $x, y \in V$,

$$< T(x), T(y) > \ = \ < x, y >$$

Equivalently, T preserves lengths and angles: for every $x \in V$, $||T(x)|| = ||x||$ and if x and y are orthogonal, then so are $T(x)$ and $T(y)$.

16. Versor u:
Let $\{u_1, u_2, \dots, u_k\}$ be a set of unit vectors.
(1) In geometric algebra, a vector isn't just a point or direction—it can also serve as an operator. Specifically, a vector u_i can be used to reflect another element through the plane perpendicular to u_i. Since any orthogonal transformation, like rotation and reflection, can be broken down into a sequence of reflections, one can represent any such transformation as a product of several unit vectors, defined as a unit versor $u = u_1 \dots u_k$.
(2) Properties:
   a. The product of unit versors is a unit versor
   b. Unit versors are their own inverse

17. A quadratic form is a polynomial with terms all of degree two.
18. Pin group and Spin group:
(1) Let V be a vector space with a non-degenerate quadratic form Q.

$$Pin(V, Q) = \{u = u_1 u_2 \dots u_k \in Cl(V, Q)\backslash\{0\} \mid u_i \in V,\ u_i^2 = \pm 1,\ k \geq 0\}$$

(2) The spin group $Spin(V, Q)$ is defined similarly, but the only difference is that k has to be even. Consequently, it is easy to see that $Spin(V, Q)$ is a subgroup of $Pin(V, Q)$.

(3) $Pin(3, 0, 1)$:
Let V be a 4-dimensional real vector space with a quadratic form Q of signature $(3, 0, 1)$, meaning that V has three non-degenerate, Euclidean directions and one degenerate/null direction.
The Pin group $Pin(3, 0, 1)$ is defined as the set of all finite products of vectors from V that are normalized with respect to Q.
i.e. $Pin(3, 0, 1) = \{u = u_1 u_2 \dots u_k \mid u_i \in V, Q(u_i) = u_i^2 = \pm 1,\ k \geq 0\}$.

Note: The multiplication in the definition is the geometric product. The condition $Q(u_i) = u_i^2 = \pm 1$ guarantees each u_i is a unit vector and is invertible in $G(3,0,1)$. The degenerate vector is not used in forming the products because its square is 0. In this way, $Pin(3,0,1)$ forms a group which is a subgroup of the group of invertible elements in $G(3,0,1)$.

19. Sandwich product:
    (1) Designed to apply a versor u to an arbitrary element x
    $$u[x] = \rho_u(x) = uxu^{-1} \; if\ u\ is\ even$$
    $$u[x] = \rho_u(x) = u\hat{x}u^{-1} \; if\ u\ is\ odd$$
    (2) u acting on a multivector through a sandwich product is linear.
    (3) u acting on a multivector through a sandwich product is an outermorphism.
    (4) u acting on a multivector through a sandwich product is grade-preserving.

20. Equivariance: A function $f\colon\ G(3,0,1) \to G(3,0,1)$ is $Pin(3,0,1) - equivariant$ with respect to the representation ρ if
    $$f(\rho_u(x)) = \rho_u(f(x))$$
    for any $u \in Pin(3,0,1)$ and $x \in G(3,0,1)$, where $\rho_u(x)$ denotes the sandwich product.

21. Binary Operation(✱) : it is one operation that combines two elements in a set to produce another element in the set. Depending on the situation, ✱ can be addition, multiplication, etc..

22. Group:
    (1) A group is a set $G \neq \emptyset$ with a binary operation $\ast\colon\ G \times G \to G$ such that:
    a. Closure: $\forall a, b \in G,\ a \ast b \in G$
    b. Associativity: $\forall a, b, c \in G,\ (a \ast b) \ast c = a \ast (b \ast c)$
    c. Existence of Identity:
       $\exists\ an\ identity\ element\ e \in G\ such\ that\ \forall a \in G,\ a \ast e = e \ast a = a$
    d. Existence of Inverse: $\exists\ an\ inverse\ a^{-1} \in G\ such\ that\ \forall a \in G,\ a \ast a^{-1} = a^{-1} \ast a = e$

    (2) Group Homomorphism:
    Let $(G, \ast)$ and $(H, \ast)$ be groups.
    A function $f\colon\ G \to H$ is a group homomorphism of G into H iff
    $$\forall a, b \in G,\ f(a \ast b) = f(a) \ast f(b)$$
    It can be easily seen that $f(e_G) = e_H$ and $\forall a \in G, f(a^{-1}) = f(a)^{-1}$

    (3) Group Isomorphism:
    Let $(G, \ast)$ and $(H, \ast)$ be groups.
    If $f \in Hom(G, H)$ and f is bijective, then f is an isomorphism of G onto H

(4) The First Isomorphism Theorem for Groups:
Let $(G, \ast)$ and $(H, \ast)$ be groups.
Let f: $G \to H$ be a homomorphism of G into H ($f \in Hom(G, H)$).
Then, $G/ker(f) \cong f(G)$.
Moreover, if f is onto, then $G/ker(f) \cong H$.

(5) Endomorphism:
Let $(G, \ast)$ and $(H, \ast)$ be groups. Let $f \in Hom(G, H)$.
If $G = H$, then $Hom(G, G) = End(G)$, and such a $f \in Hom(G, G)$ is called an endomorphism.

23. Cover and double cover:
Let G' and G be groups.
(1) A covering homomorphism π: $G' \to G$ is a surjective group homomorphism such that algebraically, every element $g \in G$ has a preimage $\pi^{-1}(g)$ that is a discrete set of constant cardinality.
(2) n-Fold cover: G' is an n-fold cover of G if for every $g \in G$, $|\pi^{-1}(g)| = n$.
Double cover/2-fold cover: G' is double cover of G if for every $g \in G$, $|\pi^{-1}(g)| = 2$. In other words, each element $g \in G$ has exactly two preimages in G'.

24. Claim: $Pin(3, 0, 1)$ is a double cover of $E(3)$.
Proof: Define the map ρ: $Pin(3, 0, 1) \to E(3)$ by the sandwich action:

$$\rho_u(x) = uxu^{-1} \; if\ u\ is\ even$$

$$\rho_u(x) = u\hat{x}u^{-1} \; if\ u\ is\ odd$$

for every element x.
Part 1: ρ is a Group Homomorphism
Let $u, v \in Pin(3, 0, 1)$:

Case 1: Both u and v are even.
Then, $\rho_u(x) = uxu^{-1}$ and $\rho_v(x) = vxv^{-1}$.
Then, $\rho_u(\rho_v(x)) = u(vxv^{-1})u^{-1} = (uv)x(v^{-1}u^{-1}) = (uv)x(uv)^{-1} = \rho_{uv}(x)$

Case 2: One of u or v is odd. WLOG, suppose u is odd and v is even.
Then, $\rho_u(x) = u\hat{x}u^{-1}$ and $\rho_v(x) = vxv^{-1}$.

$$\rho_u(\rho_v(x)) = u(vxv^{-1})^{hat}u^{-1} = u(v\hat{x}v^{-1})u^{-1} = (uv)\hat{x}(v^{-1}u^{-1}) = (uv)\hat{x}(uv)^{-1} = \rho_{uv}(x)$$

Case 3: Both u and v are odd.
Then, $\rho_u(x) = u\hat{x}u^{-1}$ and Then, $\rho_v(x) = v\hat{x}v^{-1}$.

$$\rho_u(\rho_v(x)) = u(v\hat{x}v^{-1})^{hat}u^{-1} = (uv^{hat})x(uv^{hat})^{-1} = (u(-v))x(u(-v))^{-1} = (uv)x(uv)^{-1} = \rho_{uv}(x)$$

Part 2: ρ is surjective

By the Cartan–Dieudonné theorem, every orthogonal transformation in the nondegenerate (Euclidean) subspace of $G(3, 0, 1)$ can be written as a product of reflections, where each reflection is given by the sandwich action of a unit vector. In the projective setting, the inclusion of the degenerate vector e_0 allows the representation of translations as well. Thus, every element of $E(3)$ can be realized by an appropriate $u \in Pin(3, 0, 1)$ via the sandwich action ρ_u. Hence, ϱ is surjective.

