NEURAL INFORMATION PROCESSING SYSTEMS

IBM Research

Quantum Doubly Stochastic Transformers

Spotlight

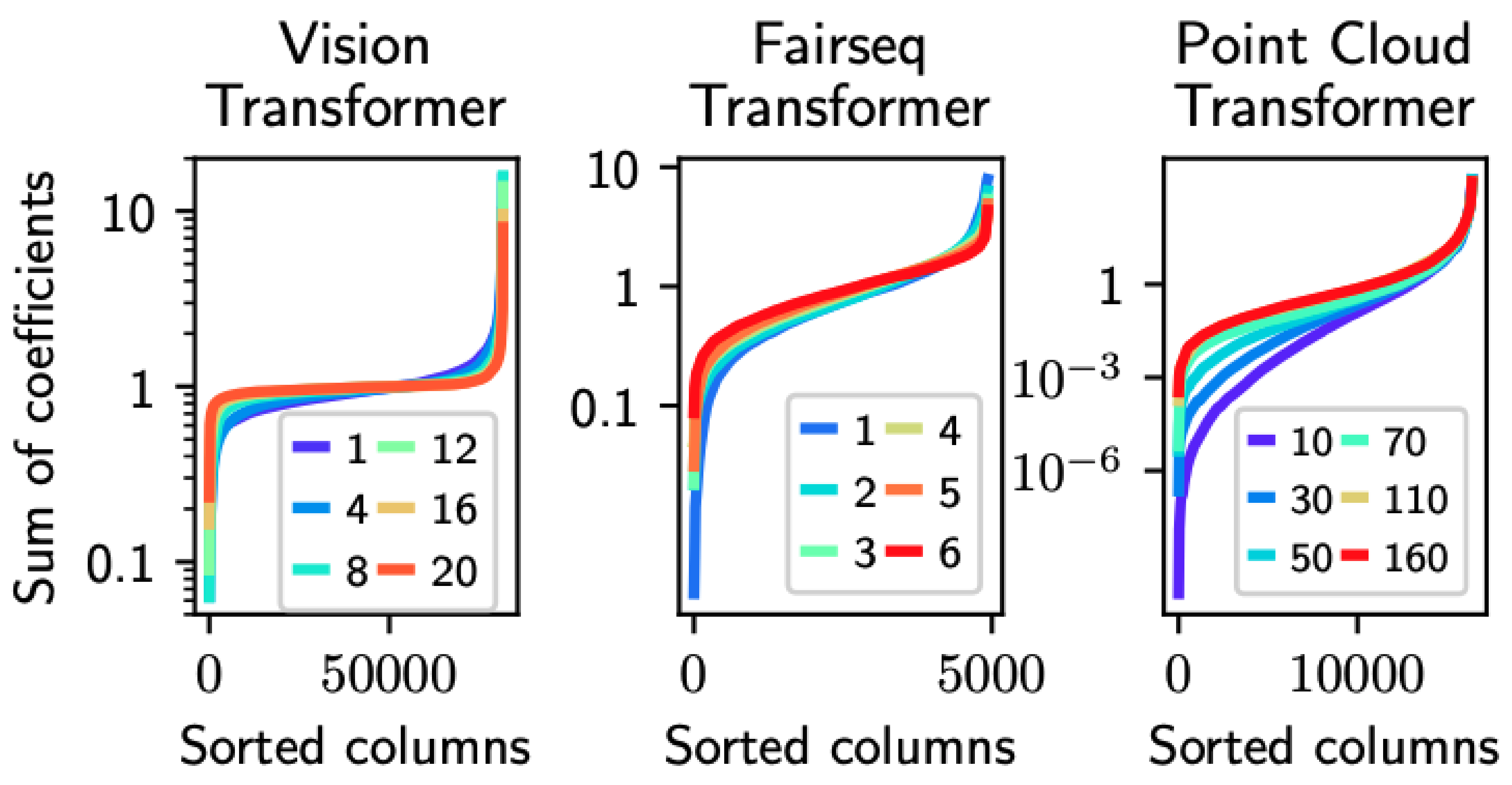
Advances in Neural Information Processing Systems (NeurIPS) 2025

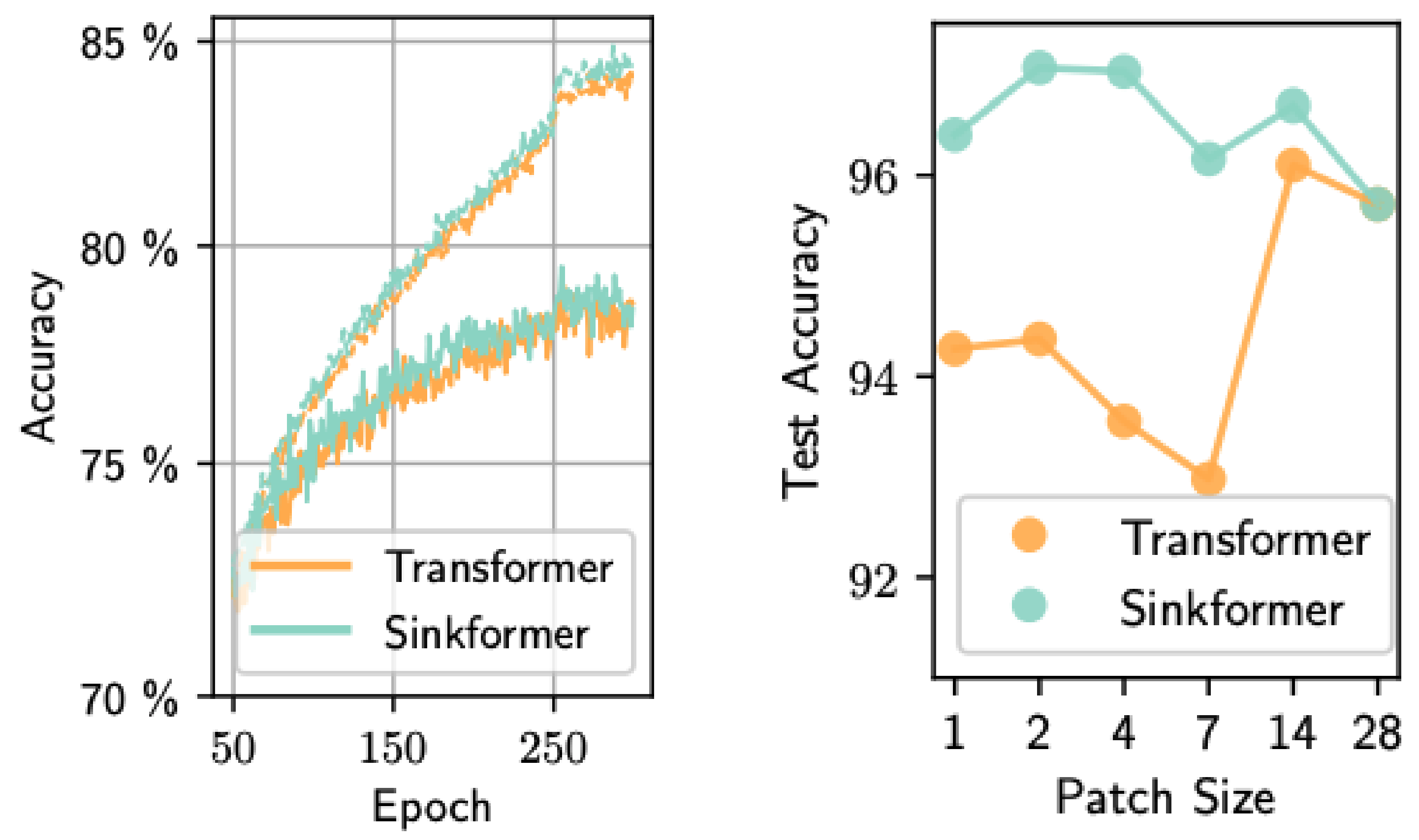
Jannis Born · Filip Skogh · Kahn Rhrissorrakrai · Filippo Utro · Nico Wagner · Aleksandros Sobczyk

IBM Research | © 2025 IBM Corporation

Right-stochastic attention can destabilize Transformer training

- Entropy collapse (ICML 2023), Rank collapse (NeurIPS 2022), Token uniformity (ICML 2021), Eureka Moments (ICML 2024)
- Attention converges to doubly stochastic matrices as training progresses
- Sinkformer – Doubly stochastic attention via Sinkhorn’s Algorithm
- Sinkformer consistently improves performance across domains





