Banya AI 7

The Rotation-Operator Gate: Reading the Operation Plane as Euler Rotations for Depthless Logic Composition

Hyukjin Han (bokkamsun@gmail.com) · ORCID 0009-0007-4132-1707

Banya Research Series, Part 7 (an extension of Part 3, the bi-token)

Abstract

This paper starts from two expressive limits of the bi-token gate of Part 3 [3]. The sigmoid gate's range of 0 to 1 expresses only passing and blocking and cannot express negation (a sign flip) in one hop, and the operation-plane components act diagonally, each adjusting only its own channel, so there is no rotation. This paper proposes the rotation-operator gate, which reads the operation plane as Euler rotation angles. The grounding is the factorization of the fractional derivative (differentiation generalized to a real order). In the frequency domain, $(i\omega)^\alpha$ splits into the product of the amplitude $|\omega|^\alpha$ and the Euler rotation $e^{i\alpha\pi/2}$, and it is the rotation that carries the logic. In the rotation gate, negation becomes a primitive operation of 180 degrees, the composition of operations becomes angle addition, the norm is preserved, and the adjoint (the transpose operation that flows gradients backward) comes out in closed form as the conjugate rotation $R(-\theta)$. In particular, the angle gradient closes as the two-dimensional cross product of the incoming credit and the rotation result, so the forward-pass input never needs to be stored. This adjoint was verified against finite differences with an inner-product identity difference of 0 and a median gradient ratio of 1.000000. The measurements were made with depth-excluded discrimination, restricting depth to one block so that the gate is the only remaining variable. On a task of reading three kinds of numbers and stating the mod 3 state of the running sum, teaching only 14 of the 27 length-3 combinations and hiding 13 (the blanks), an ablation contrast that swaps only the gate on the same one-block engine splits the ranking. Self-feedback accuracy on the blanks is 100 percent on all three seeds for the rotation gate and the tanh gate, 90 percent for the sigmoid gate, 62 percent with the operation plane blocked, and 49 percent for a standard one-layer transformer (chance 33 percent). At the unseen length 6, the rotation gate is highest at an average of 82 percent with one seed reaching 100 percent, while the standard one-layer collapses to chance. Convergence is also fastest for the rotation gate at an average of 317 steps, with the smallest variation across seeds. Since mod n addition is isomorphic to the discrete rotation group that divides the circle into n parts, the algebra of this task is rotation itself, and the reason this discrimination is deliberately biased is that no order-3 operation exists in a real scalar gate. This paper then extends the same discrimination one rung up. The composition of pairwise planar rotations closes as angle addition and is therefore commutative; a non-commutative composition, where order changes the result, does not exist in that gate. The prescription is the quaternion gate, which reads four components of the operation plane as a unit quaternion and left-multiplies blocks of four channels. Left multiplication by a unit quaternion rotates two overlapping planes at once, so compositions do not commute, and the adjoint is again a storage-free closed form that closes through the conjugate left product and a recovery term (inner-product identity difference 0, median gradient ratio 1.000000). The discrimination task moves to the symmetries of a marked equilateral triangle, the smallest non-commutative group: reading moves (a 120-degree turn, a 240-degree turn, a flip) one at a time, the model states one of the six triangle states at every position. In the composition trial, which exposes all 18 rule-table entries and hides 13 of the 27 length-3 combinations (six seeds), the quaternion gate is highest on the blanks at a mean of 97.5 percent, converges fastest at a mean of 288 steps with the smallest variation across seeds, is the most uniform on the unseen lengths 4 to 6 at 94 to 100 percent forced across all six seeds, and declines most gently under self-feedback. The remaining failures of both geometric gates localize entirely to the rules of the starting state, and every engine group scores far above the 38-percent commutative ceiling that bounds any order-blind solver, because the loop externalizes order into the states written in the window. In the inference trial, which leaves part of the rule table unexposed, no group sustains a score above the 69-percent share of blanks solvable from seen rules alone across the seeds, and the rotation and quaternion gates never fall below that share on any of the six seeds, separating the composition of learned rules from the inference of untaught rules as different abilities. A commutative task needs only rotation; on a non-commutative task, rotation in overlapping planes is the fastest and most stable vessel. That the algebra of the task determines the gate's primitive is the shared conclusion of the two discriminations. We state explicitly the limitation of a small synthetic task and that training at the published model size remains a next task.

Measured

All quantitative figures in this paper are measured values obtained in the writing environment (environment in Appendix B). The entire training and measurement is reproduced with a single bundled probe; no separate model or corpus download is needed (Appendix C). The theoretical background of the orthogonal view is in the separate axiom document [1]; the claims of this paper come from measurement, not from deduction from those axioms.

Contents

Terminology Legend

Terms are defined in the tables below. One concept goes by one name only, and the text and figures of all seven parts follow these tables. The in-code column gives the identifier that implements the concept in the enclosed code, and the standard-term column gives the closest name in common use; cells are filled only where they differ.

Terminology legend 1. Engine and credit

Term	Meaning	In code	Standard term
exact adjoint	The backward computation derived by hand as closed forms for every operation; the proper name of this engine and its credit chain	`banya_core.py`	manual back-propagation
adjoint	The partner operation of the forward pass; it carries the gradient exactly backward	`bwd` kernels	adjoint
credit	The loss gradient at a hidden state; the responsibility a parameter bears for the outcome	delta (δ)	gradient
credit chain	The skeleton that walks the cache in reverse, applying each adjoint to carry credit back to the input	`backward`	back-propagation
adjoint agreement	The check that the adjoint formulas produce the same values as finite differences	per-part probes	gradient check
circulant-matrix channel mixing	Channel mixing with a circulant matrix, computed as componentwise products under rfft		circular convolution
relative-distance mixing	Mixing positions with shared coefficients that depend only on distance	`m_mat_w_lag`	Toeplitz mixing
temporal mixing	The umbrella name for any path by which other positions' information flows into the current one		token mixing
mixing heads	The number of parallel heads a mixing operation is split into		attention heads
filter ladder	The convolutional channel transform whose filter scale varies with depth, from a few wide filters to many narrow ones	`m_mat_w_filter`	convolutional filterbank
window	The token positions the model sees at once; as a measurement unit, a fixed-length text slice	`T`	context window
probe	The verification script that replays every measured figure together with its target value	`probe` folder	reproduction script
frozen model	A developmental-stage model whose training is stopped and fixed	`model` npz	frozen checkpoint
holdout	Evaluation text never used in training		held-out set

Terminology legend 2. Bi-token and the gate

Term	Meaning	In code	Standard term
bi-token	The structure that gives one token an orthogonal pair of embeddings, a data plane and an operation plane	`USE_BITOKEN`	
data plane	The embedding plane carrying a token's content; it enters at the add slot	`m_mat_w_data_axis`	token embedding
operation plane	The embedding plane carrying a token's operation instruction; it enters only at the multiply slot	`m_mat_w_operate_axis`	
add slot, multiply slot	The residual-addition position where the data plane lands and the gate-multiplication position where the operation plane acts		residual connection, gating
gate	The valve that sets each channel's pass-through between 0 and 1; the previous token's operation plane is injected into it	`m_mat_w_gate`	gating
one-step shift	The one-position displacement by which the previous token's operation plane is injected into the next position's gate		
operation-as-weight	The phenomenon in which operation-plane embeddings emerge as weights without any labels		
consistent operator	An operator that acts in a consistent direction on any data		
direction consistency	The mean cosine of the directions in which one operation bends different data; the quantitative sign of operator emergence	`p_direction_consistency`	