Part 3: Kernel of ϱ

Goal: if an element $u \in Pin(3, 0, 1)$ satisfies $\rho_u(x) = x$,, then $u = \pm 1$.

Let V be a vector space, and $W \subset V$ a non-degenerate subspace spanned by e_1, e_2, e_3.

We can see that those basis vectors are of grade 1 so, for any $x \in W$, $\hat{x} =- x$.

Case 1: u is even.

Assume $u \in Pin(3, 0, 1)$ satisfies $\rho_u(x) = x$ for all $x \in W$.

By definition of map ρ, $\rho_u(x) = uxu^{-1}$

$\Rightarrow uxu^{-1} = x$ for all $x \in W$

$\Rightarrow uxu^{-1}u = xu$ for all $x \in W$

$\Rightarrow ux = xu$ for all $x \in W$

Since W generates the non-degenerate part of the Clifford algebra, we use the established fact that the only elements that commute with every vector in W are the scalars (Pertti).

Thus, $\exists \lambda \in R$ such that $u = \lambda$.

Since u is the product of unit vectors and so it is invertible, $||u|| = ||\lambda|| = 1$.

$\Rightarrow \lambda^2 = 1$

$\Rightarrow \lambda = \pm 1$

$\Rightarrow u = \pm 1$

Case 2: u is odd.

Assume $u \in Pin(3, 0, 1)$ satisfies $\rho_u(x) = x$ for all $x \in W$.

By definition of map ϱ, $\rho_u(x) = u\hat{x}u^{-1}$

$\Rightarrow u\hat{x}u^{-1} = x$ for all $x \in W$

$\Rightarrow u(- x)u^{-1} = x$ for all $x \in W$

$\Rightarrow - uxu^{-1} = x$ for all $x \in W$

$\Rightarrow uxu^{-1} =- x$ for all $x \in W$

$\Rightarrow u(uxu^{-1})u^{-1} =- uxu^{-1}$ for all $x \in W$

$\Rightarrow u^2x(u^{-1})^2 =- (- x) = x$ for all $x \in W$

$\Rightarrow u^2x = xu^2$ for all $x \in W$

Thus, u^2 commutes with every $x \in W$.

By a similar reason from Case 1, we know that $\exists\lambda \in R$ such that $u^2 = \lambda$

$\Rightarrow u^4 = 1$

$\Rightarrow u = \pm 1$

Thus, $ker(\rho) = \{1, -1\}$.

By the First Isomorphism Theorem,

$$Pin(3, 0, 1)/\{1, -1\} \cong E(3)$$

Thus, $Pin(3, 0, 1)$ is a double cover of $E(3)$.

Moreover, restricting to even elements with a similar argument gives

$$Spin(3, 0, 1)/\{1, -1\} \cong SE(3)$$

25. Any transformation $u \in Pin(n, 0, r)$ gives a homomorphism of the geometric algebra because for any multivectors x, y, $u[xy] = u\widehat{xy}u^{-1} = u\widehat{x}\widehat{y}u^{-1} = u\widehat{x}u^{-1}u\widehat{y}u^{-1} = u[x]u[y]$.

IV. Geometric Algebra Transformer

1. Overview:

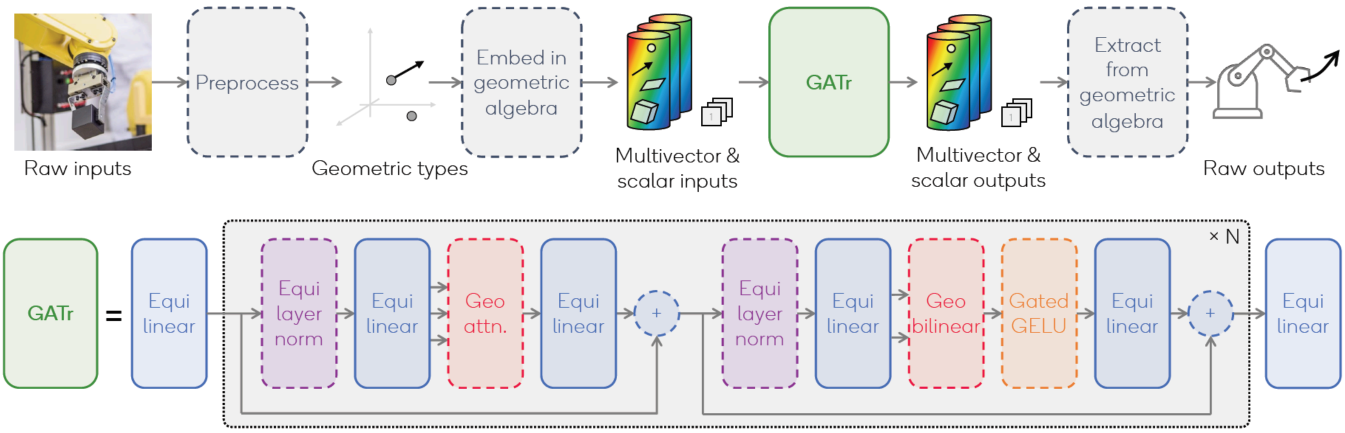

Figure 2

Operator	Scalar	Vector		Bivector		Trivector		PS
	1	e_0	e_i	e_{0i}	e_{ij}	e_{0ij}	e_{123}	e_{0123}
Scalar $\lambda \in R$	λ	0	0	0	0	0	0	0
Pseudoscalar $\mu \in R$	0	0	0	0	0	0	0	μ
Point $p \in R^3$	0	0	0	0	0	p	1	0
Point reflection through $p \in R^3$	0	0	0	0	0	p	1	0
Plane w/ normal $n \in R^3$, origin shift $d \in R$	0	d	n	0	0	0	0	0
Reflection through plane w/ normal $n \in R^3$, origin shift $d \in R$	0	d	n	0	0	0	0	0
Line w/ direction $n \in R^3$, orthogonal shift $s \in R^3$	0	0	0	s	n	0	0	0
Translation $t \in R^3$	1	0	0	$t/2$	0	0	0	0
Rotation expressed as quarternion $q \in R^4$	q_0	0	0	0	q_i	0	0	0

Table 5

Note: For *Figure 2*,

a. The top row shows the complete workflow: raw inputs are optionally converted into geometric types and then embedded as multivectors in the $G(3, 0, 1)$. For the sake of efficiency, common objects are embedded following *Table 5*. Then, these multivector tokens are processed by the GATr network.

b. The bottom row details the network architecture itself, which is built from N neural network transformer blocks. Each block contains an equivariant multivector LayerNorm, an equivariant linear transformation, a multivector self-attention mechanism, another equivariant linear transformation, a residual connection, another LayerNorm, another equivariant linear transformation, an equivariant MLP that uses geometric bilinear interactions, an activation function GELU, and another equivariant linear transformation—finished off with a second residual connection. In essence, the architecture is like a standard pre-layer normalized transformer but adapted to handle multivector data while ensuring $E(3)$ equivariance.

c. Also, in the diagram, boxes with solid borders indicate trainable components. In other words, they have parameters (weights, biases, etc.) that are adjusted during the training process. On the other hand, boxes with dashed borders denote fixed, non-trainable parts. These fixed parts typically include predetermined operations or functions, such as fixed preprocessing modules or specific normalization steps, that do not change during the training process.

