

Transformers as Effective Fields

A quantum-physics theory of AI



Feature



State



Architecture



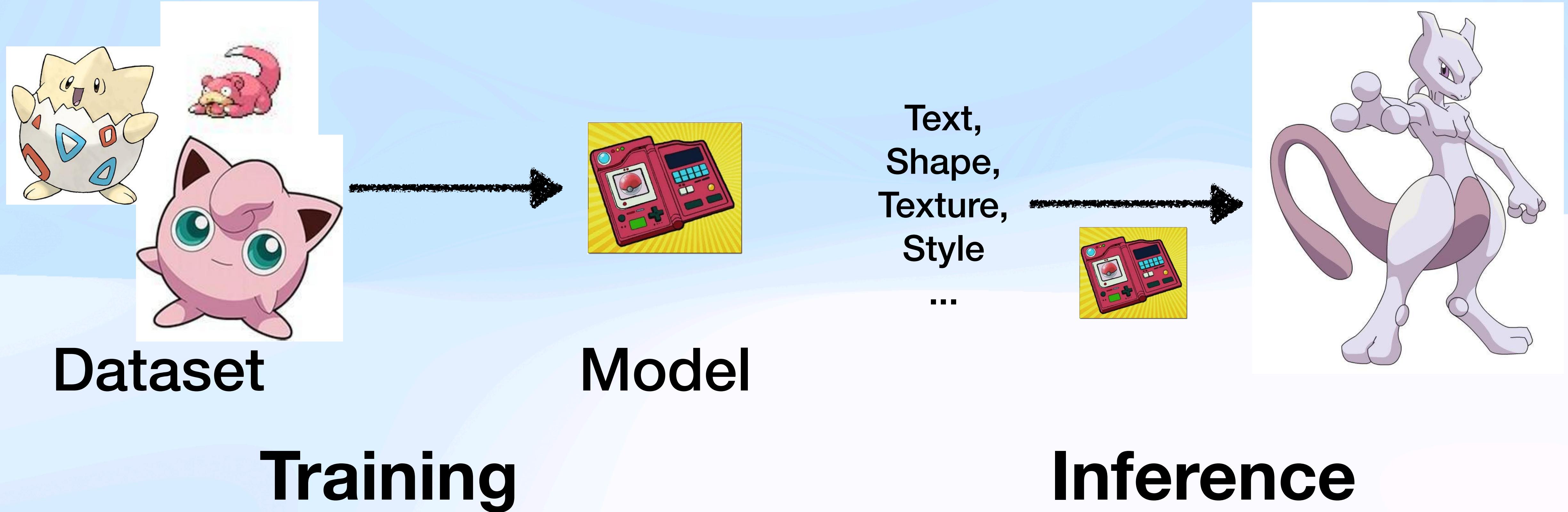
Learning



Evaluation

Changqing Fu, presented at ML Collective, November 2025

Generative Models

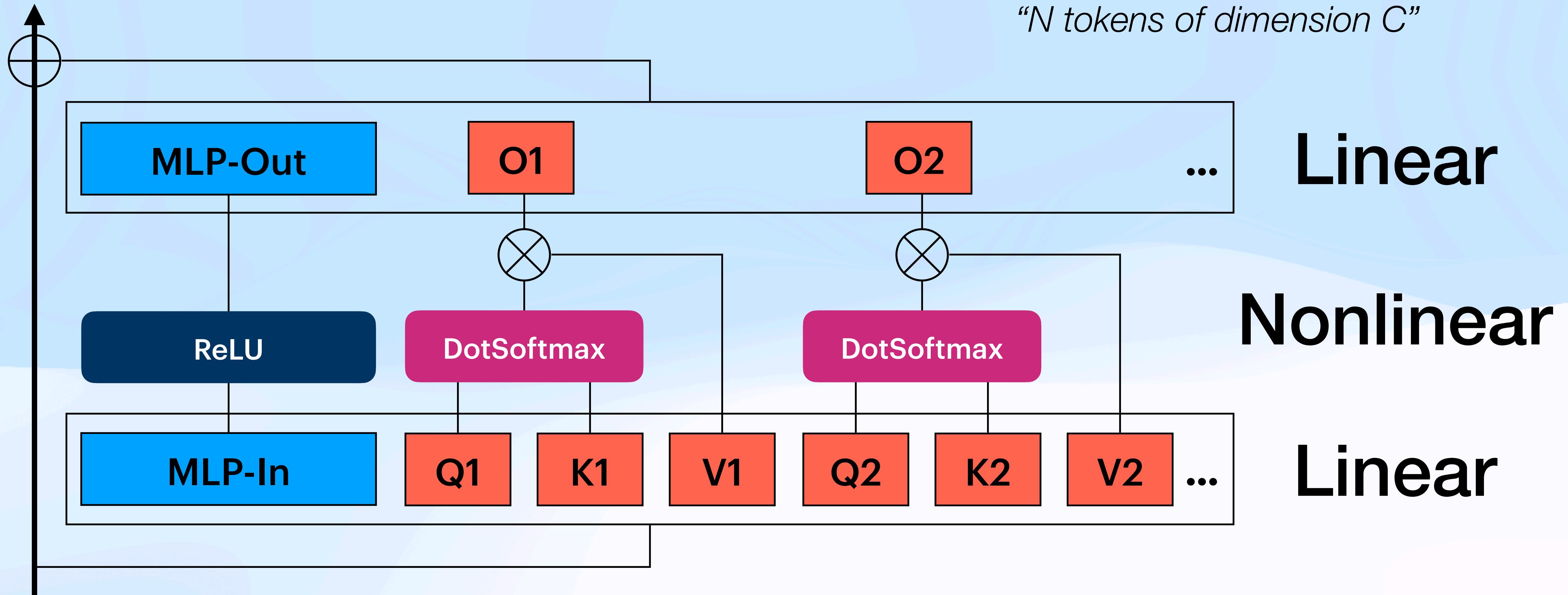




What happened inside?