Terminology legend 2 (continued). Bi-token and the gate

Term	Meaning	In code	Standard term
rotation-operator gate	The gate that reads the operation plane as Euler rotation angles; abbreviated in the text as the rotation gate	`paper7_rotation.py`	
quaternion gate	The gate that reads four operation-plane components as a unit quaternion and left-multiplies blocks of four channels	`install_quaternion`	
primitive operation	An operation the gate can express directly in a single-hop instruction		primitive
primitive ladder	The extension series of gate primitives running rotation, quaternion, and Clifford algebras above them		
depth-excluded discrimination	The deliberately biased measurement that restricts depth to one block so the gate is the only remaining variable	Part 7 probe	
rule table	The table giving the next state for every state-and-move pair; 9 entries in the mod 3 task, 18 in the triangle task		transition table
triangle task	The non-commutative discrimination task of reading three kinds of moves and stating one of six states of a marked triangle at every position	Part 7 probe	
inference trial	The random-split discrimination that leaves part of the rule table unexposed, measuring the inference of untaught rules		
composition trial	The coverage-guaranteed-split discrimination that exposes the whole rule table, measuring pure composition of learned rules		
commutative ceiling	The accuracy upper bound an order-blind solver can reach on the blanks		
self-feedback	The scoring mode that feeds the model's own output back as input and scores the final state		autoregressive rollout

Terminology legend 3. Rumination and development

Term	Meaning	In code	Standard term
Rumination	Self-organizing learning that pulls embeddings toward their neighbors without computing any gradient	Part 2 probes	self-organization
same-kind group	The set of close-in-meaning concepts gathered from a seed by nearest-cosine traversal		cluster
centroid pull, decay	The two forces of Rumination: pulling toward the mean and restoring toward the original; equilibrium emerges from their balance	`ALPHA`, decay rate	
World-first	The curriculum order that develops sensory, spatial, and logical axes before language	`banya_world_data`	curriculum learning
developmental stage	A checkpoint stage defined by step count and named after human growth		
axis	A property scale formed as a direction inside the embedding: sensory, ordinal, categorical		representation axis

Terminology legend 4. Vocabulary and metacognition

Term	Meaning	In code	Standard term
syllable atom	The syllable-level token forming the bottom layer of the vocabulary	`AtomTokenizer`	character-level token
bundle	A concept token made by binding syllable atoms into the dictionary; used only as a vocabulary unit		subword
cache-dictionary tokenization	The two-layer vocabulary system that puts a bundle dictionary on top of the atom layer	Part 5 probes	contrast with subword tokenization
certain, vague, unknown	The three knowing states separated by the model's own output distribution	pmax, entropy	uncertainty estimation
routing	Assigning an answer's processing path according to the state		routing
reinterpretation	The procedure that re-runs a vague hidden state together with its context to lift it to certain	Part 6 probes	
similarity consolidation	The axis along which concepts are organized by data-plane similarity		
inference traversal	The axis along which inference is carried forward on the operation plane		
self-utterance	The target state that weaves similarity consolidation and inference traversal so the model continues speech by itself		
Banya framework	The fifteen-axiom system forming the series' background; not the ground of the claims in the text		

Symbol Legend

This part defines the symbols used jointly by its formulas and text. Local symbols used within a single equation are given on the symbol line beneath that equation.

Symbol	Name	Definition and meaning	Type
ω	frequency	Frequency variable in the frequency domain; one differentiation multiplies each component by $i\omega$	scalar
α	differentiation order	The order of the fractional derivative, generalizing the number of differentiations to a real number	scalar
i	imaginary unit	i itself is $e^{i\pi/2}$, a 90-degree rotation, so two differentiations give negation	scalar
n	modulus	The modulus of mod n addition and the number of divisions of the circle; an order-n cycle is $360/n$ degrees	integer
σ	sigmoid	The function of the Part 3 valve gate; its range of 0 to 1 expresses only passing and blocking	function
o_{t-1}	previous token's operation plane	In the rotation gate its even components become the rotation angle θ_k	vector
$R(\theta)$	rotation operator	Rotates each channel pair by its angle; preserves the norm, adjoint is the conjugate rotation $R(-\theta)$	operator
θ	rotation angle	The collection of rotation angles read from the operation plane; identity when the plane is 0	vector
θ_k	rotation angle of the k-th pair	Read from the $2k$-th component of the previous token's operation plane	scalar
x_f	residual stream	The residual stream at the current position, the data-plane add slot	vector
x_m	context from time mixing	The context gathered by time mixing, to which the gate applies the rotation before adding to the residual	vector
h	gate output	The gate forward-pass output; in the rotation gate $h = x_f + R(\theta)x_m$	vector
δ^h	credit on the output side	The loss gradient at the output h, the input to the adjoint computation	vector

Symbol	Name	Definition and meaning	Type
M_0, M_1, M_2	number tokens	The number tokens read one at a time in the mod 3 task; values 0, 1, 2	integer
S_0, S_1, S_2	state tokens	The answer cells stating the running sum mod 3: 0, 1, 2	integer
a	operation-plane block	The block of four operation-plane components of the previous token; the input of the quaternion gate	vector
u	offset sum	$(1 + a_0,\, a_1,\, a_2,\, a_3)$; dividing by its norm gives q	vector
q	unit quaternion	The rotor read from the operation-plane block; identity when the plane is 0	quaternion
$\bar{q}$	conjugate quaternion	The quaternion with the sign of its three vector components flipped; the inverse of a unit quaternion	quaternion
$\otimes$	quaternion product	The product of two quaternions; swapping the order changes the result	operator
v	channel block	A block of four channels of the time-mixing context x_m, the object q left-multiplies	vector
R, R^2, F	move tokens	The three moves of the triangle task: a 120-degree turn, a 240-degree turn, a flip	integer
P_0 to P_5	state tokens	The answer cells stating the six states of the marked equilateral triangle	integer
Cl(n)	Clifford algebra	The algebra built from geometric products of n directions; names the ladder rungs in the series preview	algebra

1. Introduction

Part 3 [3] measured that placing a token as an orthogonal pair of a data plane (an additive slot) and an operation plane (a multiplicative gate slot) makes the operation plane emerge as a consistent operator without labels. Yet dissecting that gate by measurement reveals expressive limits in two places. First, the sigmoid's range of 0 to 1 means the valve can only open or close; it cannot flip. Negation, the most basic logical operation, does not exist as a one-hop instruction. Second, the operation-plane components act diagonally, each adjusting only its own channel's valve, so there is no rotation connecting the components.

This paper resolves these two limits with a single prescription: read the operation plane not as valve openness but as rotation angles. Pairing the channels two by two and rotating each pair by the angle the operation plane gives, negation is expressed as 180 degrees, the composition of operations as angle addition, and a cyclic operation of arbitrary order (how many repetitions return to the start) as $360/n$ degrees. And this primitive follows the discipline of Part 1 [2] as it stands: the transpose of a rotation is the conjugate rotation, so the adjoint comes out in closed form and is verified against finite differences.