2. General inputs without optimization in GATr:
In $G(3,0,1)$ with basis vectors $\{e_0, e_1, e_2, e_3\}$, where e_0 is degenerate and $\{e_1, e_2, e_3\}$ represents the usual Euclidean directions,
$x = a + b_0 e_0 + b_1 e_1 + b_2 e_2 + b_3 e_3 + c_{01} e_{01} + c_{02} e_{02} + c_{03} e_{03} + c_{12} e_{12} + c_{31} e_{31} + c_{23} e_{23} + d_{021} e_{021} + d_{013} e_{013} + d_{032} e_{032} + d_{123} e_{123} + f_{0123} e_{0123},$
In practice, common objects are embedded following *Table 5* as mentioned above.

3. Linear Layers:
Any linear map Φ: $G(3,0,1) \to G(3,0,1)$ equivariant to $Pin(3,0,1)$ has the form:

$$\Phi(x) = \sum_{k=0}^{4} w_k \langle x \rangle_k + \sum_{k=0}^{3} v_k e_0 \langle x \rangle_k,$$

- $w \in R^5$, $v \in R^4$.
- $\langle x \rangle_k$ is the grade-k projection of the multivector x, which sends all non-grade-k components to 0.

In other words, $E(3)$-equivariant linear maps between $G(3,0,1)$ multivectors can be parameterized using 5 coefficients for the grade projections and 4 for the homogeneous multiplication—9 parameters in total per channel pair.

4. Geometric Bilinears
Since linear maps alone cannot mix the different grades (scalar, vector, bivector, etc.) effectively, two bilinear operations are introduced:
a. Geometric product:

$$x, y \mapsto xy \text{ where } xy = x \cdot y + x \wedge y.$$

b. Equivariant Join:

$$x, y, z \mapsto EquiJoin(x, y; z) = z_{0123}(x^* \wedge y^*)^*$$

- x^* is the dual of x, y^* is the dual of y.
- $z_{0123} \in R$ is the pseudoscalar component of a reference vector z chosen based on the input data rather than the hidden representations. To be specific, it is chosen to be the mean of all inputs to the network.

Remark: dualization is necessary, because, without it, even simple functions like the Euclidean distance between two points cannot be represented as such a function relies on combining information across different grades.

c. Geometric bilinear layer: The combination of a and b:

$$Geometric(x, y; z) = Concatenate_{channels}(xy, EquiJoin(x, y; z))$$

5. Nonlinearities and equivariant normalization:
   a. $GELU$ (Gaussian Error Linear Unit) nonlinear activation function:
   $$GELU(x) = x \cdot P(X \leq x) = x \cdot \Phi(x) = x \cdot [1 + erf(x/\sqrt{2})] / 2$$
   Remark: $\operatorname{erf} z = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, dt.$

   b. $GELU(x)$ can be approximated by
   $$0.5 \cdot x \cdot (1 + tanh[(\sqrt{(2/\pi)}) \cdot (x + 0.044715 \cdot x^3)])$$
   c. Scalar-gated $GELU$ activation function:
   $$GatedGELU(x) = GELU(x_1) \cdot x$$
   - x_1 is the scalar component of the multivector x.

   d. Normalization: $E(3)$-equivariant $LayerNorm$:
   $$LayerNorm(x) = x / E_c^{0.5}[\langle x, x\rangle]$$
   - $\langle x, x\rangle$ is the invariant inner product in $G(3, 0, 1)$.
   - E_c is the expectation over channels/key length.

6. Attention:
   Let $q,\ k,\ v$ be the query, key, and value tensors, respectively, with indices $i,\ i'$ over tokens/items and $c,\ c'$ over channels/key length. The $E(3)$-equivariant multivector attention is:
   $$\text{Attention}(q, k, v)_{i'c'} = \sum_i \text{Softmax}_i\left(\frac{\sum_c \langle q_{i'c}, k_{ic}\rangle}{\sqrt{8n_c}}\right) v_{ic'}$$
   - $\langle q_{i'c}, k_{ic}\rangle$ is the invariant inner product, which uses the 8 dimensions in $G(3, 0, 1)$ not involving e_0
   - n_c is the key length
   - n_i is the number of tokens

- $\sqrt{(8n_c)}$ is the scaling factor.

7. Auxiliary scalar representation:
While the multivector representation is ideal for geometric information, many tasks also involve high-dimensional non-geometric data, such as sinusoidal positional encodings and time indices. GATr handles these by incorporating an auxiliary scalar to the hidden states of GATr that parallels the multivector stream. In this case, each layer can produce both scalar and multivector inputs and outputs, which have the same batch dimension and item dimension but not necessarily key length.

The integration occurs in following ways:
a. In linear layers: The scalar features are allowed to mix with the scalar component of the multivectors.
b. In the attention layer: the $Attention(q, k, v)_{i'c'}$ is used for computing attention weights from the multivectors, and the normal scaled dot-product attention is used for computing attention weights from the auxiliary scalars. Then, these weights maps are summed before the softmax is applied, and normalization is properly done. This ensures that both geometric and non-geometric information contribute to the attention weights.

$$Attention(q, k, v)_{i'c'} = \mathrm{Softmax}_i \left(\frac{\sum_c \langle q_{i'c}^{MV}, k_{ic}^{MV} \rangle + \sum_c q_{i'c}^{s} k_{ic}^{s}}{\sqrt{8n_{MV} + n_s}} \right)$$

- q^{MV} and k^{MV} are query and key multivector representations
- q^{s} and k^{s} are query and key scalar representations
- n_{MV} is the number of multivector channels
- n_s is the number of scalar channels

c. In all other layers: the multivector information is processed separately from the scalar information. For example, regular $GELU$ functions are applied to auxiliary scalars while equivariant $GatedGELUs$ are applied to multivectors in the phase of nonlinear transformation/activation.

8. Distance-aware dot-product attention:
Since the $E(3)$-equivariant multivector attention only uses the 8 dimensions in $G(3, 0, 1)$ not involving e_0 and the 8 dimensions involving e_0 vary under translations which means the normal Euclidean inner product violates equivariance, it is necessary to extend the attention mechanism with addition, nonlinear features to incorporate capturing more relevant geometric relationships inherited in the 8 dimensions involving e_0.
In GATr, for each channel c in the query and key, the auxiliary, nonlinear query features $\Phi(q)$ and key features $\Psi(k)$ are defined and the $E(3)$-equivariant multivector attention weights are extended as $\langle q_{i',c'}, k_{ic} \rangle \rightarrow \langle q_{i',c'}, k_{ic} \rangle + \Phi(q_{i'c}) \cdot \Psi(k_{ic})$:

a. $\Phi(q_{i'c})$ and $\Psi(k_{ic})$ are learned nonlinear mappings from the query and key channels to scalars with the following choices:

$$\phi(q) = \omega(q_{\backslash 0}) \begin{pmatrix} q_{\backslash 0}^2 \\ \sum_i q_{\backslash i}^2 \\ q_{\backslash 0}\, q_{\backslash 1} \\ q_{\backslash 0}\, q_{\backslash 2} \\ q_{\backslash 0}\, q_{\backslash 3} \end{pmatrix} \quad \text{and} \quad \psi(k) = \omega(k_{\backslash 0}) \begin{pmatrix} -\sum_i k_{\backslash i}^2 \\ -k_{\backslash 0}^2 \\ 2k_{\backslash 0}\, k_{\backslash 1} \\ 2k_{\backslash 0}\, k_{\backslash 2} \\ 2k_{\backslash 0}\, k_{\backslash 3} \end{pmatrix} \quad \text{with} \quad \omega(x) = \frac{x}{x^2+\epsilon}$$

The index $\backslash i$ means the trivector component of the query or key input with all indices except for i. With this choice,

$\Phi(q) \cdot \Psi(k) = - \omega(q_{\backslash 0}) \times \omega(k_{\backslash 0}) \times ||k_{\backslash 0}\vec{q} - q_{\backslash 0}\vec{k}||^2_{R^3}$, where $\vec{q} = (q_{\backslash 1}, q_{\backslash 2}, q_{\backslash 3})^T$ and $\vec{k} = (k_{\backslash 1}, k_{\backslash 2}, k_{\backslash 3})^T$.