Transformers $X : [0, 1] \rightarrow \mathbb{C}^{N \times C}$

" N tokens of dimension C "



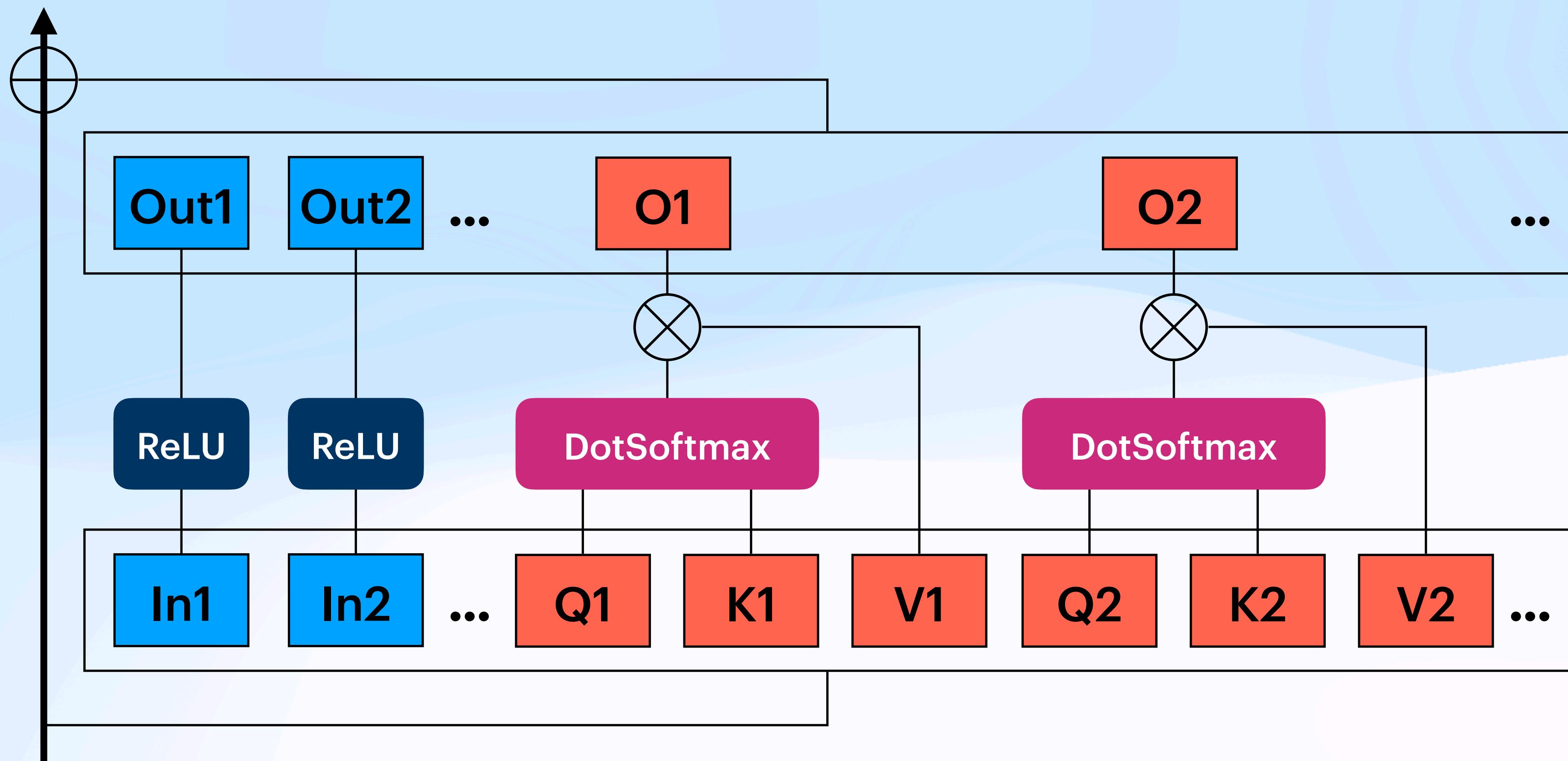
Discrete

$$X(t+1) = X(t) + W'(t) \circ \lambda \circ W(t)(X(t))$$

Continuous $\dot{X} = W' \circ \lambda \circ W(X)$

$$X \in \mathbb{C}^{N \times C}, W_{\text{in},i}, W_{\text{out},i}, Q_i, K_i, V_i, O_i \in \mathbb{C}^{C \times r}$$