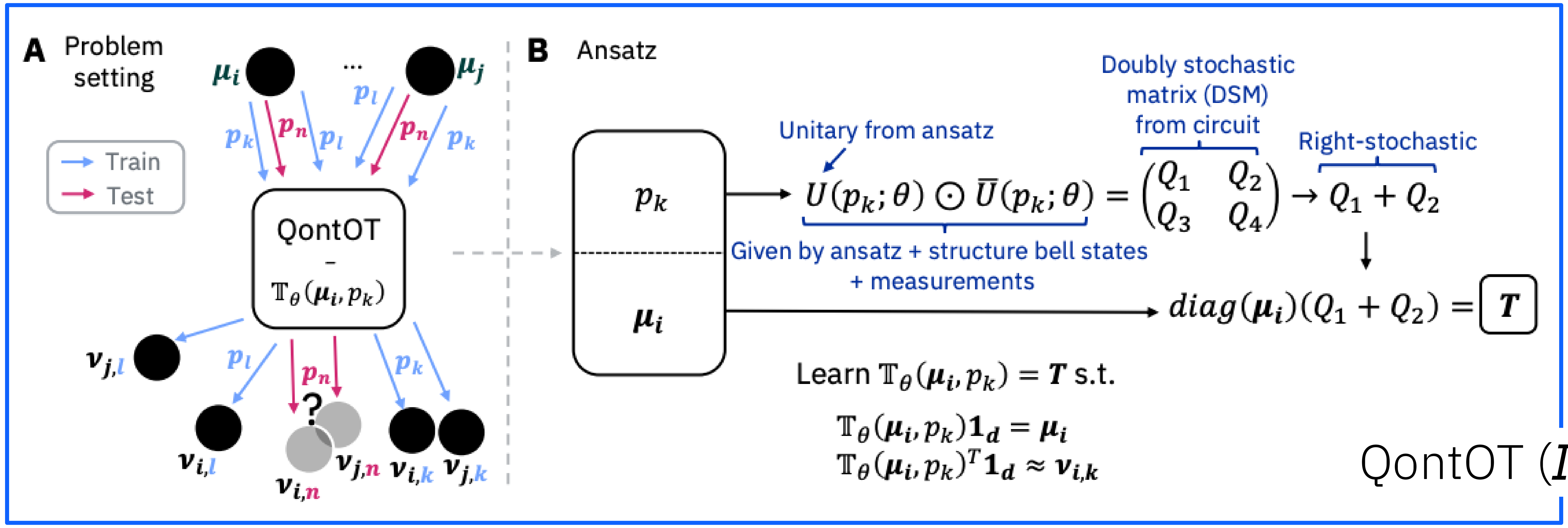
Sinkformer (2022, *AISTATS*)

IBM Research | © 2025 IBM Corporation

Linking attention with quantum computing

- Sinkhorn's algorithm is approximative, non-parametric, iterative and not gradient-friendly

- Doubly stochastic matrices (DSM) can be obtained with quantum circuits (. $\mathrm{U} \odot \bar{\mathrm{U}}$ is a DSM)



QontOT (*ICML* 2024)

- QontOT - Amortized optimization of Optimal Transport maps through a novel variational circuit

- Requires only $\approx 4\log_2(\mathrm{n})$ qubits and lacks a classical analogue

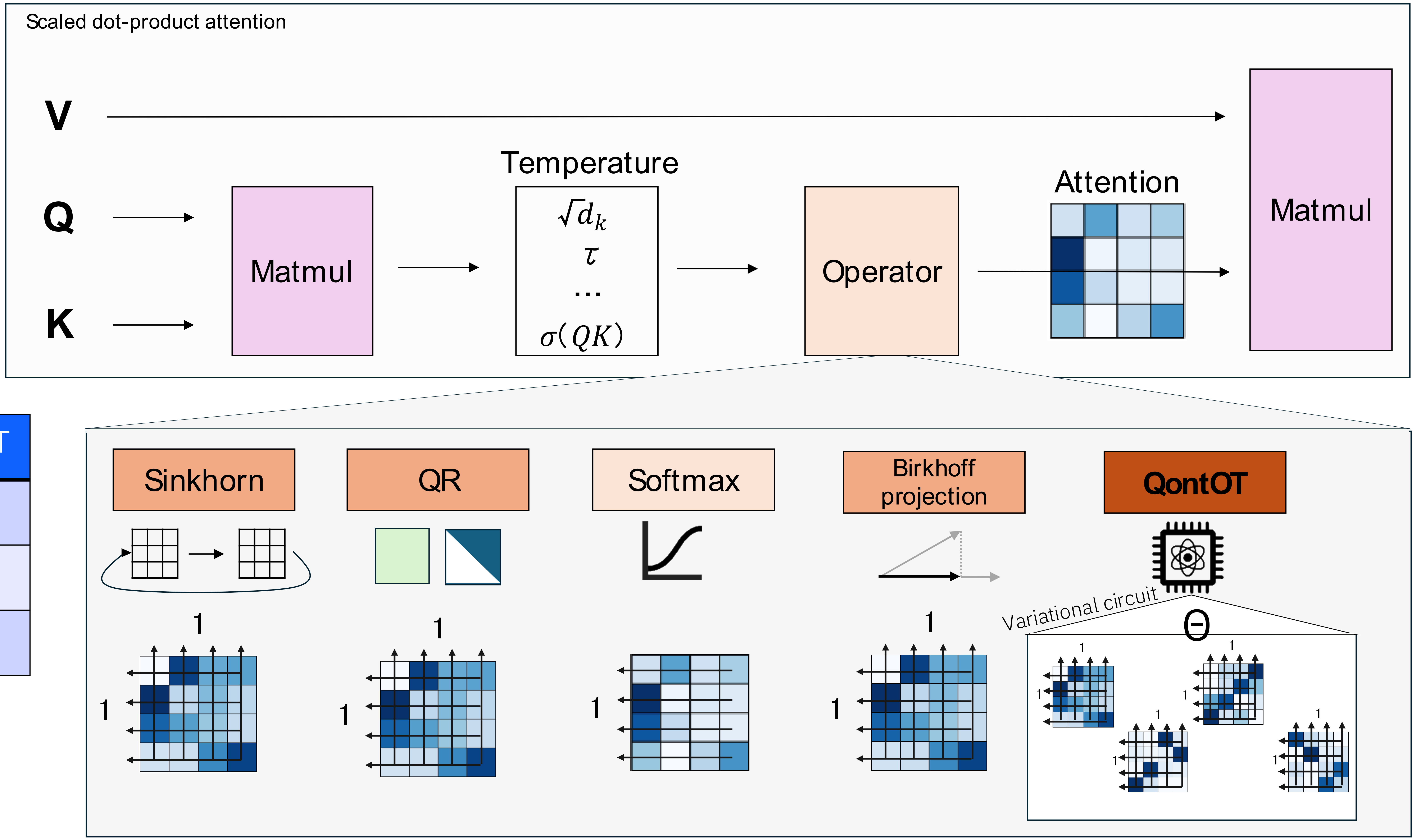
IBM Research | © 2025 IBM Corporation

Quantum Doubly Stochastic Transformer (QDSFormer)

- Replace softmax in Transformer attention with QontOT
- Compare QDSFormer alternative operators