b. If the trivector components of the queried and keys encode 3D points where $q_{\backslash 0} = k_{\backslash 0} = 1$, then $\Phi(q) \cdot \Psi(k) \propto -||p_q - p_k||^2$. In other words, it is proportional to the pairwise negative squared Euclidean distance between inputs.

c. For a trivector $q = (q_{\backslash 0}, \vec{q})$, the rotation transformation follows $R \in O(3)$ and the translation transformation follows $(q_{\backslash 0}, \vec{q}) \mapsto (q_{\backslash 0}, R\vec{q} + q_{\backslash 0}t)$ where $t \in R^3$.

d. Such a construction and process produces the invariance property of $\Phi(q) \cdot \Psi(k)$:

$$\Phi(q) \cdot \Psi(k) \mapsto -\omega(q_{\backslash 0}) \times \omega(k_{\backslash 0}) \times ||k_{\backslash 0}(R\vec{q} + q_{\backslash 0}t) - q_{\backslash 0}(R\vec{k} + k_{\backslash 0}t)||^2_{R^3}$$

$$\Rightarrow \quad = -\omega(q_{\backslash 0}) \times \omega(k_{\backslash 0}) \times ||R(k_{\backslash 0}\vec{q} - q_{\backslash 0}\vec{k}) + q_{\backslash 0}k_{\backslash 0}t - q_{\backslash 0}k_{\backslash 0}t||^2_{R^3}$$

$$\Rightarrow \quad = -\omega(q_{\backslash 0}) \times \omega(k_{\backslash 0}) \times ||k_{\backslash 0}\vec{q} - q_{\backslash 0}\vec{k}||^2_{R^3}$$

$$\Rightarrow \quad = \Phi(q) \cdot \Psi(k)$$

Interpretation: Consider a general form $f(x) \mapsto f(T(x)) = f(x)$ for all x.
In the above process, $\Phi(q) \cdot \Psi(k)$ is transformed in a way such that rotating q and k does not change the final scalar value. More concretely, the Euclidean norm of the difference of two vectors is unchanged by a rotation R and a translation t, so any term involving $||k_{\backslash 0}\vec{q} - q_{\backslash 0}\vec{k}||^2_{R^3}$ remains the same when $\vec{q}$ and $\vec{k}$ are replaced by $R\vec{q}$ and $R\vec{k}$.
Remark: Rather than using a constant $\omega(x) = 1$—which would lead to a fourth-degree polynomial in the keys and queries, causing numerical instability upon exponentiation, the choice of $\omega(x) = x/(x^2 + \varepsilon)$ allows the function to behave in a way that keeps the attention score effectively quadratic in the keys and queries as $|x|$ grows large, much like standard dot-product attention, thus avoiding runaway values in the softmax.

9. GATr's overall attention mechanism:
With the construction from 6-8, GATr's overall attention mechanism consists of the following three parts:
    a. The $G(3, 0, 1)$ inner product $\langle q, k \rangle$ between the multivector queries and keys

b. The Euclidean inner product of auxiliary scalars $q_s \cdot k_s$

c. The distance-sensitive inner product of nonlinear features $\Phi(q) \cdot \Psi(k)$

Mathematically, the final attention weights are:

$$Attention(q, k, v)_{i'c'} = \mathrm{Softmax}_i \left(\frac{\alpha \sum_c \langle q^{MV}_{i'c}, k^{MV}_{ic} \rangle + \beta \sum_c \phi(q^{MV}_{i'c}) \cdot \psi(k^{MV}_{ic}) + \gamma \sum_c q^s_{i'c} k^s_{ic}}{\sqrt{13 n_{MV} + n_s}} \right)$$

- $\alpha, \beta, \gamma > 0$ are learnable, head-specific weights
- q^{MV} and k^{MV} are query and key multivector representations
- q^{s} and k^{s} are query and key scalar representations
- n_{MV} is the number of multivector channels
- n_s is the number of scalar channels
- Indices i, i' are over tokens/items and indices c, c' are over channels/key length

Adding α, β, γ are beneficial:

a. First, the learnable weights allow the network to adaptively balance the contributions from each source of information. Depending on the task or the data distribution, the model might benefit from placing more emphasis on the geometric (multivector) relationships or on the distance information provided by the nonlinear features. In other words, these weights enable the network to learn which aspects of the geometric and scalar interactions are most informative for a given problem.

b. Second, scaling each term separately can help with numerical stability during training. Without such prefactors the raw contributions might have different magnitudes, which could cause instabilities. For example, applying the softmax function over the attention logits. The prefactors allow the network to rescale the terms so that the combined attention logit remains well-behaved.

c. Third, adding these learnable weights increases the model's expressive power. Instead of being forced to treat the three terms equally, the model can learn a more flexible representation by modulating the relative importance of each geometric and scalar component. This flexibility is particularly useful when handling diverse geometric data where the relevance of each type of interaction may vary across tasks.

V. Theoretical Guarantee

1. Proposition 2: The grade projection $\langle x\rangle_k$ is equivariant.

Proof: Let u be a 1-versor.

Assume a multivector x is a l-blade.

Then, write x as $x = a_1 \wedge a_2 \wedge ... \wedge a_l$.

Since u is an outermorphism, applying u to x gives

$$u[x] = u[a_1] \wedge ... \wedge u[a_l]$$

$u[x]$ is also an l-blade because the action u is grade-preserving.

Now, consider the grade projection $\langle x\rangle_k$.

Case 1: $l \neq k$:

Since x is homogeneous of grade l and $l \neq k$, $\langle x\rangle_k = 0$.

Then, transforming $\langle x\rangle_k$ by a linear action u gives $u[\langle x\rangle_k] = u[0] = 0$.

On the other hand, since $u[x]$ is also an l-blade, its grade-k projection is also 0: $\langle u[x]\rangle_k = 0$.

Thus, $u[\langle x\rangle_k] = 0 = \langle u[x]\rangle_k$, the equivariance condition holds.

Case 2: $l = k$.

Then, x is already homogeneous of grade k so $\langle x\rangle_k = x$.

On the other hand, $u[x]$ in this case is also a k-blade, so $\langle u[x]\rangle_k = u[x]$.

Thus, $u[\langle x\rangle_k] = u[x] = \langle u[x]\rangle_k$, the equivariance condition holds.

Since grade projection is linear, the equivariance holds for any multivector.

2. Proposition 3: The map Φ: $G(3, 0, 1) \rightarrow G(3, 0, 1) : x \mapsto e_0 x$ is equivariant.

Proof: Let u be a 1-versor. (So it is odd, and it is a vector)

Then, u acts on a multivector x is the sandwich action $u[x] = u\hat{x}u^{-1}$, where $\hat{x}$, as introduced in the preliminary, is the grade involution.

Consider the action of u on e_0.

By the definition of e_0, the choice of u, and the functionality of u discussed in the sandwich product section, we have the orthogonality of u and e_0 and thus $ue_0 =- e_0 u$.

In other words, $ue_0 = u \cdot e_0 + u \wedge e_0 = 0 + (- e_0 \wedge u) =- e_0 \wedge u =- e_0 u$.

Then, $u[e_0] =- ue_0 u^{-1} = e_0 uu^{-1} = e_0$.

As we proved in the double cover section, the sandwich action u is a homomorphism, then, $u[\Phi(x)] = u[e_0 x] = u[e_0]u[x] = e_0 u[x] = \Phi(u[x])$.

Thus, the map Φ is equivariant with respect to 1-versor u.

Actually, it follows that the map Φ is equivariant to any versor u', as we can decompose u' into 1-versors and recursively apply $u[\Phi(x)] = \Phi(u[x])$ for 1-versor.