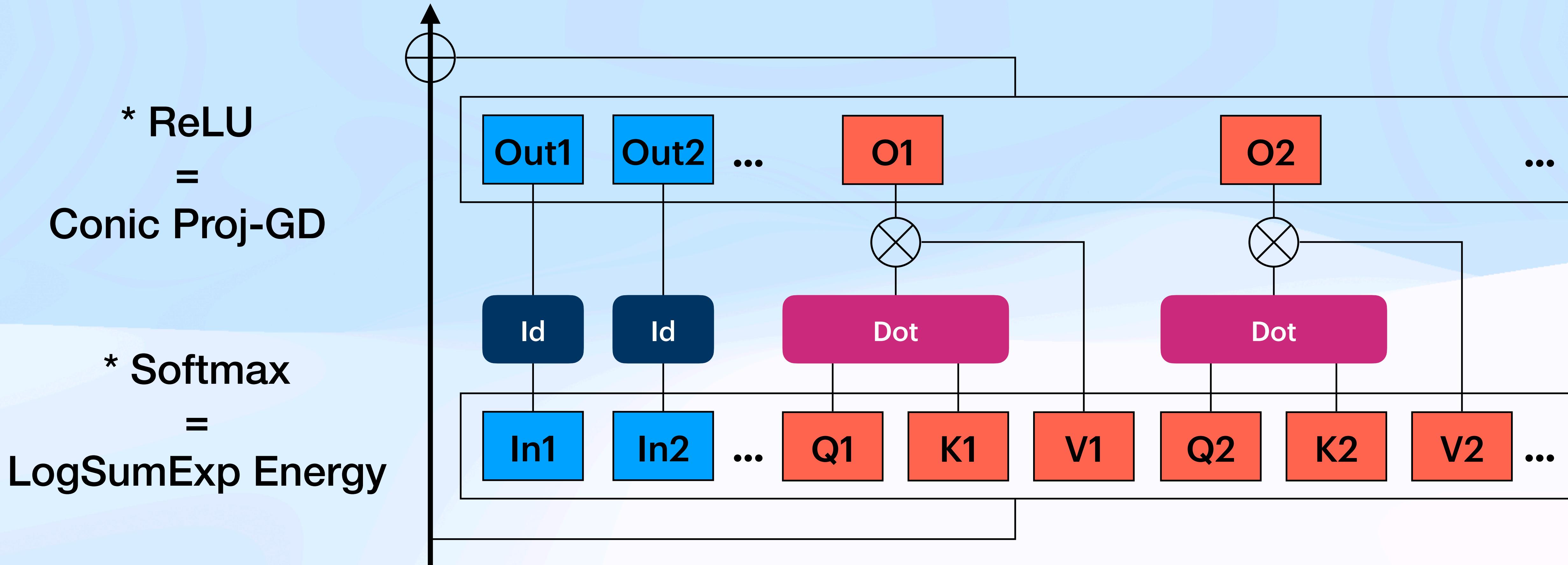
MultiHeadMLP
= Wide MLP



$$\dot{X} = \sum_i (\textcolor{red}{X} W_{\text{in},i})_+ W_{\text{out},i}^\top + \sum_i \text{Softmax}(C^{-1/2} \textcolor{red}{X} Q_i K_i^\top X^\top) \textcolor{red}{X} V_i O_i$$

Theorem: Linear Transformers are Canonical

(Without loss of Generality*)



$$\dot{\mathbf{X}} = \sum_i \mathbf{X} \mathbf{W}_{\text{in},i} \mathbf{W}_{\text{out},i}^\top + \sum_{i_6} C^{-1/2} \mathbf{X} \mathbf{Q}_i \mathbf{K}_i^\top \mathbf{X}^\top \mathbf{X} \mathbf{V}_i \mathbf{O}_i$$

“In-Context Learning” desires $\dot{X} = \frac{\partial V}{\partial X^*}$

$$V(X) = \frac{i}{2} \text{Tr}(XWXX^*) + \frac{i}{2\sqrt{C_A}} \sum_{i=1}^{n/2} \text{Tr}(XW_{Ai}X^*XW_{Bi}^*X^*)$$

$\frac{\partial V}{\partial X^*} = \text{Transformers Architecture!}$

Why Is $i\dot{X} = \frac{\partial V}{\partial X^*}$ Canonical?

Euler-Lagrange Eq
Schrödinger Eq
Feynman Path Integral

$$Z = \int \exp(iS(X)) \mathcal{D}X$$

$$S(X) := \int_0^1 \text{Tr}\left(\frac{i}{2}(\dot{X}X^* - X\dot{X}^*) - V(X)\right) dt$$

$$X = \frac{q + ip}{\sqrt{2}}, \quad X^* = \frac{q - ip}{\sqrt{2}}, \quad \frac{1}{2}(\dot{X}X^* - X\dot{X}^*) = \frac{i}{2}(X^*\dot{X} - \dot{X}^*X) = p\dot{q}$$

Legendre Transform

$$L^*(p) = \max_{\dot{q}} p\dot{q} - L(\dot{q})$$

$$\omega = \frac{i}{2}((dX)X^* - XdX^*) \quad d\omega = i dX \wedge dX^*$$

Symplectic Form

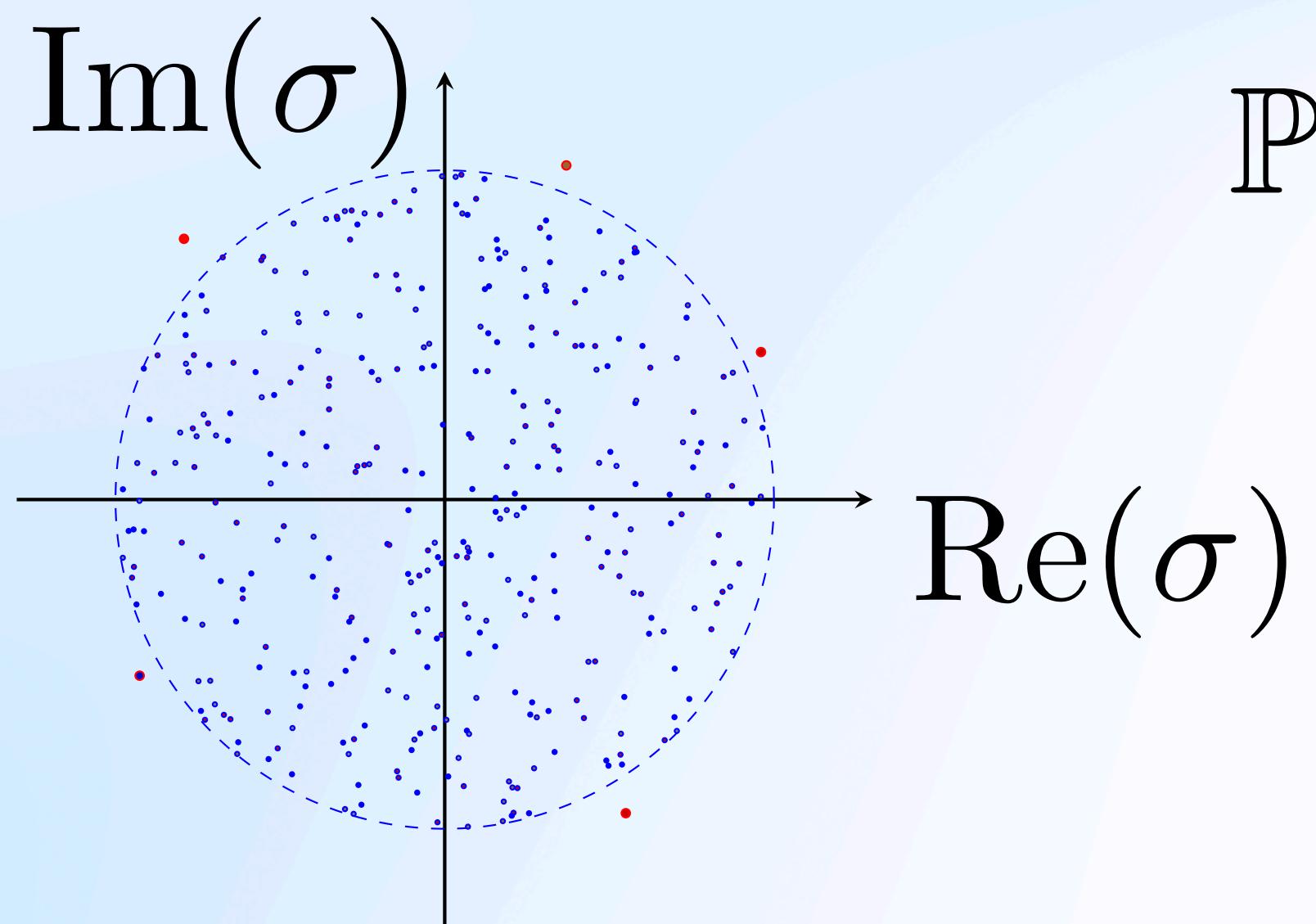
Why Is V “Effective”?