	Sinkhorn	Project	QR	QontOT
Exact	☹	✓	✓	✓
Different.	✓	☹	✓	✓
Parametric	☹	☹	☹	✓

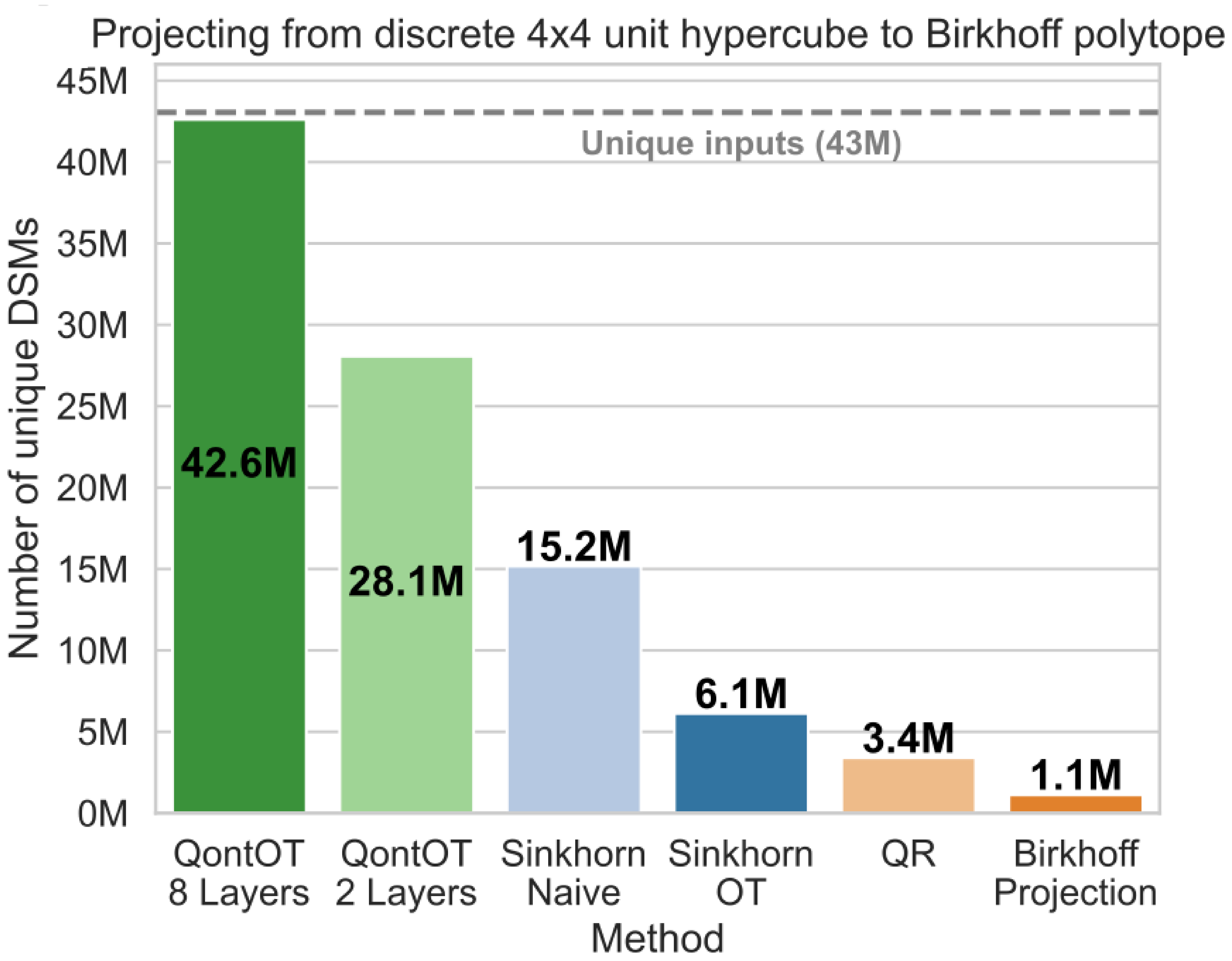
- Focus on small-scale Vision Transformers (ViT)



IBM Research | © 2025 IBM Corporation

Expressivity of different doubly stochastic activation functions

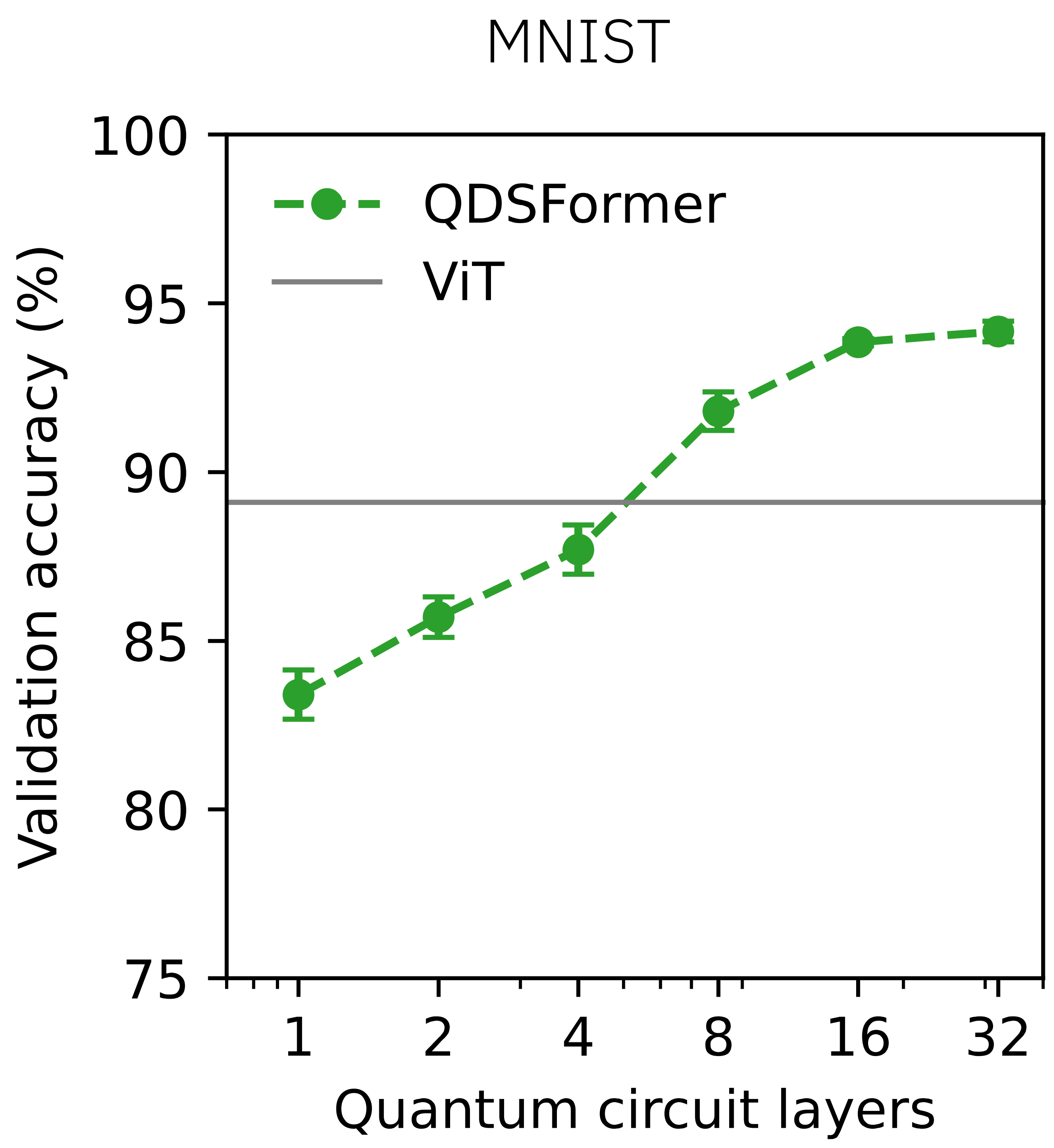
- Brute-force analysis over discretized grid of 4x4 unit hypercube
- Ideal activation function is injective
- QontOT produces more unique DSMs (even with fixed, random parametrization!)