3. Theorem 1 (Cartan-Dieuodonné): Every orthogonal transformation of an n-dimensional space can be decomposed into at most n reflections in hyperplanes.
   Proof: By mathematical induction:
   Base case: $n = 1$.
   In 1-dimension, every orthogonal transformation of R^1 is either the identity or a point reflection. Thus, any rotation in 1-dimension is just a sign flip (it could only be 180° rotation) which is exactly one reflection; the identity can be done in zero reflection; any reflection is just one reflection itself. Thus, the theorem follows trivially in R^1.

   Inductive step: Assume Cartan-Dieuodonné Theorem holds in $n - 1$ dimensions, which means any orthogonal transformation of T^{n-1} can be written as at most $n - 1$ reflections: $T^{n-1} = u_1 o\, u_2\, o \dots\, o\, u_{n-1}$, where u_i is a reflection. We will proceed to prove that the theorem holds for n dimensions.
   Proof: Let T^n be an orthogonal transformation different from the identity in n -dimensions.
   Fix the origin. Assume that $v \in R^n$ is a hyperplane where $T^n[v] \neq v$.
   Define the bisector $u_n = (v - T^n[v])'$.
   Since the reflection in the bisector u_n maps v to the same hyperplane as $T^n[v]$, we have $T^{n-1} = u_n T^n$ act as the identity on v: $T^{n-1}[v] = v$.
   Meanwhile, T^{n-1} is an orthogonal transformation on the $n - 1$ dimensional subspace orthogonal to v.
   By the inductive assumption, $T^{n-1} = u_1 o\, u_2\, o \dots\, o\, u_{n-1}$, therefore, we can rearrange $T^{n-1} = u_n T^n$ to get $T^n = u_1 o\, u_2\, o \dots\, o\, u_n$.
   Thus, with mathematical induction, we proved the theorem.

4. Lemma 1: In the n-dimensional Euclidean geometric algebra $G(n, 0, 0)$, the group $Pin(n, 0, 0)$ acts transitively on the space of k-blades of norm $\lambda \in R^{>0}$.
   Proof: WLOG, assume $\lambda = 1$ because the Pin group preserves the norm of multivectors.
   Let x be a unit norm k-blade multivector. By Gram-Schmidt process, we can write $x = v_1 \wedge v_2 \wedge \dots \wedge v_k$, where v_i orthonormal.
   Let y be another k-blade $y = w_1 \wedge w_2 \wedge \dots \wedge w_k$, where w_i orthonormal.

   By Axler 2.33 "Every linearly independent list can be extended to a basis," we can extend v_i's and w_i's by choosing $n - k$ additional orthonomal vectors $\{v_{k+1}, \dots, v_n\}$ and $\{w_{k+1}, \dots, w_n\}$ to form orthonomal bases $\{v_1, \dots, v_n\}$ and $\{w_1, \dots, w_n\}$ of R^n.

Then, by Axler 3.5 "Given $\{x_1, ..., x_n\}$ a basis for V and $\{y_1, ..., y_n\}$ any vectors in W, there exists a unique linear transformation $T: V \to W$ such that $T(x_i) = y_i$ for all i," we can see that, in our case, there exists a unique orthogonal transformation $O: R^n \to R^n$ such that $O(v_i) = w_i$ for all i.

Then, by the Cartan-Dieuodonné Theorem, $O: R^n \to R^n$ can be written as the product of at most n reflections, so there exists a $u \in Pin(n, 0, 0)$ such that $u[v_i] = w_i$ for all i.

Since the action u is an outermorphism ($u[a \wedge b] = u[a] \wedge u[b]$ for all multivectors a and b),

$u[x] = u[v_1 \wedge v_2 \wedge ... \wedge v_k] = u[v_1] \wedge u[v_2] \wedge ... \wedge u[v_k] = w_1 \wedge w_2 \wedge ... \wedge w_k = y.$

Thus, u is indeed acting transitively on targeted space.

5. Lemma 2: In the Euclidean ($r = 0$) or projective ($r = 1$) geometric algebra $G(n, 0, r)$, let x be a k-blade and let u be a 1-versor.

   Then, $u[x] = x \leftrightarrow u \rfloor x = 0$, and $u[x] =- x \leftrightarrow u \wedge x = 0$.

Proof: Let x be a k-blade and let u be a 1-versor.

Then, we can decompose u into $u = t + v$, where t, chosen to be Euclidean, is the tangential part to the subspace of x such that $t \wedge x = 0$, and v is the normal part such that $v \rfloor x = 0$ (the component of u that is perpendicular to every direction in x).

This decomposition $u = t + v$ is unique unless x is ideal in projective geometric algebra. If x is ideal, the e_0 component of u is both normal and tangential.

Note:$xt = (-1)^{k-1}tx; xv = (-1)^k vx;\ vt = -tv;\ \nexists\lambda \neq 0\ such\ that\ vtx = \lambda x$

Since u is a 1-versor, the sandwich action on x is:

$u[x] = (-1)^k(t + v)x(t + v) = (t + v)(-t + v)x = (-\|t\|^2 + \|v\|^2)x - 2vtx$

Then, we can see that:

- $u[x] \propto x \Leftrightarrow vtx = 0$
- If x is not ideal, then either $v = 0$ so $u \wedge x = 0\ and\ u[x] =- x$ or $t = 0$ so $u\rfloor x = 0\ and\ u[x] = x$
- If x is ideal, then either $v \propto e_0$ so $u \wedge x = 0\ and\ u[x] =- x$ or $t = 0$ so $u\rfloor x = 0\ and\ u[x] = x$.

6. Lemma 3: Let $r \in \{0, 1\}$. Any linear $Pin(n, 0, r)$-equivariant map $\Phi: G(n, 0, r) \to G(n, 0, r)$ can be decomposed into a sum of equivariant maps $\Phi = \Sigma_{lkm}\Phi_{lkm}$, with Φ_{lkm} equivariantly mapping k-blades to l-blades.

   If $r = 0$ (Euclidean algebra), or $k < n + 1$, such a map Φ_{lkm} is defined by the image of any one non-ideal k-blade, like $e_{12...k}$.

   Instead, if $r = 1$ (projective algebra), and $k = n + 1$, then such a map is defined by the image of a pseudoscalar, like $e_{01...n}$.

Proof: Any multivector $x \in G(n, 0, r)$ can be uniquely written as a sum of its homogeneous components: $x = \langle x\rangle_0 + \langle x\rangle_1 + \langle x\rangle_2 + ... + \langle x\rangle_n$.

Since the sandwich action u is grade-preserving, the $Pin(n, 0, r)$ group action maps k-vectors to k-vectors.

Consequently, Φ can be decomposed into $\phi(x) = \Sigma_k \phi(\langle x\rangle_k)$.

Moreover, by linearity we can further decompose each $\phi(\langle x\rangle_k)$ into components of definite output grade. We write the part of ϕ that maps grade-k elements to grade-l elements as ϕ_{lk}.

Then, we can further specify $\phi(x) = \Sigma_l \Sigma_k \phi_{lk}(\langle x\rangle_k)$, where each ϕ_{lk} is an equivariant linear map from the subspace of k-vectors to the subspace of l-vectors. Also, ϕ_{lk} has l-vectors as image, and for $k' \neq k$, all k'-vectors in the kernel.

Let x be non-ideal k-blade or pseudoscalar if $k = n + 1$.

By Lemma 1 and 4, in both Euclidean and projective geometric algebra, the span of the k-vectors in the orbit of x contains any k-vector.

Thus, ϕ_{lk} is defined by the l-vector $y = \phi_{lk}(x)$.

Since any l-vector can be decomposed as a finite sum of l-blades, we can first write $y = y_1 + ... + y_M$ and the define $\phi_{lkm}(x) = y_m$.

Then, by equivariance, the definition can be extended to all l-vectors.

Then, $\phi_{lk} = \Sigma_m \phi_{lkm}$.