NO MORE TERMS!!!

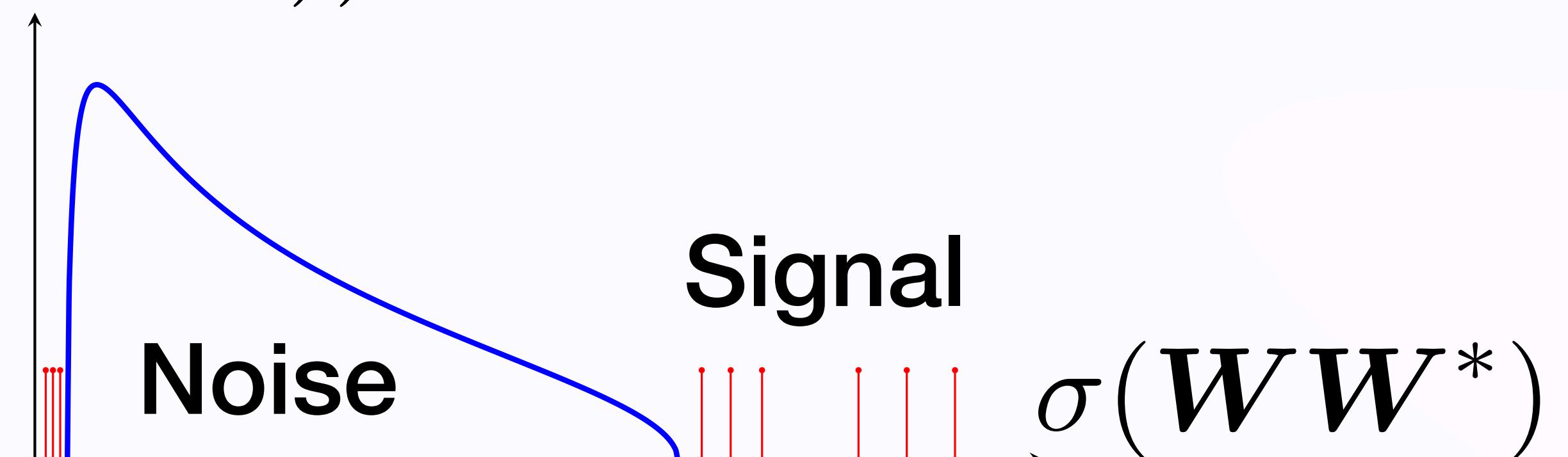
Axiom 1: Real Analyticity $V = V_2 + V_4 + O(\|X\|^6)$, $X/\hbar \xrightarrow{\text{replace}} X$
(Multi-Scale, Free+Interactive terms, RG / Phase Transition)

Axiom 2: Left Permutation Symmetry $V(X^* X) : \mathbb{C}^{N \times N} \rightarrow \mathbb{C}$
(Word Exchangeability, *broken by RoPE)

Axiom 3: Right Learnable U(1) Symmetry $V(\{X W_i X^*\}_i) : \mathbb{C}^n \rightarrow \mathbb{C}$
 $\text{Im}(\sigma)$ $\text{P}(\sigma(WW^*))$ (Hamiltonian Matrix/RMT)



$$\text{P}(\sigma(WW^*))$$



Q1: What About Softmax?