Climbing one rung by rotation raises the next question at once. The composition of pairwise planar rotations is angle addition and therefore commutative (order does not change the result). Yet not all logical composition is commutative: flipping a marked equilateral triangle and then turning it ends in a different state than turning and then flipping. Such non-commutative composition does not exist in angle addition, so an algebraic argument of exactly the same shape that separated the rotation gate from the sigmoid gate returns upon the rotation gate itself. This paper defines the quaternion gate, which reads four operation-plane components as a unit quaternion, as the next rung (Section 4), and repeats the same depth-excluded discrimination on the symmetries of the equilateral triangle, the smallest non-commutative group (Section 7). The scope of this extension was set to match the prior axiom document [1], whose first axiom states that the fundamental datatype of nature is non-commutative; by the convention of this series, the axioms remain background citation only, and the grounds for every claim here are the measurements.

The measurement is done with depth-excluded discrimination. That the design is deliberately biased is not hidden, because the bias is the hypothesis. In the standard world, logic composition is handled by depth, so restricting depth to one block in every experimental group leaves the gate as the only place to hold the logic. An ablation contrast that swaps only the gate on the same engine reveals the expressiveness ranking of that place.

1.1 Contributions

- The rotation-operator gate : we define a gate that reads the operation plane as Euler rotation angles. Negation and cyclic operations become primitives and the norm is preserved (Section 3).
- The quaternion gate : we define a gate that reads four operation-plane components as a unit quaternion and left-multiplies blocks of four channels. Non-commutative composition becomes a primitive and the norm is preserved (Section 4).
- Storage-free closed-form adjoints : we derive the adjoints of both gates. The transposes are the conjugate rotation and the conjugate left product, and the operation-plane gradients close by recovering the rotation result from the output, so no forward-pass input storage is needed. We verify them with finite differences (Sections 3, 4, 5).
- Engine-untouched injection : we show a method that attaches the new primitives by runtime kernel swapping without changing a single line of the published engine files, and prove by trajectory agreement that the swap does not change the computation (Section 5).
- Depth-excluded discrimination measurement : on a depth-stripped mod 3 task, we measure the expressiveness ranking (rotation, tanh, sigmoid, blocked, standard one-layer) with an ablation contrast that swaps only the gate (Section 6).
- The triangle discrimination and its two splits : on the triangle symmetry task, two splits that control the exposure of the rule table (the inference trial and the composition trial) measure the composition of learned rules and the inference of untaught rules as separate abilities (Section 7).
- The order argument and the primitive ladder : we explain the bias of the discrimination by the algebraic fact that no order-3 operation exists in a real scalar gate while rotation

has one (Sections 2, 6), and raise the same argument one rung: no non-commutative composition exists in commuting rotations, which the triangle discrimination measures. The algebra of the task determines the gate's primitive (Section 7).

2. Fractional Derivatives and Euler Rotation

In the frequency domain, one differentiation is multiplying each component by $i\omega$. The fractional derivative, which generalizes the number of differentiations to a real number α, multiplies by $(i\omega)^\alpha$, and this splits exactly into two factors.

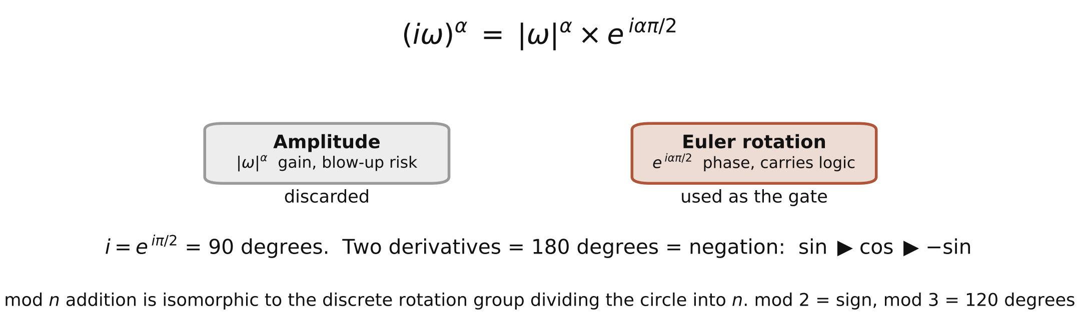
Eq. 2-1. Factorization of the fractional derivative

$$(i\omega)^\alpha \;=\; |\omega|^\alpha \times e^{i\alpha\pi/2}$$

Symbols: ω is the frequency, α is the differentiation order, $|\omega|^\alpha$ is the amplitude factor, and $e^{i\alpha\pi/2}$ is the phase-rotation factor.

How it works: i itself is $e^{i\pi/2}$, that is, a 90-degree rotation. One differentiation is 90 degrees; two make 180 degrees, the same as negation. That sin becomes $-$sin after two differentiations is that fact in substance. The amplitude factor amplifies high frequencies and risks blow-up under repeated composition, so it is discarded, and only the rotation factor, which carries the logic, is used as the gate. A pure rotation is unitary (a length-preserving transformation), so there is neither blow-up nor vanishing.

Figure 2-1. Factorization of the fractional derivative and discrete rotations



The fractional derivative splits into the product of an amplitude and an Euler rotation, and it is the rotation that carries the logic. Since i is 90 degrees, two differentiations are negation. Mod n addition is isomorphic to the discrete rotation group that divides the circle into n parts, so the algebra of the task in Section 6 is rotation itself.

One algebraic fact grounds the design of the experiment in Section 6. Mod n addition is isomorphic to the discrete rotation group of n evenly spaced marks on the circle. Mod 2 is 0 and 180 degrees, that is, sign; mod 3 is 0, 120, and 240 degrees. In the world of real scalars there is no operation that returns to the start after three multiplications while not being at the start after one (the only real solution of $g^3 = 1$ is $g = 1$). Up to negation, which has order 2, a signed scalar (tanh) can imitate it, but from order 3 on only rotation has it as a primitive.

This argument has a next rung. The composition of planar rotations is angle addition and therefore commutative, and in the world of angle addition no composition exists whose result depends on order. In exactly the shape of the argument that the sigmoid has no order 3, planar rotation has no non-commutativity. The quaternion gate of Section 4 is that rung.

3. Definition and Adjoint of the Rotation Gate

The gate of Part 3 is $h = x_f + \sigma(W_g x_f + b_g + o_{t-1})\, x_m$, where the previous token's operation plane o_{t-1} adjusts the valve inside the sigmoid. The rotation gate of this paper replaces the valve with a rotation. Channels $2k$ and $2k+1$ are tied into a pair, and the even components of the operation plane are read as that pair's rotation angle θ_k.

Eq. 3-1. Forward pass of the rotation gate

$$h \;=\; x_f \;+\; R(\theta)\, x_m, \qquad R(\theta_k)\begin{pmatrix} u_k \\ v_k \end{pmatrix} = \begin{pmatrix} \cos\theta_k\, u_k - \sin\theta_k\, v_k \\ \sin\theta_k\, u_k + \cos\theta_k\, v_k \end{pmatrix}$$

Symbols: x_f is the residual stream at the current position, x_m is the context gathered by time mixing, (u_k, v_k) is the k-th channel pair of x_m, and θ_k is the $2k$-th component of the previous token's operation plane.