7. Proposition 4: For an n-dimensional Euclidean geometric algebra $G(3, 0, 0)$, any linear endomorphism Φ: $G(n, 0, 0) \rightarrow G(n, 0, 0)$ that is equivariant to the $Pin(n, 0, 0)$ group (equivalently to $O(n)$) is of the type $\Phi(x) = \sum_{k=0}^{n} w_k \langle x\rangle_k$, for parameters $w \in R^{n+1}$.

Proof: By decomposition of Lemma 3, let Φ map from k-blades to l-blades.

Let x be a k-blade. Let u be a 1-versor.

By Lemma 2 and equivariance:

- if $u \rfloor x = 0$ (u is orthogonal to x), then $u[\Phi(x)] = \Phi(u[x]) = \Phi(x)$, and so $u \rfloor \Phi(x) = 0$ (i.e. u is also orthogonal to $\Phi(x)$).
- If $u \wedge x = 0$, then $u[\Phi(x)] = \Phi(u[x]) = \Phi(-x) = -\Phi(x)$ and $u \wedge \Phi(x) = 0$.

Thus, any vector in x is in $\Phi(x)$ and any vector orthogonal to x is orthogonal to $\Phi(x)$, this implies $\Phi(x) = w_k x$, for some $w_k \in R$.

By Lemma 3, we can extend Φ to $\Phi(y) = w_k y$ for any k-vector y.

8. Lemma 4: The Pin group of the projective geometric algebra, $Pin(n, 0, 1)$, acts transitively on the space of k-blades with positive norm $||x|| = \lambda > 0$. Additionally, the group acts transitively on the space of zero-norm k-blades of the form $x = e_0 \wedge y$ (called ideal blades), with $||y|| = \kappa$.

Proof: Case 1: Non-ideal blades.

Let $x = x_1 \wedge x_2 \wedge ... \wedge x_k$ be a k-blade with positive norm λ.

Since x_i's are all vectors and we have $Pin(n, 0, 1)$, all vectors x_i can be decomposed as $x_i = v_i + \delta_i e_0$, where v_i is a nonzero Euclidean vector with no e_0 component, and $\delta_i \in R$.

In this case, $v_i = 0 \rightarrow ||x|| = 0$.

Then, by Gram-Schmidt process, we can orthogonalize x_i's such that

$x_2' = x_2 - \langle x_1, x_2 \rangle x_1, etc.$

In this case, we can write $x' = x_1' \wedge x_2' \wedge ... \wedge x_k'$ where $x_i' = v_i' + \delta_i' e_0$ and v_i' are orthogonal.

Now, define the translation as $t = 1 + (\sum_i \delta_i' e_0 \wedge v_i')/2$.

This translation operator t acts on multivectors via the sandwich product.

t removes the e_0 components of x' and thus makes x' Euclidean:

$t[x'] = v_1' \wedge v_2' \wedge ... \wedge v_k'$.

By Lemma 1, we know that the Euclidean $Pin(n, 0, 0)$, a subgroup of $Pin(n, 0, 1)$, acts transitively on the space of Euclidean k-blades of positive norm. Thus, for any two such Euclidean k-blades $t[x']$ and $t[y']$, there exists a $g \in Pin(n, 0, 0)$ such that $g[t[x']] = t[y']$.

Since any non-ideal k-blade x can be translated by t to a purely Euclidean k-blade and then any two such Euclidean blades are related by an element $g \in Pin(n, 0, 0)$, we conclude that in projective geometric algebra, any two non-ideal k-blades with the same positive norm are related by and element of $Pin(n, 0, 1)$. In other words, the $Pin(n, 0, 1)$ acts transitively on the space of non-ideal k-blades with positive norm λ.

Case 2: Ideal blades.

Let $x = e_0 \wedge y$, where y is a $(k - 1)$-vector in the Euclidean subspace, and assume $||y|| = \kappa > 0$.

Since for any $g \in Pin(n, 0, 1)$, $g[e_0] = e_0$, the sandwich action on the ideal blade $x = e_0 \wedge y$ thus becomes $g[x] = g[e_0 \wedge y] = g[e_0] \wedge g[y] = e_0 \wedge g[y]$.

Now, consider another $x' = e_0 \wedge y'$ with $||y'|| = \kappa$ and y' Euclidean.

Again, by the transitivity of $Pin(n, 0, 0)$ on Euclidean $(k - 1)$-blades with norm κ, there exists a $g \in Pin(n, 0, 0)$ and hence $g \in Pin(n, 0, 1)$ such that $g[y] = y'$.

Then, $g[x] = e_0 \wedge g[y] = e_0 \wedge y' = x'$.

Thus, the $Pin(n, 0, 1)$ also acts transitively on the space of zero-norm ideal k-blades of the form $x = e_0 \wedge y$ with $||y|| = \kappa > 0$.

9. Proposition 5: For the projective geometric algebra $G(n, 0, 1)$, any linear endomorphism $\Phi: G(n, 0, 1) \rightarrow G(n, 0, 1)$ that is equivariant to the group $Pin(n, 0, r)$ (equivalently to $E(n)$) is of the type $\Phi(x) = \sum_{k=0}^{n+1} w_k \langle x \rangle_k + \sum_{k=0}^{n} v_k e_0 \langle x \rangle_k$ for parameters $w \in R^{n+2}$, $v \in R^{n+1}$.

Proof: By Lemma 3, any linear $Pin(n, 0, r)$-equivariant map Φ can be decomposed into linear equivariant map from k-blades to l-blades.

Case 1: For $k < n + 1$, let $x = e_{12...k}$.

Then, by Lemma 2 and proof of Proposition 4, for any $1 \le i \le k$,

$e_i \wedge x = 0$, $e_i[x] =- x$, and $e_i[\phi(x)] = \phi(e_i[x]) = \phi(- x) =- \phi(x)$, and thus $e_i \wedge \phi(x) = 0$.

Consequently, we can write $\phi(x) = x \wedge y_1 \wedge ... \wedge y_{l-k}$, where $l - k$ vector y_j orthogonal to x.

Now, consider $k < i \le n$.

Since $x = e_{12...k} = e_1 \wedge e_2 \wedge ... \wedge e_k$, any e_i with $k < i \le n$ is not in x.

Then, $e_i \rfloor x = 0$. By Lemma 2 and using equivariance, we have:

$e_i \rfloor x = 0 \Rightarrow e_i[\phi(x)] = \phi(x) \Rightarrow e_i \rfloor \phi(x) = 0 \Rightarrow \forall i,\ \langle e_i, y_j \rangle = 0.$

Thus, y_j is orthogonal to all e_i with $1 \le i \le n$.

Then, $l = k$ or $l = k + 1$ and $y_1 \propto e_0$.

The above argument imply that $\phi(x)$ has the same tangent normal structure as x and hence $\phi(x) = w_k x$.

Case 2: For $k = n + 1$, let $x = e_{012...k}$.

By a similar argument, for any invertible vector u that is tangent to x, equivariance forces $u[x]$ to be tangent to $\phi(x)$. Then, we can write $\phi(x) = x \wedge y$ for some blade y.

However, since x already spans the full space of $(n + 1)$-vectors up to scaling, the only possibility for a nonzero $\phi(x)$ is that $y \propto 1$.

Thus, $\phi(x) \propto x$.

Thus, by the above argument, Lemma 3, equivariance, linearity, we conclude that the only linear, $Pin(n, 0, 1)$-equivariant endomorphisms of $G(n, 0, 1)$ are those that, when restricted to each homogeneous grade k, act by multiplying by a scalar w_k (if the output grade is preserved) or by appending an e_0 factor (if the output grade is increased by one).

Thus, $\Phi(x) = \sum_{k=0}^{n+1} w_k \langle x \rangle_k + \sum_{k=0}^{n} v_k e_0 \langle x \rangle_k$ for parameters $w \in R^{n+2}$, $v \in R^{n+1}$.