Why Softmax = LogSumExp Potential?

$$V(\mathbf{X}) = \frac{i}{2} \sum_{i=1}^{\text{nHead}} \sum_{j=1}^N \log \sum_{k=1}^N \exp\left(\left[\frac{1}{\sqrt{r}} \mathbf{X} \mathbf{W}_i \mathbf{X}^*\right]_{jk}\right)$$

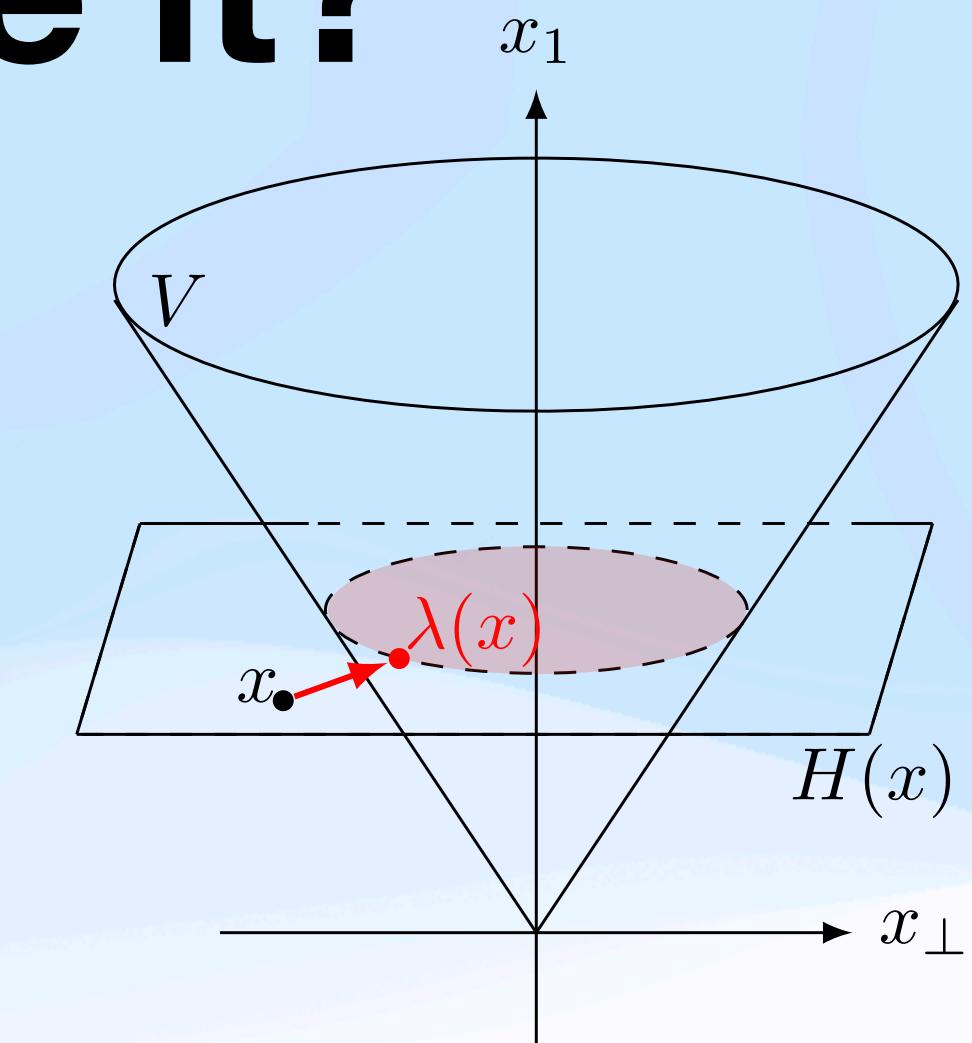
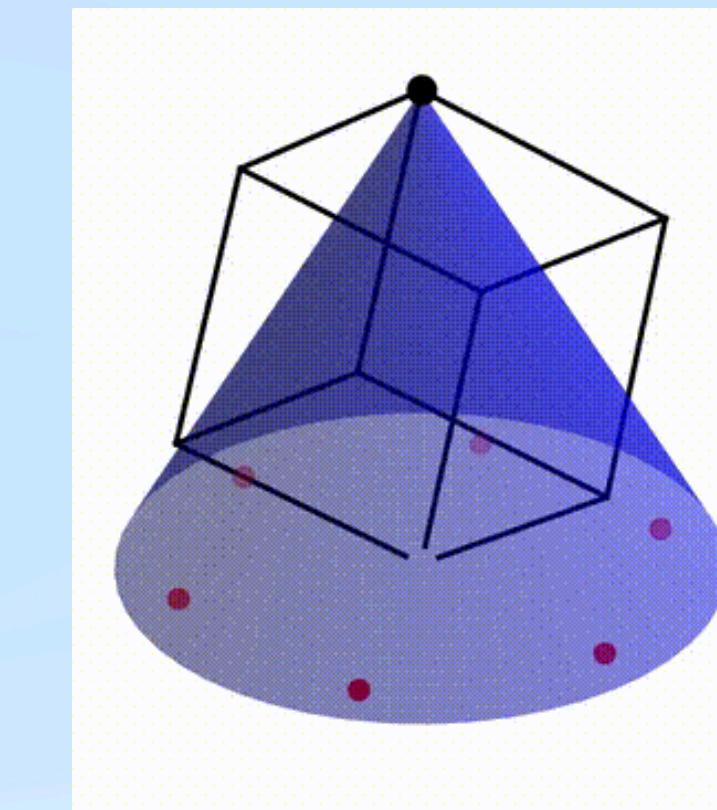
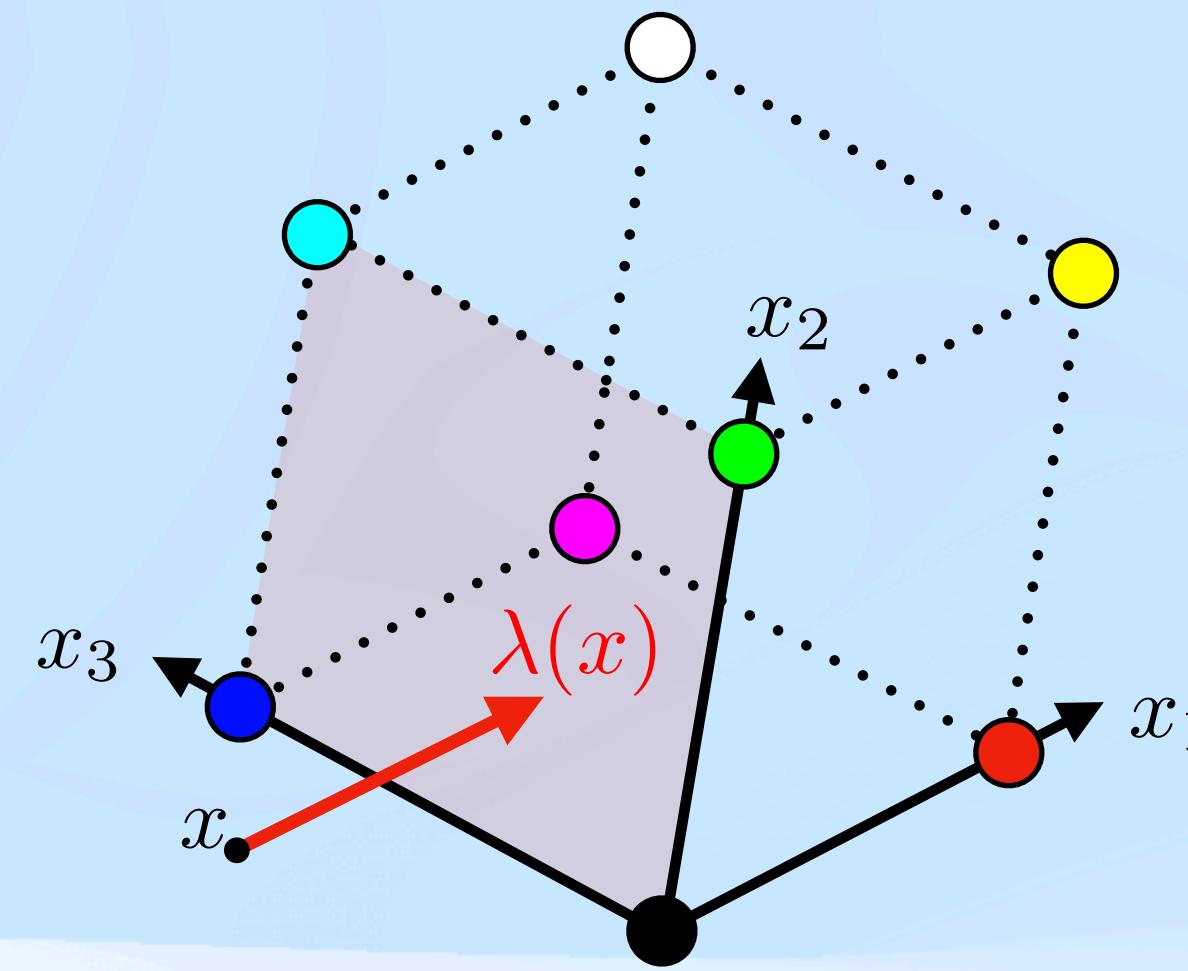
$$\frac{\partial V}{\partial \mathbf{X}^*} = \frac{i}{2} \sum_{i=1}^{\text{nHead}} \text{Softmax}\left(\frac{1}{\sqrt{r}} \mathbf{X} \mathbf{W}_i \mathbf{X}^*\right) \mathbf{X} \mathbf{W}_i$$