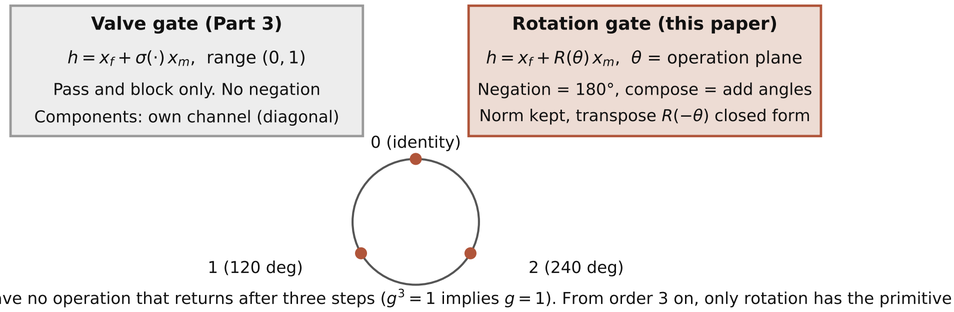
Eq. 3-2. Adjoint of the rotation gate

$$\delta^{x_m} = R(-\theta)\, \delta^h, \qquad \delta^{\theta_k} = -\,\delta^h_{u_k}\, r_{v_k} + \delta^h_{v_k}\, r_{u_k}, \qquad \delta^{x_f} = \delta^h$$

Symbols: δ^h is the credit on the output side, $(r_{u_k}, r_{v_k}) = R(\theta_k)(u_k, v_k)$ is the rotation result, and δ^{θ_k} is the angle gradient.

How it works: the transpose of a rotation is the rotation by the same angle in the opposite direction. The angle gradient is the two-dimensional cross product of the credit and the rotation result, and since the rotation result is recovered from the output as $h - x_f$, there is no need to store the forward-pass input. It is a storage-free adjoint. When the operation plane is 0, $\theta = 0$ and the rotation is the identity, the same convention as the identity start of Part 3 (the operation plane starts at 0 and emerges).

Figure 3-1. Valve gate versus rotation gate



The valve gate expresses only passing and blocking, and each component adjusts only its own channel. In the rotation gate, negation becomes the primitive of 180 degrees, composition becomes angle addition, and the norm is preserved. The circle below shows the three marks of mod 3. Since real scalars have no order-3 operation, from this algebra on only rotation has it as a primitive.

The implementation changes not a single line of the published engine files. It is an injection method that swaps in the three gate kernels (forward, backward, weight gradient) by module-attribute replacement at runtime, and that the injection does not change the computation is proved in Section 5 by showing trajectory agreement with a Python reimplementation of the original kernels. The rotation gate has no valve knob (W_g) and no bias (b_g); the rotation is driven by the operation plane alone.

4. Definition and Adjoint of the Quaternion Gate

The pairwise rotation of the rotation gate turns one plane of its own per channel pair. Different pairs never overlap, and within one pair composition is angle addition, so either way the order can be swapped without changing the result. To hold non-commutative composition as a primitive, two successive rotations must share channels across different planes. The smallest such extension is the quaternion: a number system that represents rotations with four components, with order sensitivity built into its very definition of multiplication.

The quaternion gate groups the channels in fours. The block a of four operation-plane components of the previous token gets the identity offset added to form u, is divided by its norm to give the unit quaternion q, and then left-multiplies the channel block v of the time-mixing context.

Eq. 4-1. Forward pass of the quaternion gate

$$h \;=\; x_f \;+\; q \otimes v, \qquad q = \frac{u}{\|u\|}, \qquad u = (1 + a_0,\; a_1,\; a_2,\; a_3)$$

Symbols: a is the block of four operation-plane components of the previous token, v is a block of four channels of the time-mixing context x_m, and $\otimes$ is the quaternion product. Channel 64 divides into 16 blocks, and this equation acts independently on each block.

How it works: when the operation plane is 0, $u = (1, 0, 0, 0)$ and q is the identity, the same identity-start convention as the rotation gate. Left multiplication by a unit quaternion preserves the norm, so there is neither blow-up nor vanishing. And the left product rotates the two overlapping planes shared by the four channels of the block by the same angle at once, so two successive left products give different results when their order is swapped. The non-commutative composition that the rotation gate cannot express is, in this gate, the multiplication rule itself.

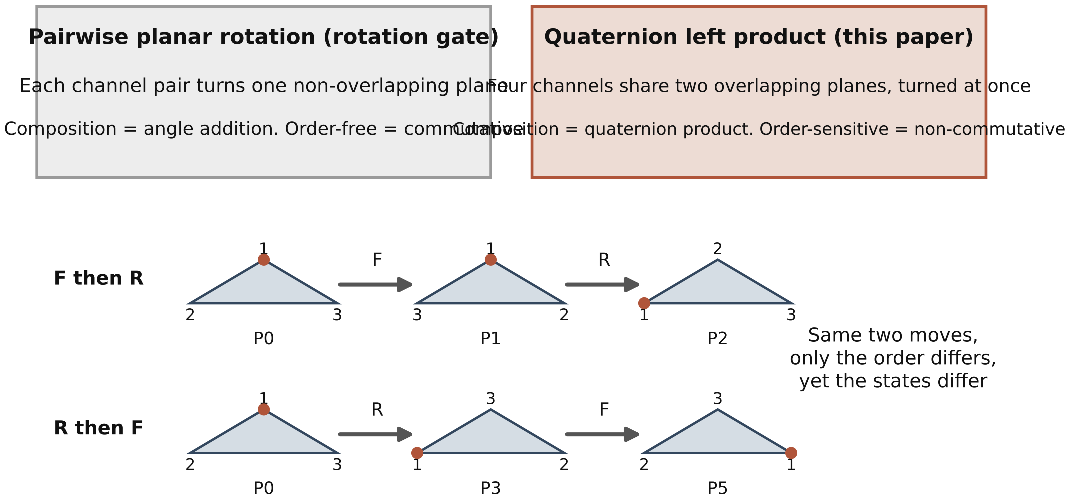
Eq. 4-2. Adjoint of the quaternion gate

$$\delta^v = \bar{q} \otimes \delta^h, \qquad \delta^q = \delta^h \otimes \bar{r} \otimes q, \qquad \delta^u = \frac{\delta^q - q\,(q \cdot \delta^q)}{\|u\|}, \qquad \delta^{x_f} = \delta^h$$

Symbols: $\bar{q}$ is the conjugate quaternion, $r = q \otimes v$ is the left-product result, $q \cdot \delta^q$ is the inner product of the four components, and δ^u is the gradient carried back through the normalization to the operation-plane block.

How it works: the transpose of a unit-quaternion left multiplication is the conjugate left multiplication. The conjugate of v needed for the quaternion gradient δ^q is recomputed from the left-product result as $\bar{v} = \bar{r} \otimes q$, and the left-product result is recovered from the output as $r = h - x_f$, so, just as in the rotation gate, no forward-pass input needs to be stored. The last term is the backward pass of the normalization: it removes the component along q from the gradient and divides by the norm. The quaternion gate also has no valve knob and no bias; the rotation is driven by the operation plane alone. The implementation is the same kernel injection as in Section 3, so the published engine files remain untouched.

Figure 4-1. Pairwise planar rotation versus quaternion left multiplication



On the left, pairwise planar rotation turns one non-overlapping plane per channel pair, so composition is angle addition and commutative. On the right, quaternion left multiplication turns the two overlapping planes shared by four channels at once, so the order of composition changes the result. Below, the two moves of the triangle task applied in swapped order end in different states.

5. Measurement Preparation: Adjoint Verification and Design Proof

Measured

The new primitives enter the experiments only after passing two checks.