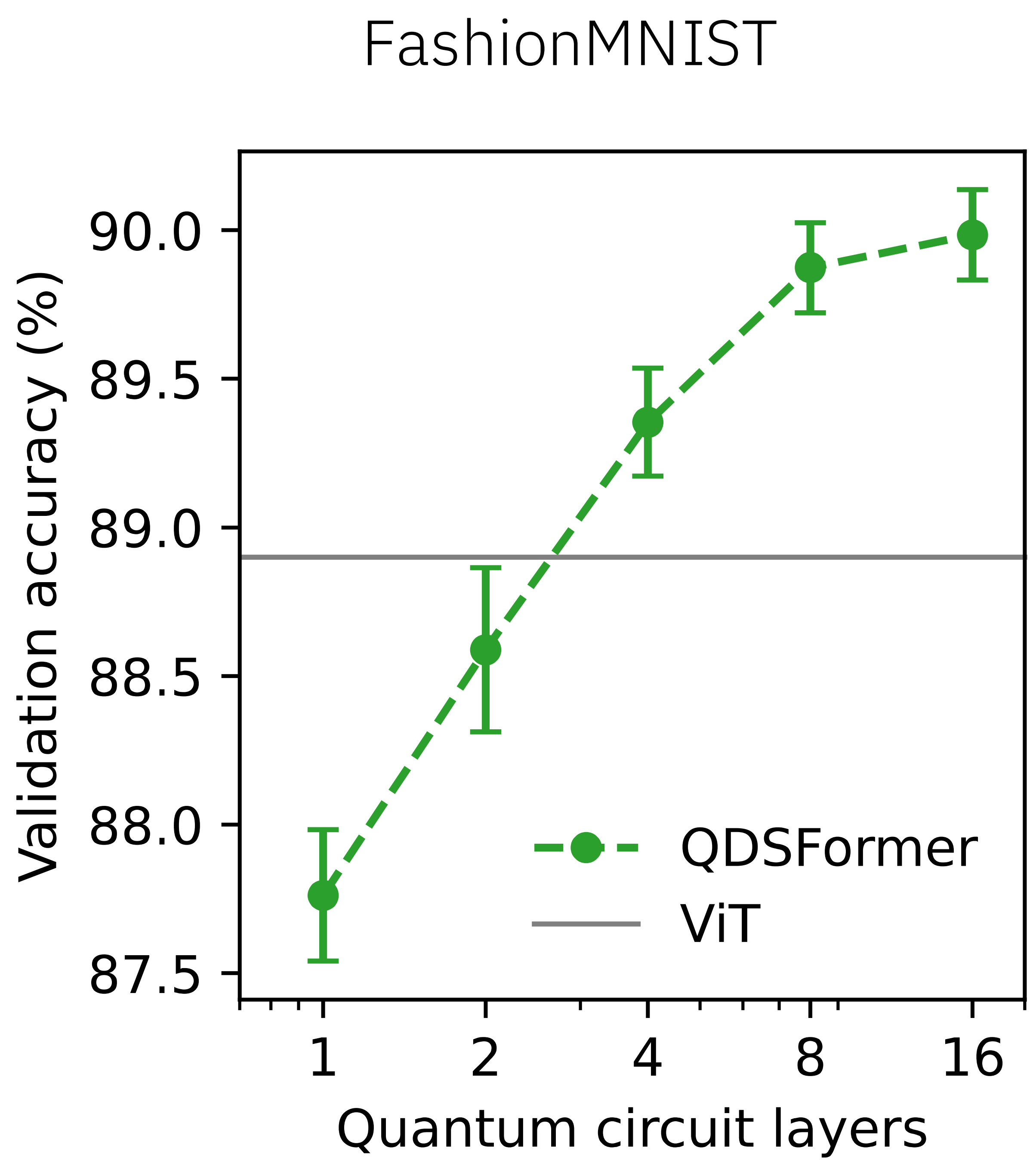


IBM Research | © 2025 IBM Corporation

Quantum Doubly Stochastic Vision Transformer

In simulation, even with few circuit layers a Vision Transformer can be outperformed





- Circuit is used statically (no online training)
- Circuit limits DSM sizes to powers of 2

Quantum Doubly Stochastic Vision Transformer

MedMNIST benchmark

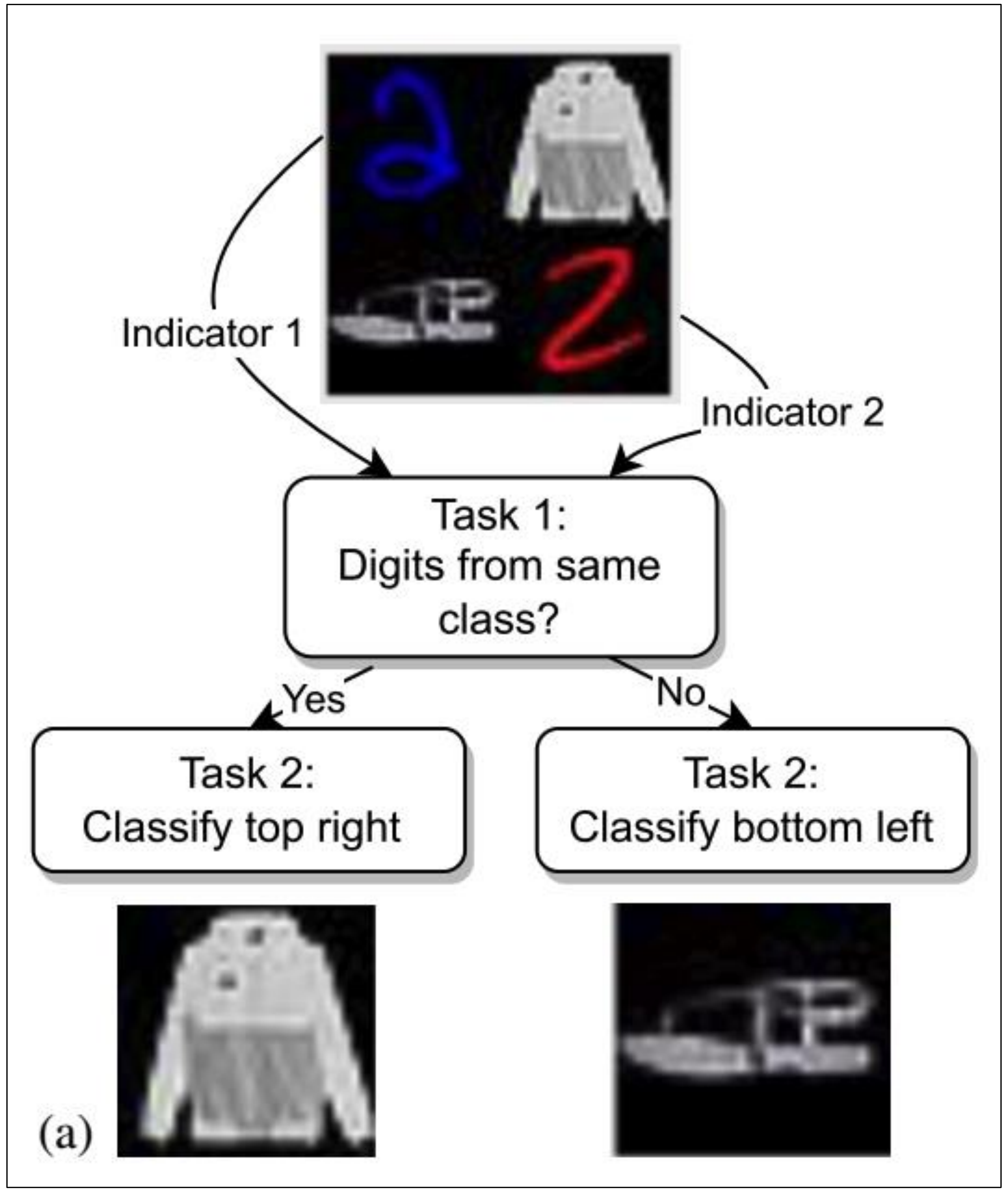
- QDSFormer best model for 5/7 datasets
- No other doubly stochastic transformer improves over standard ViT
- Up to 240k samples, up to 200k model parameters
- Small images (28x28)

Table 3. Test accuracy for different MedMNIST datasets across five attention types in a 2-layer ViT. QontOT uses 16 circuit layers.

MedMNIST dataset	Softmax	Softmax$_{\sigma^2}$	QR	QontOT	Sinkhorn
OCT	$\mathbf{64.4}_{\pm 1.6}$	$43.6_{\pm 3.0}$	$62.5_{\pm 0.9}$	$61.6_{\pm 0.6}$	$55.1_{\pm 5.2}$
Pneumonia	$84.2_{\pm 0.8}$	$84.7_{\pm 2.0}$	$84.3_{\pm 0.7}$	$\mathbf{86.1}_{\pm 1.0}$	$83.0_{\pm 1.5}$
Tissue	$60.0_{\pm 0.2}$	$49.4_{\pm 1.2}$	$59.0_{\pm 0.1}$	$\mathbf{60.6}_{\pm 0.1}$	$56.9_{\pm 2.0}$
OrganA	$78.8_{\pm 0.5}$	$73.6_{\pm 1.7}$	$78.4_{\pm 0.6}$	$\mathbf{81.2}_{\pm 0.3}$	$77.0_{\pm 2.5}$
OrganC	$79.8_{\pm 0.5}$	$71.7_{\pm 7.3}$	$79.6_{\pm 0.3}$	$\mathbf{82.7}_{\pm 0.5}$	$79.7_{\pm 1.0}$
OrganS	$64.4_{\pm 0.6}$	$59.3_{\pm 0.9}$	$62.6_{\pm 0.8}$	$\mathbf{68.1}_{\pm 0.6}$	$63.5_{\pm 0.9}$
Breast	$79.6_{\pm 2.0}$	$78.2_{\pm 2.2}$	$\mathbf{81.3}_{\pm 2.9}$	$80.0_{\pm 1.1}$	$80.1_{\pm 0.8}$
Mean	73.0	65.8	72.5	**74.3**	70.8