$$\min_{p \in \Delta^{n-1}} \mathbb{E}[E] + \tau \mathbb{E}_p[\log p] \implies p = \text{Softmax}(-E/\tau)$$

“Energy + Entropic Regularization”

Q2: What About Activation Functions? (Previous Works)

Why ReLU-MLP = Projected Gradient Descent? Or Even: How to improve it?



ReLU = Orthant Projection

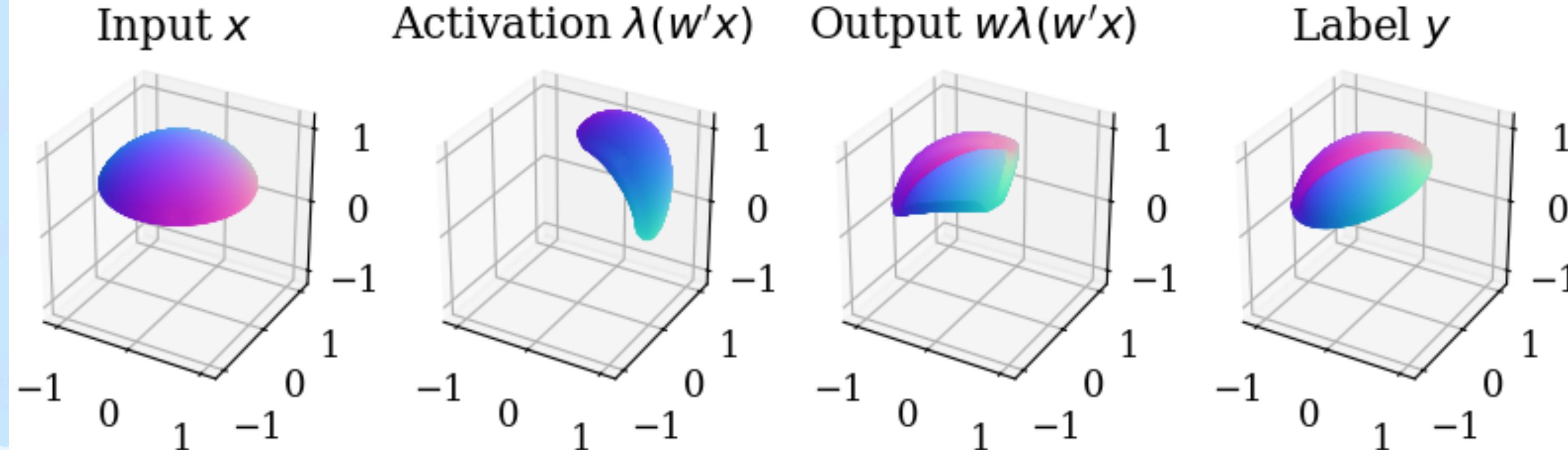
Axiom	Property	Formulation
• Component-Wise	$\forall i, \lambda\pi_i = \pi_i\lambda$	$\lambda(x_i) = \lambda(x)_i$
• Positive 1-Homogeneity	$\forall t > 0, \lambda(tx) = t\lambda(x)$	$\lambda _{\Omega} = \text{id}$
• Idempotence/Projection	$\lambda\lambda = \lambda$	$\lambda(\Omega^c) = \partial\Omega$