- Rotation-gate adjoint verification : checking the transpose of Eq. 3-2 with the inner-product identity $\langle \delta, R(\theta)x \rangle = \langle R(-\theta)\delta, x \rangle$ gives a difference of 0.00e+00 (exactly 0 in float64), and the median ratios of the mixing gradient and the angle gradient of a scalar loss against central finite differences are both 1.000000 .
- Quaternion-gate adjoint verification : checking the transpose of Eq. 4-2 with the inner-product identity $\langle \delta, q \otimes v \rangle = \langle \bar{q} \otimes \delta, v \rangle$ gives a difference of 0.00e+00, and the median ratios of the mixing gradient and the storage-free operation-plane gradient (with the left-product result recovered as $h - x_f$) against central finite differences are both 1.000000. In the same check, the mean norm of the difference after swapping the order of two left products is 1.912, a value that would be 0 for pairwise planar rotations. This measures that the gate actually owns non-commutative composition.
- Design proof : injecting a Python reimplementation of the original sigmoid kernels and training side by side for 2,000 steps on seed 0, the cross-entropy trajectory at five points (0.1141, 0.0048, 0.0002, 0.0001, 0.0000) is identical at the displayed digits, with a maximum difference below 1e-07, at float32 noise level (this floor itself

wobbles from run to run with GPU summation order). The design of kernel injection itself does not change the computation. The rotation gate and the quaternion gate both enter through this single injection.

Reproduce

The adjoint verifications, the design proof, and the entire discrimination measurements of Sections 6 and 7 are reproduced with the following one-liner after completing the preparation in Appendix C.

```
cd probe && python3 paper7_rotation_probe.py
```

6. Depth-Excluded Discrimination: the mod 3 Task

6.1 Task and conditions

The task is to read three kinds of numbers (M_0, M_1, M_2) one at a time and state, at each position, the mod 3 state of the sum so far (S_0, S_1, S_2). The task uses seven tokens: one pad that fills empty cells, the number tokens M_0, M_1, M_2 (values 0, 1, 2), and the state tokens S_0, S_1, S_2 (the running sum modulo 3: 0, 1, 2). Length is the number of numbers read in a row within one problem. A problem is built by filling the left of the 16-cell window with pad and then alternating number and state; right after each number comes the cell that holds the state at that point. The model's only job is to fill those state cells; number cells and pad are not scored.

Example. The window of the length-3 problem $M_1\,M_2\,M_0$ is ten pads followed by $M_1\,S_1\,M_2\,S_0\,M_0\,S_0$. Reading M_1 makes the sum 1, hence S_1; reading M_2 makes the sum 3, remainder 0, hence S_0; reading M_0 leaves the sum at 3, hence S_0 again. The rule is single: each number read advances the state forward by that number.

Training uses only all 9 length-2 combinations and 14 of the 27 length-3 combinations drawn with a fixed seed, 23 problems in total; the remaining 13 of length 3 are hidden. Scoring has three bins. Seen is the 23 trained problems. Blanks are the 13 hidden combinations of the same length 3, the bin that catches a solution that memorized the whole 27-entry table. Unseen lengths are the full sets of length 4 (81 items), 5 (243 items), and 6 (729 items); training contained no problem of these lengths, so only a model that holds the rule itself, not the table, can keep applying it to longer problems.

There are two scoring modes. Forced keeps the true states in the window and scores the prediction at each state cell; even when an earlier step is wrong, the next step restarts from the truth, so this isolates the per-step rule alone. Self-feedback (Self for short) blanks all state cells, writes the model's predicted first state into the window, predicts the next state from that window, and so on, feeding the model's own output back, and scores the final state. Earlier errors become later inputs and accumulate, which makes this the real loop and the honest score.

The five experimental groups are as follows. A (sigmoid gate, operation plane on), B (operation plane blocked), D (tanh gate), and E (rotation gate) are an ablation contrast that swaps only the gate on the same Banya one-block engine (channel 64, 4 mixing heads, identical filter ladder). C is a standard transformer [4] one-layer of the same width. Every group trains for 12,000 steps on three seeds (0, 1, 2). In this protocol, which proceeds while writing the state, chaining is handled by the loop, so what remains per step is a single cyclic operation, and what separates the groups is who has that single operation as a primitive.

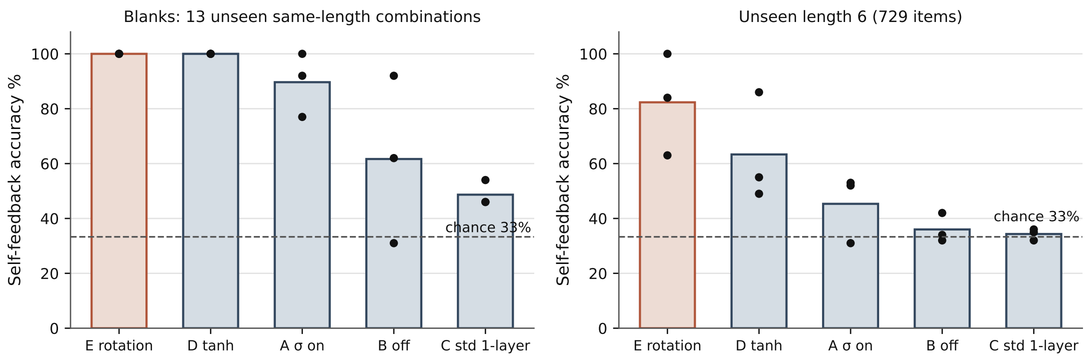
6.2 Results

Measured

Measured under the self-feedback criterion (each of the three seeds; parentheses are the mean).

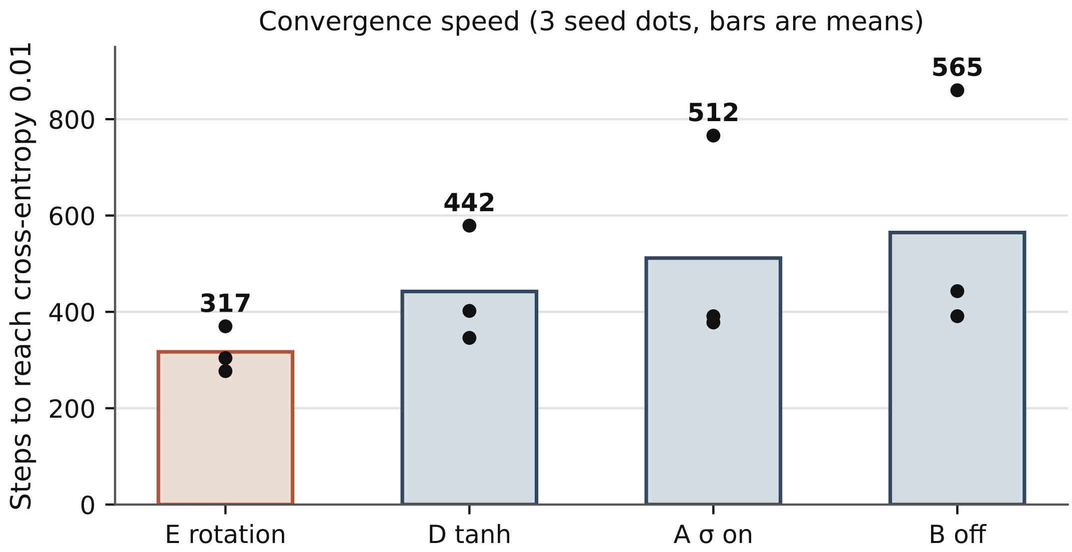
- Blanks (13 of the same length 3) : E rotation 100, 100, 100 (100.0) · D tanh 100, 100, 100 (100.0) · A sigmoid 77, 100, 92 (89.7) · B blocked 31, 62, 92 (61.7) · C standard one-layer 46, 54, 46 (48.7). Chance is 33.3 percent.
- Unseen length 6 (729 items) : E 84, 63, 100 (82.3) · D 86, 49, 55 (63.3) · A 53, 52, 31 (45.3) · B 34, 42, 32 (36.0) · C 35, 36, 32 (34.3).
- Convergence (steps to reach cross-entropy 0.01) : E 370, 304, 277 (mean 317) · D 579, 346, 402 (442) · A 766, 378, 391 (512) · B 860, 443, 391 (565). E is the fastest with the smallest variation across seeds.
- Seen : all five groups are at 100 percent in both forced and self modes. The split comes not from memorization but from the blanks and the lengths.