10. Lemma 5: In Euclidean algebra $G(n, 0, 0)$, the join is $Spin(n, 0, 0)$ equivariant.

Furthermore, it is $Pin(n, 0, 0)$ equivariant $\Leftrightarrow$ n is even.

Proof: For $u \in Spin(n, 0, 0)$ (even versors), it is easy to see that $u[I] = I$ so $u[I^{-1}] = I^{-1}$.

Since the dual is defined using I^{-1}, the dualization operation is preserved.

Since the sandwich action by u is an outermorphism, both the wedge product and the dual operations are equivariant under u.

Then, since the join is defined based on the dual and the wedge product, it follows that

$u[x \vee y] = u[((xI^{-1}) \wedge (yI^{-1}))I] = u[(xI^{-1}) \wedge (yI^{-1})]u[I] = (u[xI^{-1}] \wedge u[yI^{-1}])I$

$= ((u[x]u[I^{-1}]) \wedge (u[y]u[I^{-1}]))I = ((u[x]I^{-1}) \wedge (u[y]I^{-1}))I = u[x] \vee u[y]$.

Now, consider $I = e_1 e_2 ... e_n \in Pin(n, 0, 0)$ not in $Spin(n, 0, 0)$, which is the point reflection. It negates vectors of odd grades by the grade involution: $I[x] = \hat{x}$.

Let x be a k-vector and y an l-vector.

Then $x \vee y$ is a vector of grade $n - ((n - k) + (n - l)) = k + l - n$ and zero if $k + l < n$

Since the join is bilinear, the inputs transform a $(-1)^{k+l}$ under the point reflection, while the transformed output gets a sign $(-1)^{k+l-n}$.

Then, only when $(-1)^{k+l} = (-1)^{k+l-n}$ can we maintain the equivariance property, which means it has to be $(-1)^{-n} = 1$. Thus, n has to be even.

Thus, if u is odd, then the join does not commute with $Pin(n,\ 0,\ 0)$ group action due to the sign mismatch, so the join cannot be $Pin(n,\ 0,\ 0)$-equivariant under odd u.

To address this, a pseudoscalar $z = \lambda I$ is introduced, and the equivariant Euclidean join is defined as $EquiJoin(x,\ y,\ z\ =\ \lambda I) := \lambda(x \vee y)\ =\ \lambda(x^* \wedge y^*)^{-*}$.

11. Proposition 6: In Euclidean algebra $G(n, 0, 0)$, the equivariant join $EquiJoin$ is $Pin(n, 0, 0)$ equivariant.

Proof: The $EquiJoin$ is designed to be a multilinear operation.

Then, for k-vector x and l-vector y, under a point reflection, the input gets a sign $(-1)^{k+l+n}$ while the output is still a $k + l - n$ vector and gets the sign $(-1)^{k+l-n}$.

The above two signs always differ by $(-1)^{2n} = 1$, so $EquiJoin$ is $Pin(n, 0, 0)$ equivariant.

12. Lemma 6: In the algebra $G(n, 0, 0)$, let v be a vector and x, y be multivectors.

$$\text{Then, } v \rfloor (x \vee y) = (v \rfloor x) \vee y \text{ and } x \vee (v \rfloor y) = -(-1)^n \widehat{v \rfloor x} \vee y.$$

Proof: For the first equation:

Let v be a vector, a be a k-vector and b an l-vector.

From the proof of Lemma 5, we know that $a \vee b$ is a $(k + l - n)$-vector, or a scalar if $k + l < n$. WLOG, let $k + l - n \geq 0$.

Then,

$a \vee b = (a^* \wedge b^*)^{-*} = \langle a^* b I^{-1} \rangle_{2n-k-l} I = \langle a^* b \rangle_{n-(2n-k-l)} I^{-1} I = \langle a^* b \rangle_{k+l-n} = a^* \rfloor b$

$(v \rfloor a)^* = \langle va \rangle_{k-1} I^{-1} = \langle v a I^{-1} \rangle_{n-k+1} = \langle v a^* \rangle_{n-k+1} = v \rfloor (a^*)$.

Combining the above two relationships with proper choice of a and b:

$(v \rfloor x) \vee y = (v \rfloor x)^* \rfloor y = v \rfloor (x^*) \rfloor y = v \rfloor (x \vee y). \qquad \cdots(1)$

For the second equation:

Note that swapping k-vector a and l-vector b gives:

$a \vee b = (a^* \wedge b^*)^{-*} = (-1)^{(n-k)(n-l)}(b^* \wedge a^*)^{-*} = (-1)^{(n-k)(n-l)}(b \vee a).$

Then,

$$\begin{aligned} a \vee (v \rfloor b) &= (-1)^{(n-k)(n-l-1)} (v \rfloor b) \vee a \\ &= (-1)^{(n-k)(n-l-1)} v \rfloor (b \vee a) \\ &= (-1)^{(n-k)(n-l-1)+(n-k)(n-l)} v \rfloor (a \vee b) \\ &= (-1)^{(n-k)(n-l-1)+(n-k)(n-l)} (v \rfloor a) \vee b \\ &= (-1)^{(n-k)(2n-2l-1)} (v \rfloor a) \vee b \\ &= (-1)^{k-n} (v \rfloor a) \vee b \\ &= -(-1)^{k-n-1} (v \rfloor a) \vee b \\ &= -(-1)^{-n}((-1)^{k-1}(v \rfloor a)) \vee b \\ &= -(-1)^{n}\widehat{(v \rfloor a)} \vee b \qquad \cdots(2) \end{aligned}$$

Generalize to multivectors x and y and the second equation is proved.

Dual and join in projective algebra:

Given $G(n, 0, 1)$ with degenerate e_0, for a bijective dual that yields the complementary indices on basis elements, the right complement has to be used. In this case, we have to choose an orthogonal basis and then for a basis k-vector x to define the dual x^* to be the basis $n - k + 1$-vector such that $x \wedge x^* = I$, where I is the pseudoscalar $I = e_{01\ldots n}$. For example, given e_{01}, the dual is $e_{01}^* = e_{23}$ so $e_{01} \wedge e_{23} = e_{0123}$

Note: Even though we can calculate this dual easily by numerical methods, we can't derive it using only the intrinsic operations provided by the geometric algebra. This limitation makes it harder to prove the equivariance property.

13. Proposition 7: In the algebra $G(n, 0, 1)$, the join $a \vee b = (a^* \wedge b^*)^{-*}$ is equivariant to $Spin(n, 0, 1)$.

    Proof: Although the dual operation is not well-defined in $G(n, 0, 1)$, since $G(n, 0, 1)$ actually contains $G(n, 0, 0)$ as a subalgebra and the properties and functionalities of the dual have been proved earlier, it is then possible to define the join:

    Let x be a k-vector and decomposed x as $x = t_x + e_0 p_x$

    - t_x is the Euclidean k-vector (no e_0 component) so it behaves exactly like a k-vector in the standard Euclidean algebra $G(n, 0, 0)$
    - p_x is the Euclidean $(k - 1)$-vector (no e_0 component)
    - $e_0 p_x$ is a k-vector that encodes the translational information of the geometry

    Define the Euclidean join of vectors a, b in the projective algebra to be equal to the join of the corresponding vectors in the Euclidean algebra:

$$a \vee_{\mathrm{Euc}} b := ((a\,\widetilde{e_{12\ldots n}}) \wedge (b\,\widetilde{e_{12\ldots n}}))e_{12\ldots n}$$

    Then, following the original paper and Dorst, we have:

$(t_x + e_0p_x) \vee (t_y + e_0p_y) = ((t_x + e_0p_x)^* \wedge (t_y + e_0p_y)^*)^{-*}$

$= t_x \vee_{Euc} p_y + (-1)^n \widehat{p_x} \vee_{Euc} t_y + e_0(p_x \vee_{Euc} p_y)$ $\cdots(3)$

By Lemma 5, the operation $a \vee_{Euc} b$ is $Spin(n, 0, 0)$ equivariant.