Permutation X
Permutation+Rotation ✓

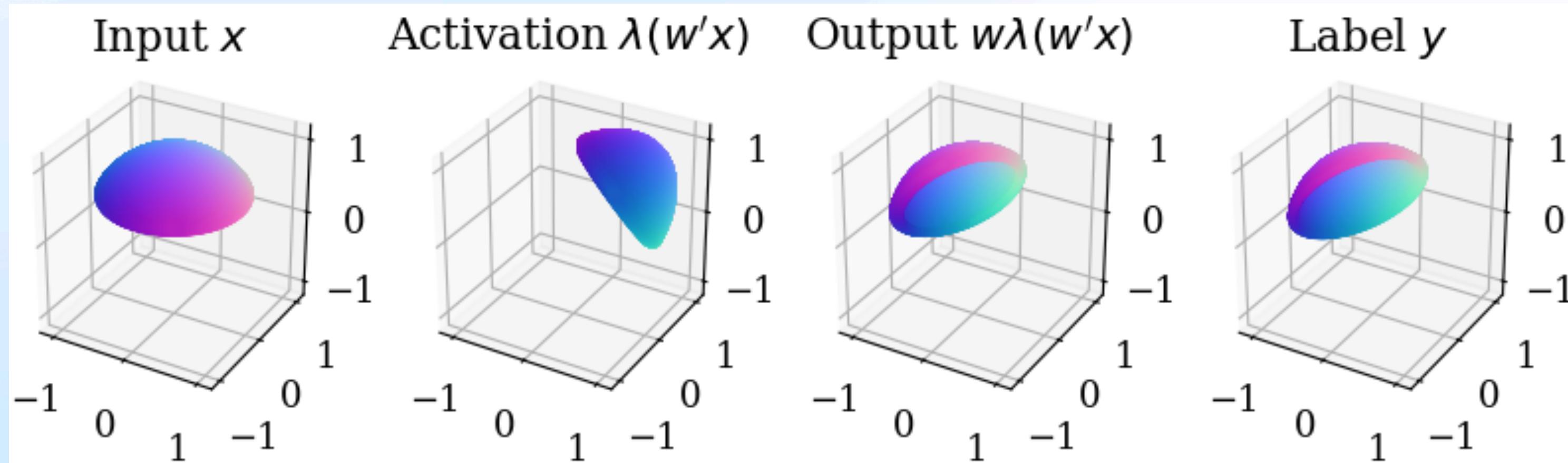
Conic Activation Functions

A symmetry constraint on generative models for improved **generalization property** and better **learning and performance**.