Figure 6-1. mod 3 discrimination results



Left: the 13 blanks of the same length; right: the unseen length 6. Bars are the mean over three seeds and dots are the individual seeds. The ranking is rotation, tanh, sigmoid, blocked, standard one-layer, following exactly the expressiveness of the operation held in the gate. The standard one-layer is at 100 percent only on what it has seen and is near chance from the blanks on.

Figure 6-2. mod 3 convergence speed



Steps to reach cross-entropy 0.01. Bars are the mean over three seeds and each dot is one seed's value. The side that holds the right primitive wanders less. The rotation gate is fastest at a mean of 317 steps, with the smallest deviation across seeds.

The reading is the order argument as it stands. The per-step operation of this task is a cycle of order 3. The rotation gate has that operation as a primitive, so it fills the blanks on all three seeds and is also highest on the unseen lengths. tanh has primitives only up to sign (order 2), so it fills the blanks but wavers on the lengths; the sigmoid lacks even sign and relies on roundabout representations; and blocking the operation plane removes the instruction itself, forcing reliance on the local memorization of the filter ladder. The standard one-layer, with its depth confiscated, memorizes only what it has seen and drops toward chance from the blanks on.

6.3 Group-structure check and the limits of the instrument

We read the angles directly from the operation plane of the trained rotation gate and computed circular statistics. The angle differences between number tokens cluster around a mean of 0 degrees with a concentration of 0.98 to 0.99, meaning the number tokens' operation planes stayed at identity (0) because that slot receives no training signal; this is measured evidence of the identity-start convention. The angle differences between state tokens have a low concentration of 0.54 to 0.79, and their means did not show a 120-degree structure either. However, this check is a circular mean over all 32 channel pairs, so the structure of the few channels carrying the state machine can be diluted by the noise of the rest. Reopening the check with only the important channels weighted is a next task, and the accuracy and convergence of Section 6.2 stand independently of this check.

6.4 Comparison across three protocols

Running the same task family under three protocols yields one structural proposition. Under the one-shot protocol (demanding only the final answer with no intermediate state), all five groups memorize the training data but do not generalize beyond the length at all. Inside a block, time mixing reads only the block input, so the gate's processing result does not carry over to the next position within the same block; chaining is therefore the business of block depth or of the autoregressive loop. Under the loop protocol, which proceeds while writing the state, the per-step task reduces to a local rule and the gate differences reappear. Reduced to a proposition: depth and the loop set the number of chainings, and the gate sets the quality of the operation carried in one chaining. The depth-excluded discrimination of this paper measured that quality.

7. Depth-Excluded Discrimination: the Triangle Task

7.1 Task and the two splits

The task is an equilateral triangle with a marked vertex. Reading the three kinds of moves one at a time, namely R (a 120-degree turn), R^2 (a 240-degree turn), and F (a flip), the model states the triangle's state at that point (P_0 to P_5) at every position. The states are exactly the six combinations of three turns and two flip states, and P_0 is the starting state. The symmetry system formed by these six states and three moves is the smallest non-commutative group. Flipping first makes subsequent turns appear reversed, so the state after F then R equals the state after R^2 then F (P_2), while R then F ends elsewhere (P_5). The rule table has 18 entries, one per pair of the six states and three moves. The window layout, the training amount (all 9 length-2 combinations and 14 of the 27 length-3 combinations), the 13 blanks, the full unseen lengths 4 (81), 5 (243), and 6 (729), and the two scoring modes, forced and self-feedback, are all identical to Section 6, and chance is one state in six, 16.7 percent.

> Example. The window of the length-2 problem $F\,R$ is twelve pads followed by $F\,P_1\,R\,P_2$: reading F gives the flipped state P_1, and then reading R gives the flipped-and-turned state P_2. The order-swapped problem $R\,F$ has the window $R\,P_3\,F\,P_5$, ending in a different state. In the mod 3 task, swapping the order never changed the final state; this difference is the whole of what separates the two tasks.

Two splits are run. The reason the same experiment is run twice with only the split changed is that measurement revealed that what a split exposes changes the very ability being tested.

- The inference trial : the same fixed-seed random split as Section 6. A posterior check shows this split exposes only 16 of the 18 rule-table entries to training. Of the 13 blanks, only 9 (69 percent) are solvable from the exposed rules alone; the remaining 4 require inferring 2 rule entries that never appeared in training. This split therefore tests not the composition of learned rules but also the inference of untaught ones.
- The composition trial : a coverage-guaranteed split chosen by search so that all 18 rule-

table entries are exposed. All 13 blanks are solvable purely by composing learned rules. The discrimination that ranks the gates' expressiveness is this one.

How far an order-blind solver can go is also fixed by arithmetic in advance. A solver that sees only the multiset of moves (what was read, how many times) can give one answer per multiset, so it is sure only on blanks whose final state is the same across every ordering. This commutative ceiling is 3 blanks (23 percent) under the inference-trial split and 5 blanks (38 percent) under the composition-trial split. Any score above it is evidence that order information was actually used.

The experimental groups are the five of Section 6 plus Q (the quaternion gate). Q is likewise an ablation contrast that swaps only the gate on the same Banya one-block engine as A, B, D, and E, with channel 64 divided into 16 quaternion blocks. The seeds are increased to six (0 to 5) so that seed-to-seed variation is measured as well; all other conditions (12,000 steps, batch 32) match Section 6.

7.2 Results

Measured

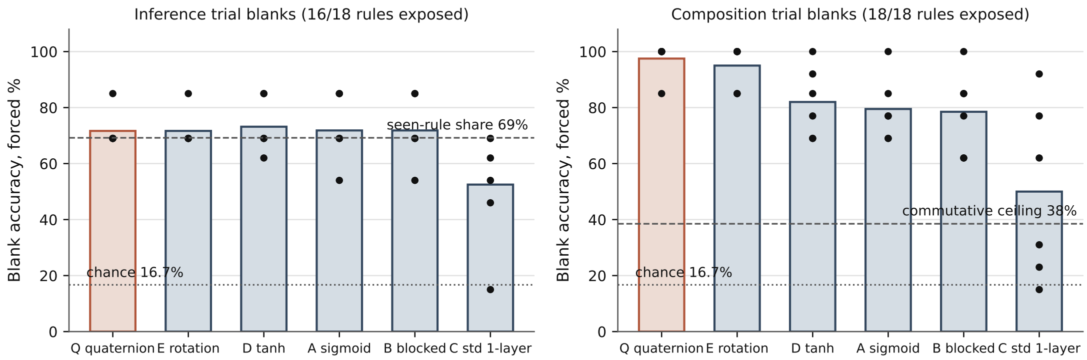
Blanks are measured under the forced criterion (each of the six seeds; parentheses are the mean). This bin asks whether the rule table was learned, so forced scoring, which isolates the per-step rule, is the criterion; the five engine groups' self-feedback scores on the blanks were identical to their forced scores, and the separate axis of long-problem self-feedback is listed in its own item below.