© 2025 IBM Corporation

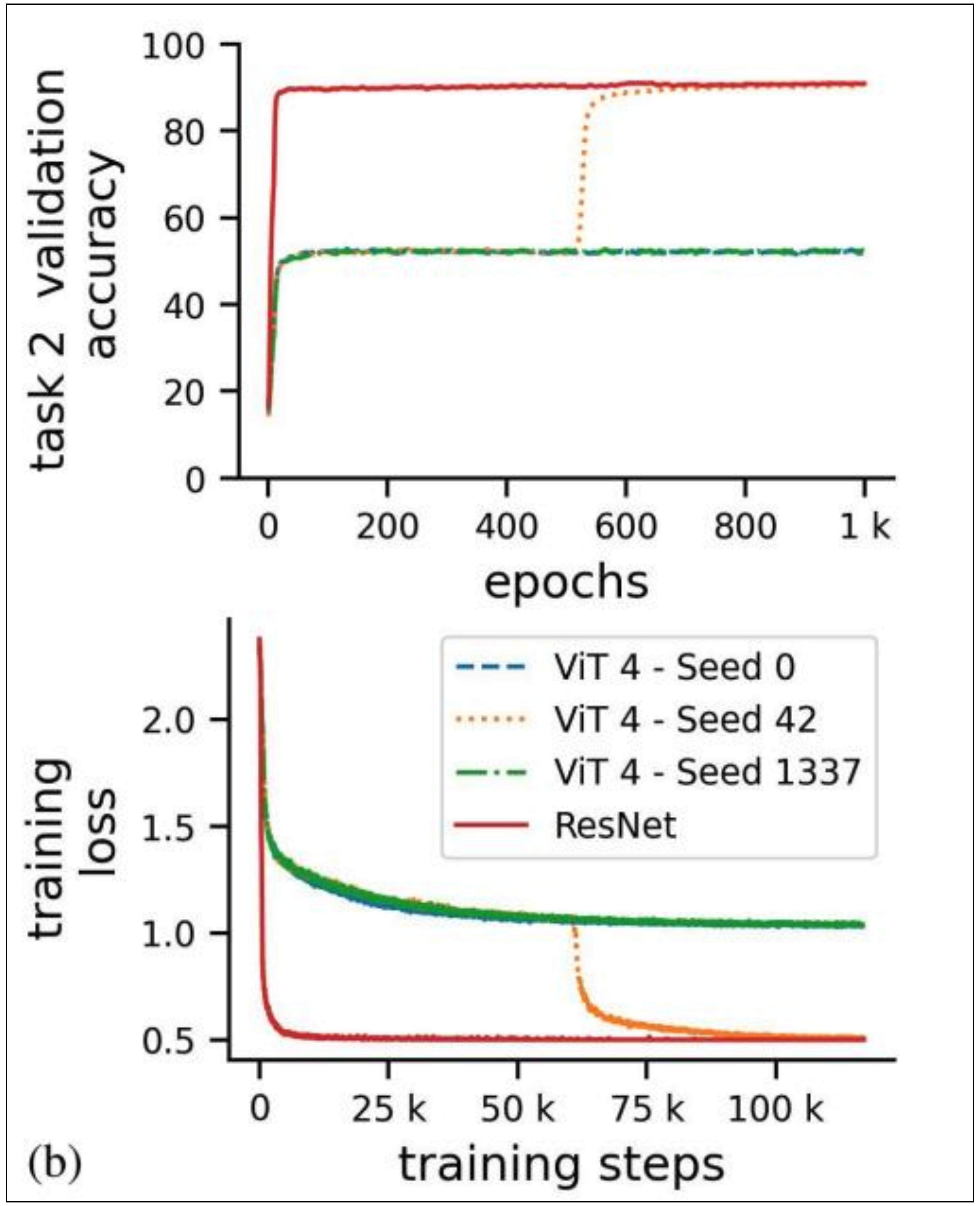
QDSFormer on a compositional vision task



Hofmann et al. (2024, *ICML*)

- Classical ViTs are unstable to train – especially in compositional tasks

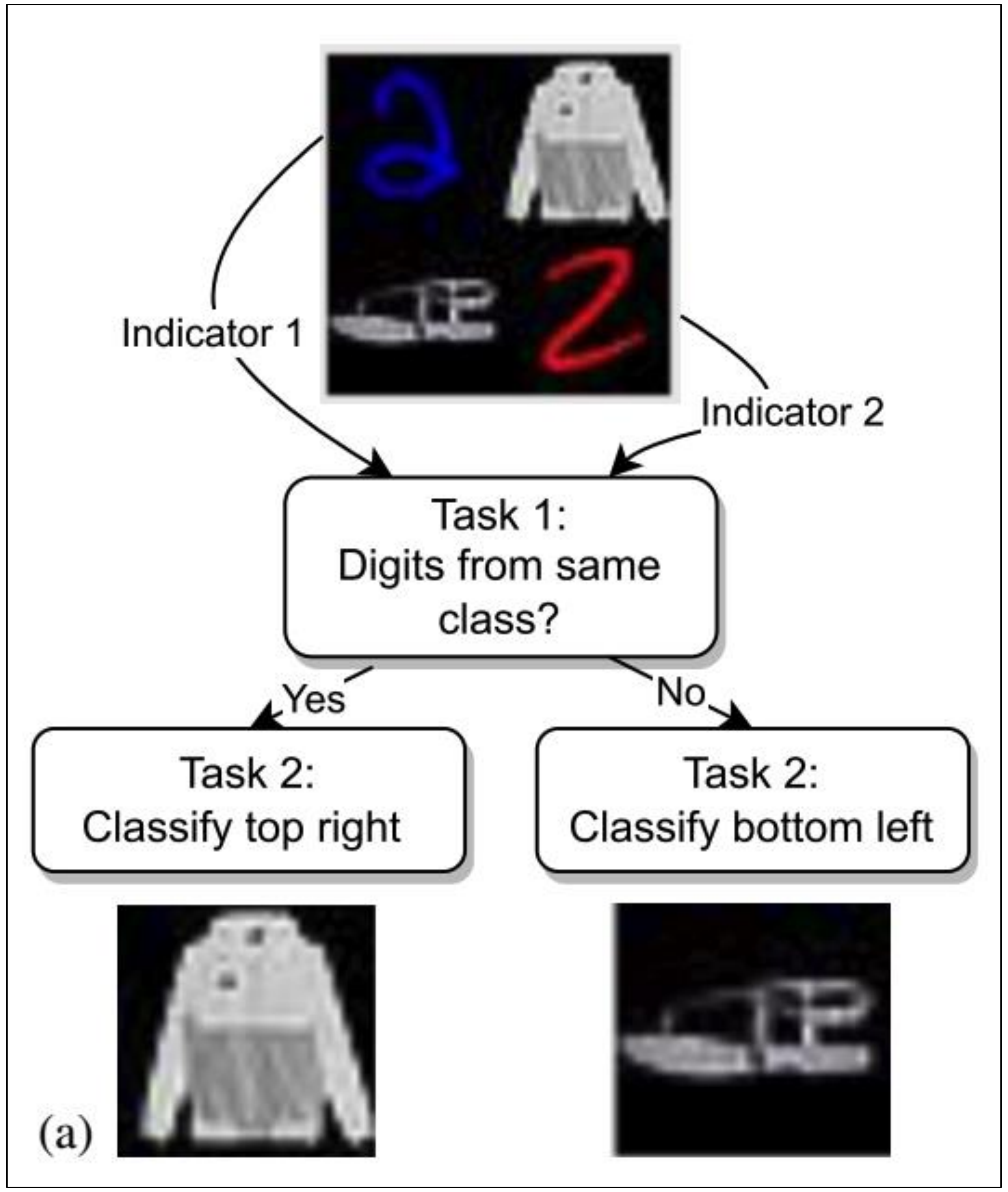
- More stable training → Earlier Eureka Moment



Hofmann et al. (2024, *ICML*)