Embedding Space Favors Orthogonal Symmetry

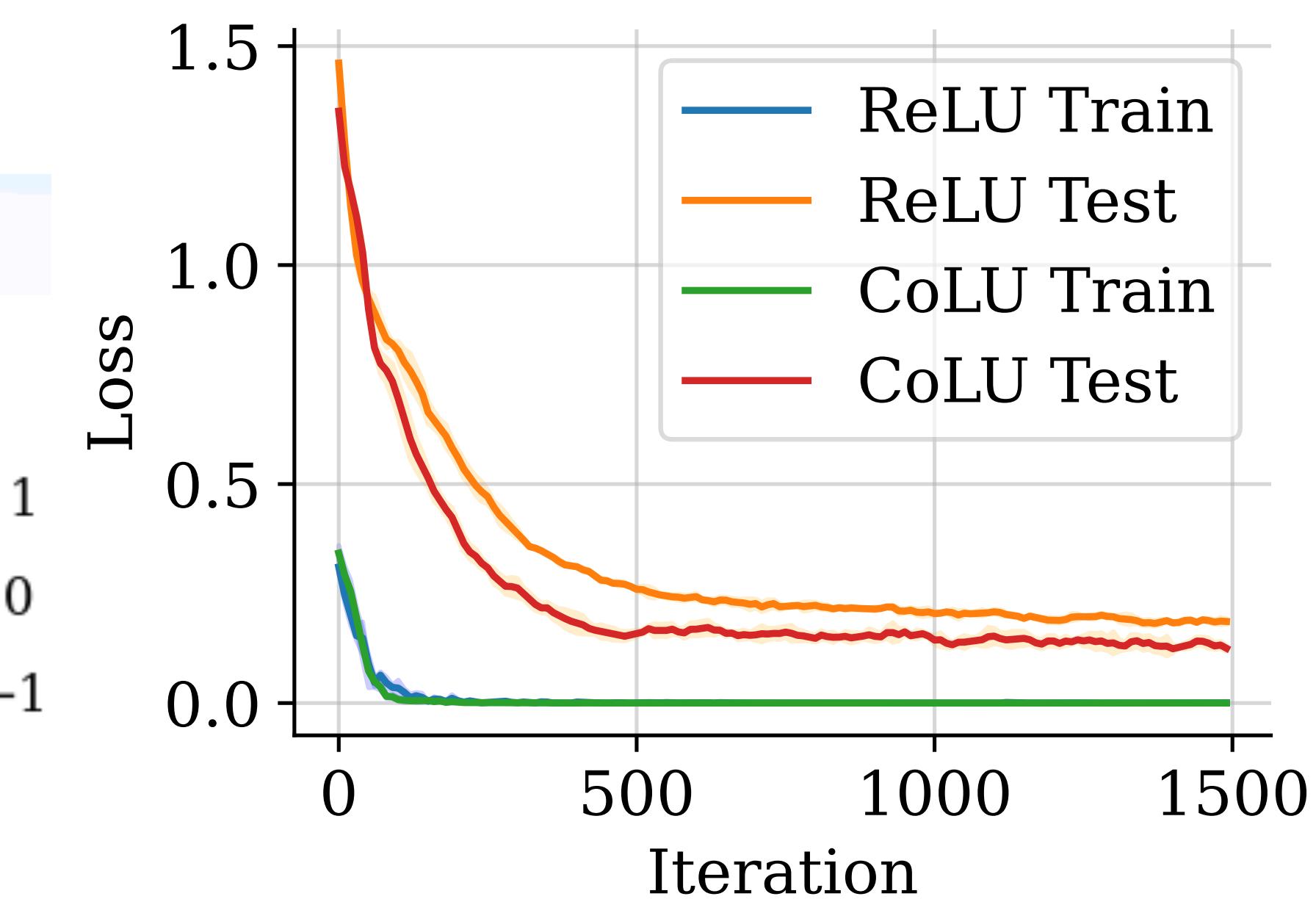


ReLU: permutation symmetry



CoLU: orthogonal/rotary symmetry

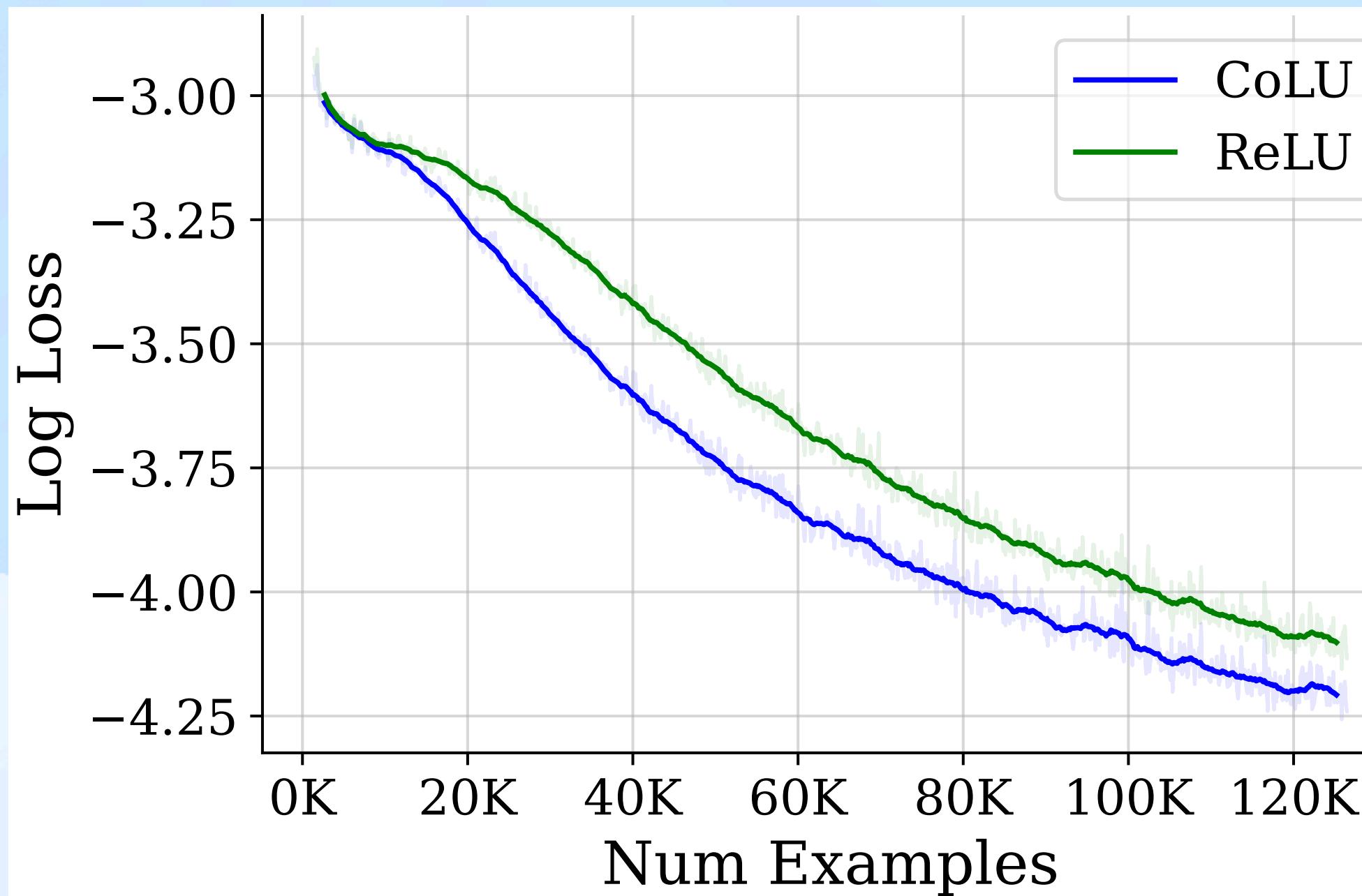
- Minimal Example (C=3)
- Improved Generalization



Conic Activation Functions

A symmetry constraint on generative models for improved **generalization property** and better **learning and performance**.

Embedding Space Favors Orthogonal Symmetry



Diffusion Transformer 0.8B
(Oxford102)

GPT2 MLP (FineWeb10M)		ReLU	CoLU
Forward FLOPs		39.064M	39.101M
Test Loss		3.4569 ± 0.1182	3.3804 ± 0.1159
ResNet-56 (CIFAR10)		ReLU	CoLU
Forward FLOPs		0.252M	0.257M
Test Accuracy		92.7282 ± 0.357	93.5851 ± 0.442
Diffusion Model (CIFAR10)		ReLU	CoLU (Faster)
Train Loss		0.1653	0.1458
Early Samples			

Conic linear units: improved model fusion and rotational-symmetric generative model. VISAPP 2024.
Conic activation functions. UniReps@NeurIPS 2024. PMLR 2025.