- Inference-trial blanks (16/18 rules exposed) : E rotation 69, 69, 69, 85, 69, 69 (71.7) · Q quaternion 69, 69, 69, 85, 69, 69 (71.7) · A sigmoid 54, 69, 85, 85, 69, 69 (71.8) · B blocked identical to A on all six seeds (71.8) · D tanh 62, 69, 85, 85, 69, 69 (73.2) · C standard one-layer 69, 54, 46, 15, 62, 69 (52.5). The two geometric gates are identical to each other seed by seed and never fall below the 69-percent share of seen rules on any seed. A, B, and D have seeds that fall below the share (54, 62). The 85s above the share appear jointly across several groups on seeds 2 and 3 and vanish on the rest, so there is no evidence beyond that seed's training order happening to match the 2 untaught rule entries. Together with A and B being identical seed by seed, what separates the groups under this split is not the gate.
- Composition-trial blanks (18/18 rules exposed) : Q quaternion 85, 100, 100, 100, 100, 100 (97.5) · E rotation 85, 85, 100, 100, 100, 100 (95.0) · D tanh 69, 92, 69, 100, 85, 77 (82.0) · A sigmoid 69, 85, 69, 100, 77, 77 (79.5) · B blocked 62, 85, 62, 100, 85, 77 (78.5) · C standard one-layer 77, 62, 23, 15, 31, 92 (50.0). Chance is 16.7 percent. The ranking is the ladder itself.
- Convergence (steps to reach cross-entropy 0.01) : Q 289, 280, 274, 275, 299, 309 (mean 288) · E 332, 346, 259, 307, 395, 366 (334) · the other three engine groups average 396 to 413. The side that holds the right primitive wanders least, with the

smallest variation across seeds.

- Unseen lengths 4, 5, 6, forced : Q stays uniform at 94 to 100 percent across all six seeds and all three lengths. E ranges over 73 to 99 percent depending on the seed, and the remaining groups swing wider. Only a group that has learned the rule table stably keeps applying it evenly to longer problems.
- Self-feedback at long lengths : at length 6 under self-feedback, Q 91, 43, 54, 25, 98, 87 (66.3) and E 11, 21, 30, 41, 54, 61 (36.3). With six states, one wrong step snowballs faster than in mod 3; this is a compounding-error axis separate from rule acquisition, and here too Q declines most gently.

Figure 7-1. Triangle discrimination results



Left: the inference-trial blanks, where every group's mean sits near the 69-percent share of seen rules and no score above the share persists across seeds. Right: the composition-trial blanks, ranked quaternion, rotation, tanh, sigmoid, blocked, standard one-layer. Bars are the mean over six seeds and each dot is one seed's value.

7.3 Ceiling autopsy

The ceiling of the inference trial is not a defect of any gate. The 2 unexposed rule entries appeared nowhere in any window, so the 4 blanks that need them sit where no training signal reaches. That the rotation and quaternion gates never fall below the share of seen rules on any seed while never sustaining a score above it across seeds shows that a gate's expressiveness is the ability to compose learned rules, not the ability to invent untaught ones. Expressiveness and inference are different abilities, and a discrimination design must not conflate them. The composition trial is that correction.

The split in the composition trial is autopsied by rule-table scoring. The trained model is run over all 27 length-3 problems with the true states shown, and every state cell is scored under the rule-table entry it uses. The autopsy has two layers. First, the remaining failures of both geometric gates are entirely entries of the starting state P_0 (E misses one or two entries on two seeds and Q one entry on one seed; every other seed is correct on all 18 entries). The training exposures of the P_0 entries sit mostly at the first move position of the window, and positions where the state returns to P_0 mid-problem are rare in training, so these failures are entangled not with non-commutativity itself but with position generalization, reusing the same rule at a different position. The conclusive autopsy of the failed entries is left as a next task. Second, every engine group's blank score sits far above the 38-percent commutative ceiling because the loop protocol externalizes order into the states written in the window, reducing the per-step task to a lookup of state and move. Order handling is carried by the loop; the split comes from how fast and how stably the non-commutative rule table is acquired.

7.4 The primitive ladder

Overlaying the two discriminations leaves one proposition. The per-step operation of the mod 3 task is a commutative cycle, and the rotation gate, which holds that operation as a primitive, was perfect on three seeds with the fastest convergence in Section 6. The per-step operation of the triangle task is non-commutative composition, and the quaternion gate, which holds that composition as its multiplication rule, is first on every item: blank mean, convergence speed, uniformity on unseen lengths, and the gentlest self-feedback decline. The rotation gate, closing under angle addition, can still learn the whole table on some seeds by leaning on the loop, but its seed-to-seed swing is large and failing seeds remain. The valve has passing and blocking; the signed scalar has order 2; planar rotation has commutative cycles of any order; the quaternion has non-commutative composition. The gates' primitives form a ladder, learning is fastest and most stable when the algebra of the task matches the rung, and the discrimination is the instrument that confirms that fit by measurement.

8. Limitations and Next Measurement Tasks

The reference class of the baseline is a single developer, a self-built exact-adjoint backpropagation engine, and one consumer graphics processing unit. The following are not defects but the next measurement tasks.

- Small synthetic task : the measurements in this paper use a one-block engine of channel 64 and the mod 3 and triangle synthetic tasks. The gain on natural language has not yet been measured. Training the rotation gate and the quaternion gate at the published model size (channel 1024, 16 blocks) and repeating the entire measurement of Part 3 is the next task.
- Seeds and statistics : three seeds per group in the mod 3 discrimination and six in the triangle discrimination. The triangle composition-trial blanks swing widely across seeds, which is why the seeds were increased to six; the ranking of the means matches the ladder, but increasing the repetitions further to pin down the variance is a next task.
- Direct observation of the group structure : the circular mean of Section 6.3 does not weight the important channels, so it could not adjudicate the 120-degree structure of the state tokens. A recheck weighted by channel contribution is a next task.

- Identity of the failed entries : the rule-table scoring of Section 7.3 showed the rotation gate's failure localized to specific entries but could not adjudicate why those entries. Tracing the algebraic common ground of the failed entries is a next task.
- Self-feedback compounding error : in long triangle problems the gap between forced and self-feedback scores widens. This is a compounding-error axis separate from rule acquisition, and the recovery mechanism after a wrong state lands in the window is a separate capability this paper does not treat.
- Complex exponential gate : with base e, magnitude and rotation cohabit in a single exponent ($e^{a+i\theta} = e^a e^{i\theta}$). This paper treated only the blow-up-free pure rotation ($a = 0$); the extension that learns a tightly bounded magnitude factor alongside is a next task.
- Time-delay rotation : applying the rotation not to channel pairs but to the frequency coefficients of time mixing becomes a time delay by the shift theorem. This is a separate capability, an operator instructing which past to look at, so it is left for a separate experiment.