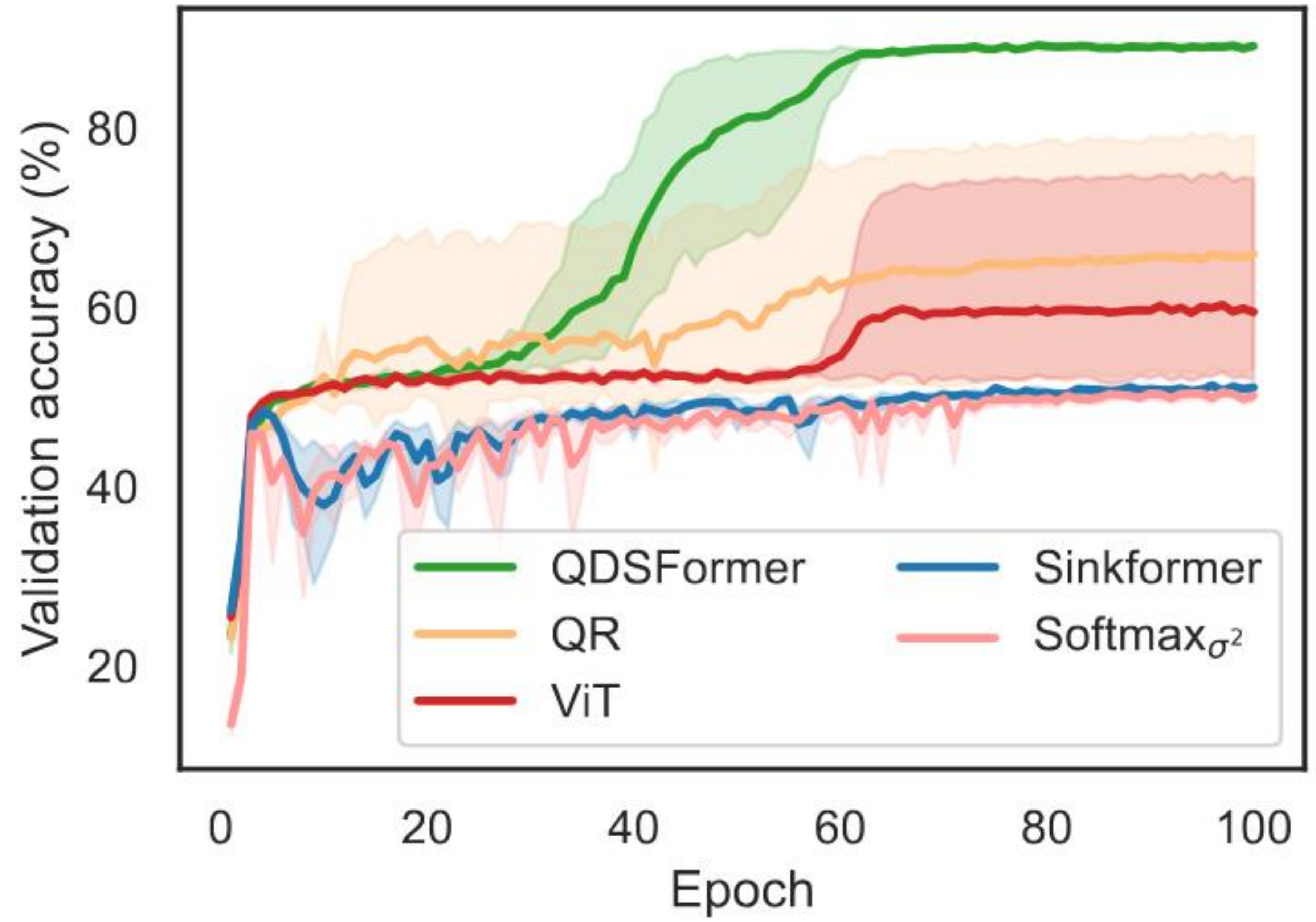
IBM Research | © 2025 IBM Corporation

QDSFormer on a compositional vision task



Hofmann et al. (2024, ICML)

- Classical ViTs are unstable to train – especially in compositional tasks
- More stable training → Earlier Eureka Moment
- QDSFormer stabilizes training and converges faster



Model	Accuracy
Softmax	$61.1_{\pm 14.6}$
Softmax$_{\sigma^2}$	$51.0_{\pm 0.52}$
Sinkhorn	$51.6_{\pm 0.24}$
QR	$66.4_{\pm 15.6}$
QDSFormer	$\mathbf{89.4}_{\pm 0.09}$

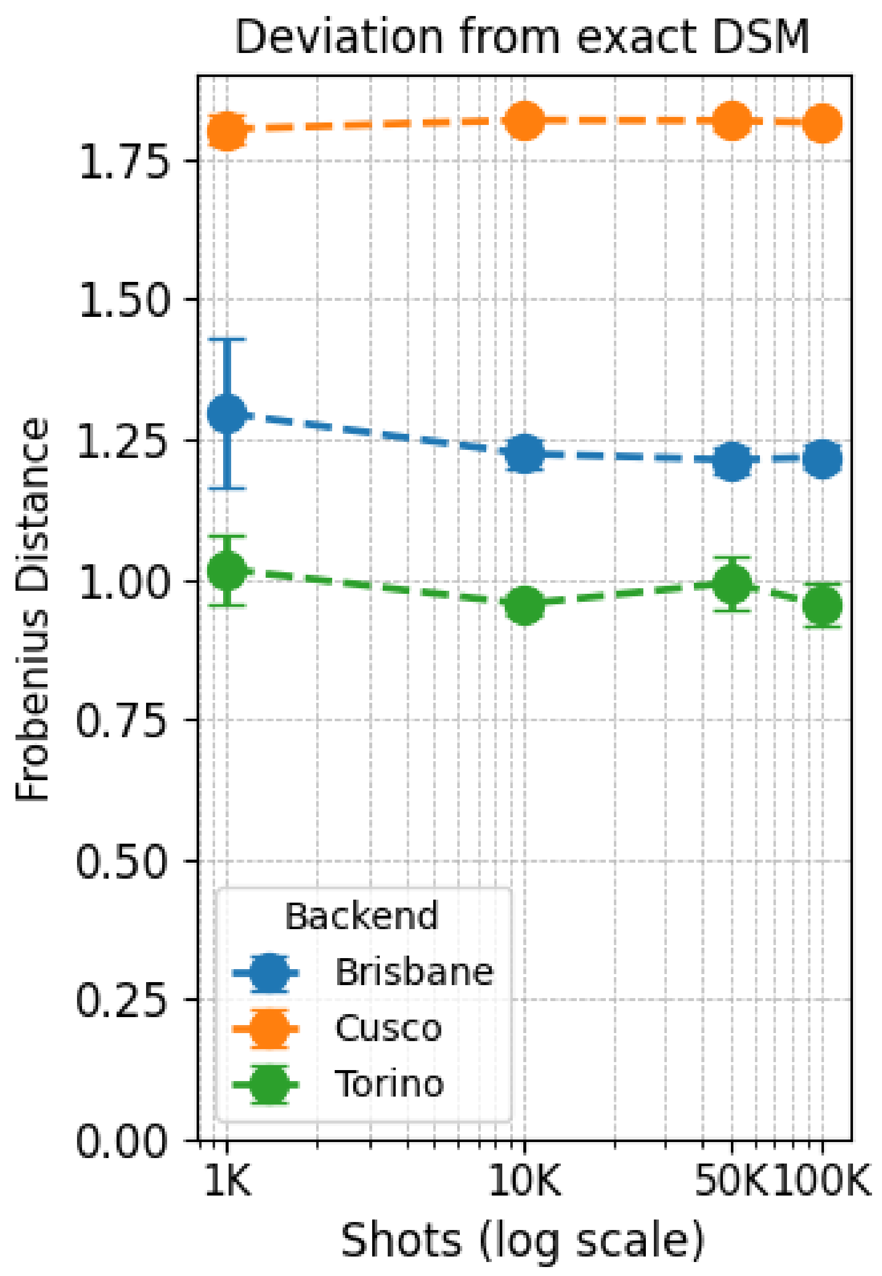
Quantum hardware experiments with 1-layer QDSFormer