Also, note that any rotation $r \in Spin(n, 0, 0) \subset Spin(n, 0, 1)$ is Euclidean.
Thus, $r[a \vee_{Euc} b] = r[a] \vee_{Euc} r[b]$.

Together with $r[a^*] = r[a]r[I^{-1}] = (r[a])I^{-1} = (r[a])^*$, we can naturally conclude that the dual in (1) is equivariant to the rotational subgroup $Spin(n, 0, 0) \subset Spin(n, 0, 1)$.
Next, equivariance to translations should be proved.
Let v be Euclidean vector and $\tau = 1 - e_0v/2$ a translation.

Specifically, translations act by shifting with e_0 times a left contraction:

$\tau[x] = x - e_0(v \rfloor x)$.

For $x = t_x + e_0p_x$, $\tau[t_x + e_0p_x] = \tau[t_x] + e_0p_x = t_x + e_0(p_x - v \rfloor t_x)$.

Then, consider $x = t_x + e_0p_x$ and $y = t_y + e_0p_y$

$\tau[x] \vee \tau[y] = (\tau[t_x] + e_0p_x) \vee (\tau[t_y] + e_0p_y)$

$= (t_x + e_0(p_x - v \rfloor t_x)) \vee (t_y + e_0(p_y - v \rfloor t_y))$

By (3) and linearity,

$= x \vee y - t_x \vee_{Euc} (v \rfloor t_y) - (-1)^n \widehat{v \rfloor t_x} \vee_{Euc} t_y$

$- e_0 (p_x \vee_{Euc} (v \rfloor t_y) + (v \rfloor t_x) \vee_{Euc} p_y)$

By (2),

$= x \vee y - e_0 (p_x \vee_{Euc} (v \rfloor t_y) + (v \rfloor t_x) \vee_{Euc} p_y)$

By (2),

$= x \vee y - e_0(-(-1)^n \widehat{v \rfloor p_x} \vee_{Euc} t_y + (v \rfloor t_x) \vee_{Euc} p_y)$

$= x \vee y - e_0((-1)^n (v \rfloor \widehat{p_x}) \vee_{Euc} t_y + (v \rfloor t_x) \vee_{Euc} p_y)$

By (1),

$= x \vee y - e_0(v \rfloor \{(-1)^n \widehat{p_x} \vee_{Euc} t_y + t_x \vee_{Euc} p_y\})$

$= \tau[x \vee y]$

Thus, the join is equivariant to both rotations and translations, so it is equivariant to $Spin(n, 0, 1)$.

For $Pin(n, 0, 1)$:

(1) Action of $g \in Pin(n, 0, 1)$ on x and I: Even g: $g[I] = I$ and odd g: $g[I] = -I$

(2) How the dual transform: Given $g[xy] = g[x]g[y]$, $g[x^*] = g[x]g[I^{-1}]$

For even g, $g[I^{-1}] = (g[I])^{-1} = I^{-1} \Rightarrow g[x^*] = g[x]g[I^{-1}] = g[x]I^{-1} = (g[x])^*$

For odd g, $g[I^{-1}] = (g[I])^{-1} = -I^{-1} \Rightarrow g[x^*] = -g[x]I^{-1} = (g[x])^*$

Thus, for even g, $g[x^*] = g[x]I^{-1} = (g[x])^*$.

(3) Application to the join: Since the join is defined by $a \vee b = \langle a^* b I^{-1} \rangle_{2n-k-l} I$ (equivalent form), which uses the dual $a^* = aI^{-1}$ and then projects onto a particular grade before multiplying by I.

Then, the same reasoning applies: under any $g \in Pin(n, 0, 1)$, the extra sign from $g[I^{-1}]$ when g is odd cancels with the sign changes coming from $g[a]$ and $g[b]$. In the end, we obtain $g[a \vee b] = g[a] \vee g[b]$.

Therefore, we arrive at full $Pin(n, 0, 1)$ equivariance via multiplication with a pseudoscalar, and the $EquiJoin$ defined from above can still be used in the projective geometric algebra.

14. Lemma 7: For the algebra $G(n, 0, r)$, for multivectors x, y, $||xy|| = ||x||\, ||y||$.
Proof: $\|xy\|^2 = xy\widetilde{xy} = xy\tilde{y}\tilde{x} = x\|y\|^2\tilde{x} = x\tilde{x}\|y\|^2 = \|x\|^2\|y\|^2$

VI Discussion

1. Question: Given enough computational resources, is there a way to improve the general capability of GATr by reducing its inductive bias?
Answer: Since in GATr, the inductive bias is brought by the inherent structure of the (projective) geometric algebra, it seems not feasible to reduce the inductive bias, because otherwise, we directly deny the underlying bedrock of GATr. But if we are able to excavate more potential of GATr (a.k.a use (projective) geometric algebra to represent the information inherited from a wider range of data, not limited in spatial and non-spatial ones mentioned by the paper which includes translation, rotation, reflection, sinusoidal positional encodings and time indices), then the general capability of GATr could be further enhanced.

References

[1] Johann Brehmer, Pim de Haan, Sönke Behrends, and Taco Cohen. 2023. *Geometric Algebra Transformer.* arXiv:2305.18415. (PDF)

[2] Taco Cohen. 2023. *Speech: "Geometric Algebra Transformers: Revolutionizing Geometric Data with Taco Cohen, Qualcomm AI Research."*
Available at https://www.youtube.com/watch?v=nPIRL-c88_E. (Video)

[3] Pertti Lounesto. 2001. *Clifford Algebras and Spinors.* 2nd edition. Cambridge University Press. (PDF)

[4] Steven De Keninck and Charles Gunn. 2019. *3DPGA.* (PDF)

[5] Dan Hendrycks and Kevin Gimpel. 2016. *Gaussian Error Linear Units (GELUs).* arXiv:1606.08415. (PDF)

[6] Martin Roelfs and Steven De Keninck. 2021. *Graded Symmetry Groups: Plane and Simple.* arXiv:2107.03771v. (PDF)

[7] Jean Gallier. 2012. *Clifford Algebras, Clifford Groups, and a Generalization of the Quaternions: The Pin and Spin Groups.* (PDF)

[8] Leo Dorst. 2020. *A Guided Tour to the Plane-Based Geometric Algebra PGA.* (PDF)

[9] Leo Dorst, Daniel Fontijne, and Stephen Mann. 2007. *Geometric Algebra for Computer Science: An Object-Oriented Approach to Geometry.* Morgan Kaufmann. (PDF)

[10] David Ruhe, Jayesh K. Gupta, Steven de Keninck, Max Welling, and Johannes Brandstetter. 2023. *Geometric Clifford Algebra Networks.* arXiv:2302.06594. (PDF)

[11] Quadratic form. n.d. https://en.wikipedia.org/wiki/Quadratic_form (Website)

[12] Pin group. n.d. https://en.wikipedia.org/wiki/Pin_group (Website)

[13] Geometric algebra. n.d. https://en.wikipedia.org/wiki/Geometric_algebra (Website)

[14] Quadratic form. n.d. https://en.wikipedia.org/wiki/Quadratic_form (Website)

[15] Exterior algebra. n.d. https://en.wikipedia.org/wiki/Exterior_algebra (Website)

[16] Error function. n.d. https://en.wikipedia.org/wiki/Error_function (Website)

[17] Blade (geometry). n.d. https://en.wikipedia.org/wiki/Blade_(geometry) (Website)

[18] Grade Projection and Extraction. n.d. https://library.fiveable.me/geometric-algebra/unit-5/grade-projection-extraction/study-guide/YoFJFJqb5YYpSxRT (Website)

[19] Geometric Algebra in Three Dimensions. n.d. https://www.cv.nrao.edu/~mmorgan2/resources/geo3.html (Website)