9. Conclusion

This paper derived the rotation-operator gate from the factorization of the fractional derivative, showed that its adjoint is a storage-free closed form and verified it with finite differences, and presented the method that attaches the new primitive by injection without changing a single line of the published engine, together with the proof of its harmlessness. In the depth-excluded mod 3 discrimination, the ablation contrast that swaps only the gate revealed the expressiveness ranking: rotation, tanh, sigmoid, blocked, standard one-layer. Only rotation and tanh filled the blanks on all three seeds, and on unseen lengths and convergence speed rotation was highest and fastest. The order argument is the algebraic reason for this ranking. Negation has order 2, so a signed scalar can imitate it, but from order 3 on only rotation holds the primitive, and since mod n is isomorphic to the discrete rotation group, the natural coordinate system of cyclic logic is the angle.

The same question was then raised one rung. We defined the quaternion gate, which reads four operation-plane components as a unit quaternion, verified that its adjoint is likewise a storage-free closed form closing through the conjugate left product and a recovery term, and discriminated on the triangle task, the smallest non-commutative group, under two splits that control the exposure of the rule table. In the composition trial, which exposes every rule, the quaternion gate was first on every item: blank mean, convergence speed, uniformity on unseen lengths, and the gentlest self-feedback decline, and the remaining failures of both geometric gates localized entirely to the starting-state rules. In the inference trial, which hides part of the table, no score above the share of seen rules persisted across seeds and the two geometric gates never fell below it, so that expressiveness and inference were measured apart as different abilities. A commutative task needs only rotation; on a non-commutative task, rotation in overlapping planes is the fastest and most stable vessel. The algebra of the task determines the gate's primitive. Reading the operation plane as rotation angles is the path that resolves with a single prescription the two expressive limits left by the gate of Part 3 (absence of sign, diagonal action); the quaternion is the next rung of the primitive ladder; and its value was established by measurement in these small discrimination settings.

10. Series Preview

Series preview

- Part 8 (planned) — The Clifford ladder : planar rotation corresponds to the even part of the Clifford algebra (the algebra built from geometric products of directions) Cl(2), and the quaternion to the even part of Cl(3). The next part designs a discrimination task per rung from Cl(4) to Cl(7) and confirms this ladder by measurement, rung by rung. One prediction that the mathematics hands down in advance is fixed as a scoring criterion: the even part of Cl(4) decomposes into two quaternions in parallel, so a Cl(4) gate must hold no new primitive beyond two copies of the quaternion gate (zero gain). The prior axiom document [1] says the same in advance: no Cl(4) operator exists. The decomposition theorem of mathematics and the prophecy of the axioms point at the same hole, and whether that hole is confirmed by measurement is the first scoring of the next part. The Clifford formalization in the background axiom document is used as citation only; the claims come from measurement.
- Final part (unfinished) — Combining similarity consolidation and inference traversal : self-utterance weaving data-plane similarity consolidation (Part 2) with operation-plane inference traversal. The property of this paper, that the composition of rotations is angle addition, is a candidate for the arithmetic of that traversal. Not yet covered, as the final assembly is incomplete.

References

[1] Han, H. (2026). Banya Framework: An Axiom-Based Science Mining Engine for Deriving Fundamental Physical Constants. Zenodo. https://doi.org/10.5281/zenodo.20301304 . (Background natural philosophy of this series; prior axiom document)

[2] Han, H. (2026). A Language-Model Training Engine Using Hand-Derived Closed-Form Exact Adjoints Without Automatic Differentiation. Banya Research Series, Part 1. Zenodo. https://doi.org/10.5281/zenodo.21385447

[3] Han, H. (2026). Orthogonal Data-Operation Tokens and Convolutional Scale-Depth Structure: Label-Free Emergence of Operation-as-Weight. Banya Research Series, Part 3. Zenodo. https://doi.org/10.5281/zenodo.21385676

[4] Vaswani, A. et al. (2017). Attention Is All You Need. NeurIPS. arXiv:1706.03762.

Appendix A. Experimental Configuration

Table A-1. Depth-excluded discrimination configuration

Item	Value	Item	Value
Channel width H	64	Context length T	16
Block count N	1	Channel transforms per block	2
Mixing heads	4	Vocabulary	8 (pad, 3 numbers, 3 states)
Training steps	12,000	Batch	32
SGD learning rate	0.2	Adam learning rate	0.002
Seeds	0, 1, 2	Standard one-layer (C)	same width, AdamW 0.001

A, B, D, and E differ only in the gate on the same engine with the configuration above. C is a standard transformer one-layer of the same width and the same data flow.

Table A-2. Triangle discrimination configuration (only items differing from Table A-1)

Item	Value
Vocabulary	10 (pad, 3 moves, 6 states)
Groups	A, B, D, E, Q, C
Splits	inference (random), composition (coverage-guaranteed)
Chance	16.7%
Quaternion blocks	channel 64 in 16 blocks of 4
Rule table	6 states $\times$ 3 moves = 18 entries
Seeds	0 to 5 (six)

Items not listed (channel width, block count, steps, batch, learning rates) match Table A-1.

Appendix B. Measurement Environment

Table B-1. Computer specifications used for measurement

Category	Specification
Graphics processing unit (GPU)	NVIDIA GeForce RTX 3090 Ti (24 GB)
Operating system	Ubuntu 24.04.4 LTS
Software	Python 3.12.3, numpy 2.5.1, cupy 14.1.1, CUDA 13.3.1, (optional) torch 2.13.0
Measurement target	87 small one-block ablation trainings (72 engine, 15 standard one-layer) and the adjoint and design checks

Even with fixed seeds, figures may vary in the last digit depending on GPU summation order.

Appendix C. Reproduction Guide

Reproducing this paper requires neither model checkpoints nor corpora, because the task data are generated inside the probe.

Step 1, requirements. On an NVIDIA GPU (CUDA) with Python 3.12, install the following. To reproduce the standard one-layer contrast (C) as well, install torch alongside; if it is absent, the probe skips only that group and says so.

```
pip install -r requirements.txt
```

Step 2, get the code. Download and unzip the repository archive.

```
wget https://ubmscoin.github.io/banya/banya_ai_paper_code/banya_ai_paper_code.zip
unzip banya_ai_paper_code.zip
```

Step 3, run the probe.

```
cd banya_ai_paper_code/probe && python3 paper7_rotation_probe.py
```

Table C-1. Correspondence between the probe and the figures in the text

Probe	Figures reproduced	Output to check
paper7_rotation_probe.py	Section 5 adjoint verifications (rotation, quaternion) and design proof, entire Sections 6 and 7 discriminations	inner-product identity difference 0, median gradient ratio 1.000000, trajectory agreement, mod 3 blanks E 100% on three seeds, triangle composition-trial blanks Q mean 97.5% and convergence mean 288 steps

The rotation-gate and quaternion-gate primitives themselves are in paper7_rotation.py and attach by runtime injection without changing the published engine files. The package's folder layout and per-file descriptions are on the index page.

Banya Research Series, Part 7 (the rotation-operator gate). All figures were measured in the environment of Appendix B. The background natural philosophy is cited as the prior axiom document [1] but is not the basis of this paper's claims. Part 1 (engine) and Part 3 (bi-token) are cited; this paper derives from the gate's expressive limits in Section 6 of Part 3.