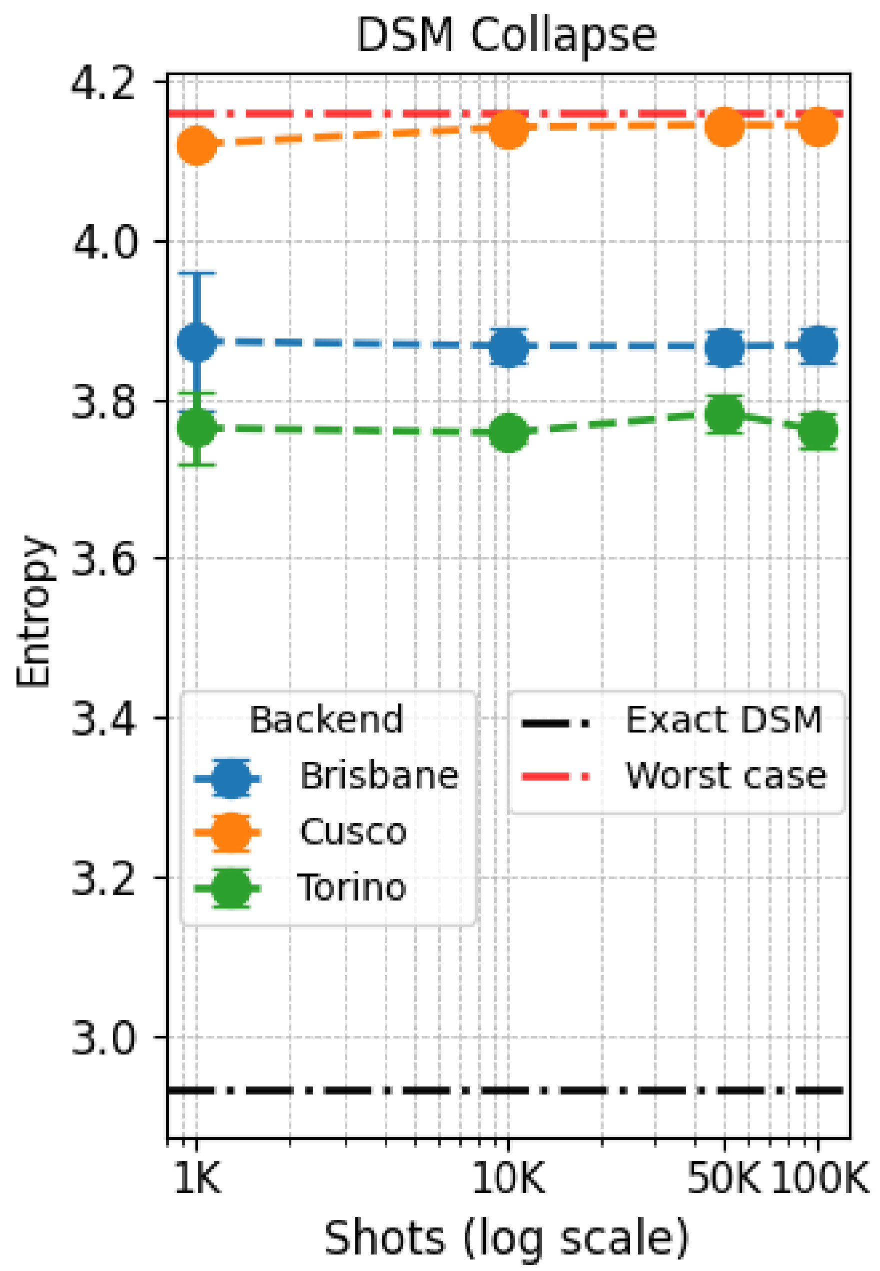
Compare exact DSM to
- Per device: 10 runs with 1k, 10k, 50k & 100k shots
- Calculate the Frobenius Distance between HW DSM and exact DSM

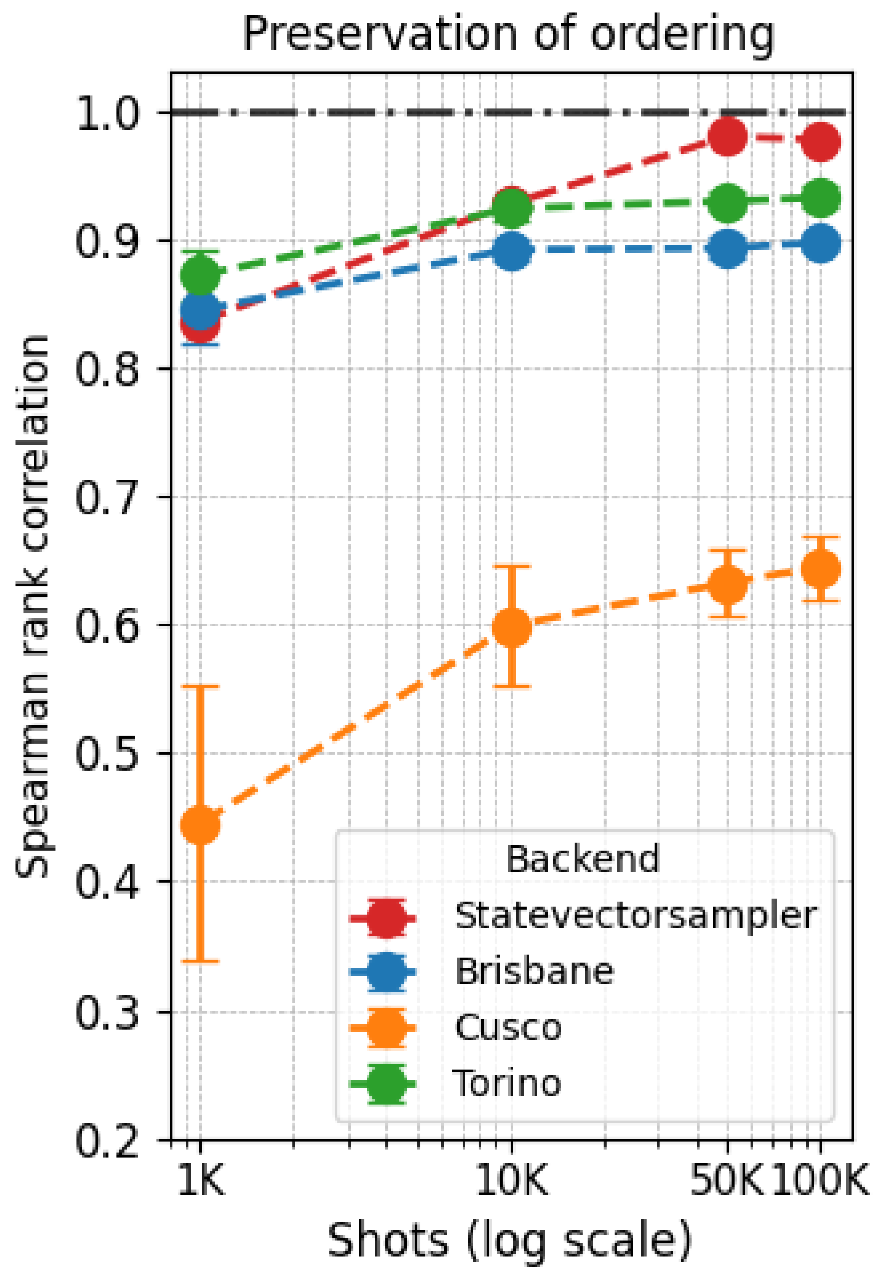
Error mitigation:
- Dynamical decoupling
- Pauli twirling
- Birkhoff projection

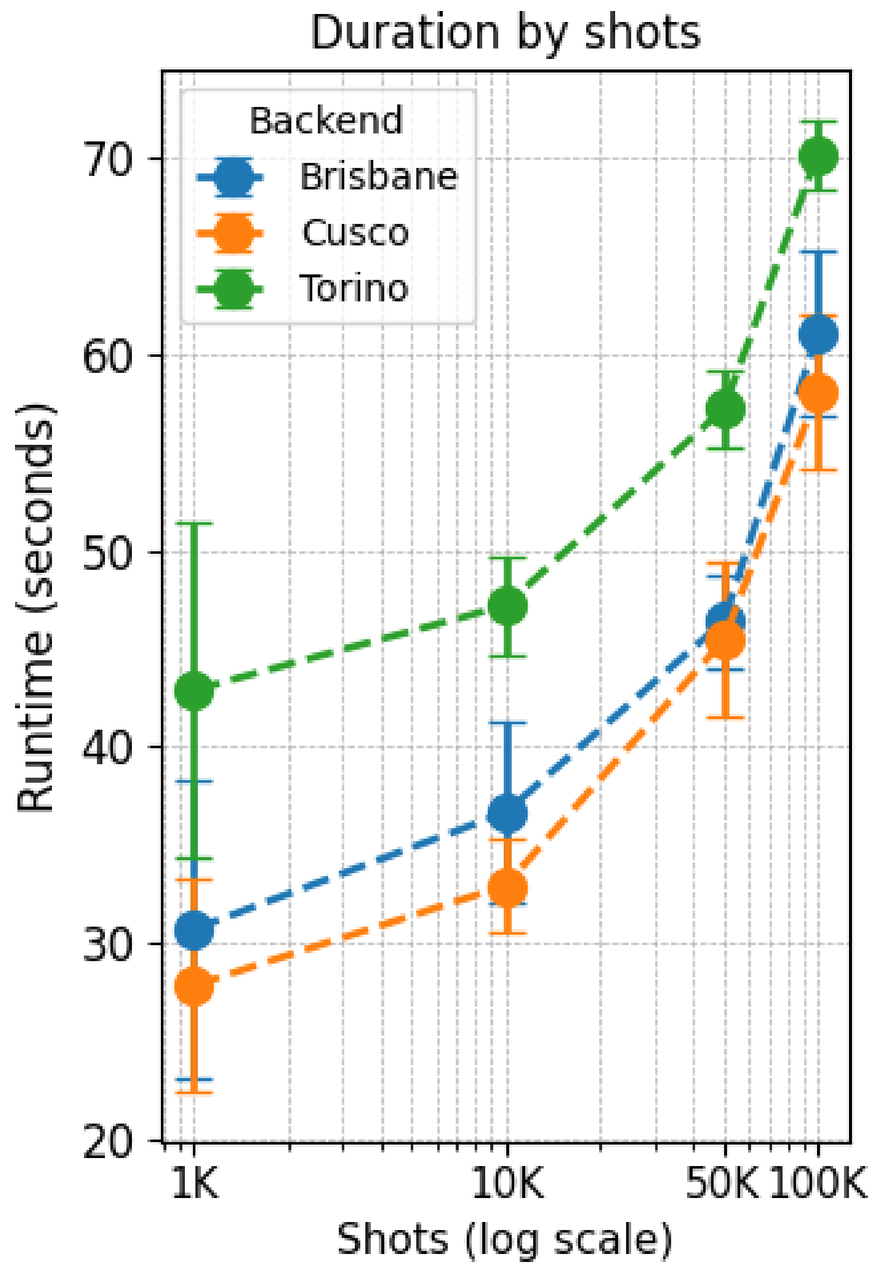
Transpilation opt. level 1*

	1 Layer	8 Layer
Qubits	14	14
Ansatz parameter	69	405
Transp. 2q-depth	15	38
Transp. 2q-op_count	52	166









Hardware specs:

Torino: Heron R1 (133 qubits, EPLG: 1.3%)

Brisbane: Eagle R3 (127 qubits, EPLG: 2.2%)

Cusco: Eagle R3 (127 qubits, EPLG: 6.8%)

IBM Research | © 2025 IBM Corporation

NEURAL INFORMATION PROCESSING SYSTEMS

IBM Research

Thank you for your attention!

Visit our spotlight poster:

@SanDiego: Wed 3.12. 4:30 p.m. CST – 7:30 p.m. (Hall C/D/E). By Filip & Kahn

@Copenhagen: EurIPS Wed 3.12 to Fri 5.12. By Jannis

IBM Research | © 2025 IBM Corporation

1. NTL consistently improves performance on mathematical tasks

Table 1: Validation accuracy of L-layered ViT on FashionMNIST and MNIST for different attention methods. QontOT uses 16 circuit layers. Mean/std computed from 5 trainings.

	FashionMNIST					MNIST				
L	**Softmax**	**Softmax$_{\sigma^2}$**	**QR**	**QontOT**	**Sinkhorn**	**Softmax**	**Softmax$_{\sigma^2}$**	**QR**	**QontOT**	**Sinkhorn**
1	$\underline{86.5}_{\pm 0.2}$	$75.3_{\pm 4.6}$	$\mathbf{87.1}_{\pm 0.3}$	$85.6_{\pm 0.1}$	$84.2_{\pm 3.6}$	$89.1_{\pm 12.5}$	$66.7_{\pm 22.5}$	$\mathbf{96.6}_{\pm 0.1}$	$93.9_{\pm 0.1}$	$\underline{94.3}_{\pm 2.0}$
2	$88.9_{\pm 0.1}$	$84.6_{\pm 2.1}$	$\underline{89.3}_{\pm 0.1}$	$\mathbf{90.0}_{\pm 0.2}$	$89.1_{\pm 0.7}$	$98.1_{\pm 0.3}$	$93.0_{\pm 4.6}$	$\underline{98.3}_{\pm 0.1}$	$\mathbf{98.4}_{\pm 0.1}$	$98.2_{\pm 0.3}$
3	$\underline{89.4}_{\pm 0.3}$	$86.3_{\pm 2.7}$	$\underline{89.4}_{\pm 0.1}$	$\mathbf{90.3}_{\pm 0.1}$	$\underline{89.4}_{\pm 0.8}$	$\underline{98.6}_{\pm 0.1}$	$97.7_{\pm 0.7}$	$\underline{98.6}_{\pm 0.1}$	$\mathbf{98.7}_{\pm 0.1}$	$\underline{98.6}_{\pm 0.1}$
4	$\underline{89.7}_{\pm 0.3}$	$87.1_{\pm 1.2}$	$89.5_{\pm 0.1}$	$\mathbf{90.3}_{\pm 0.1}$	$89.1_{\pm 1.1}$	$\mathbf{98.8}_{\pm 0.1}$	$97.9_{\pm 0.7}$	$\underline{98.7}_{\pm 0.1}$	$\mathbf{98.8}_{\pm 0.1}$	$97.9_{\pm 1.6}$

IBM Research | © 2025 IBM Corporation

From NTL to general world knowledge

Training with cross entropy
=
all mistakes are equally bad

- Frequently violated:
  - Image recognition
  - molecule or protein generation
  - numbers or units in text
  - Spherical coordinates
  - Algorithmic design (e.g., trigonometric functions)

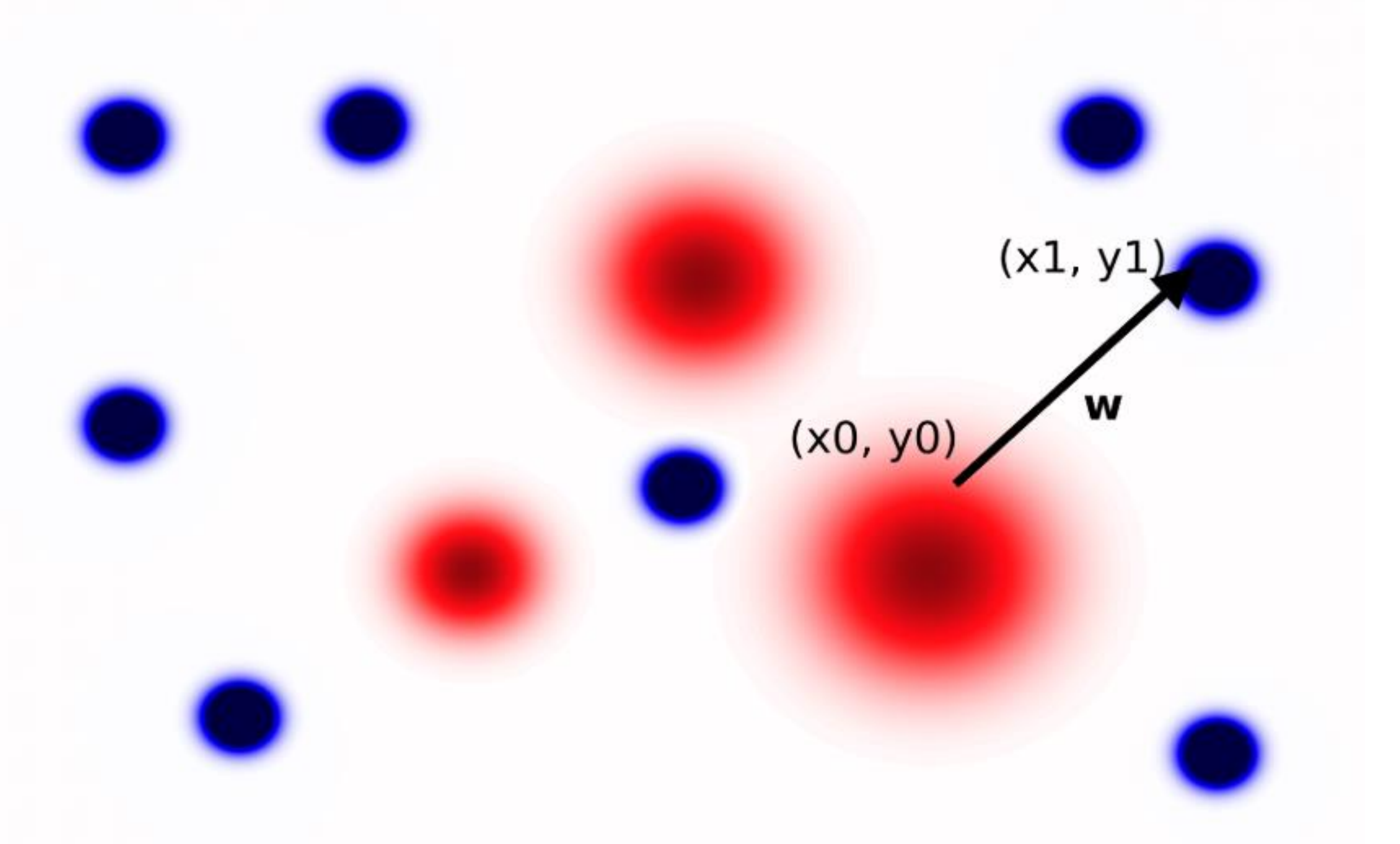
- Observation: NTL only requires a pairwise distance matrix between a subset of vocabulary tokens

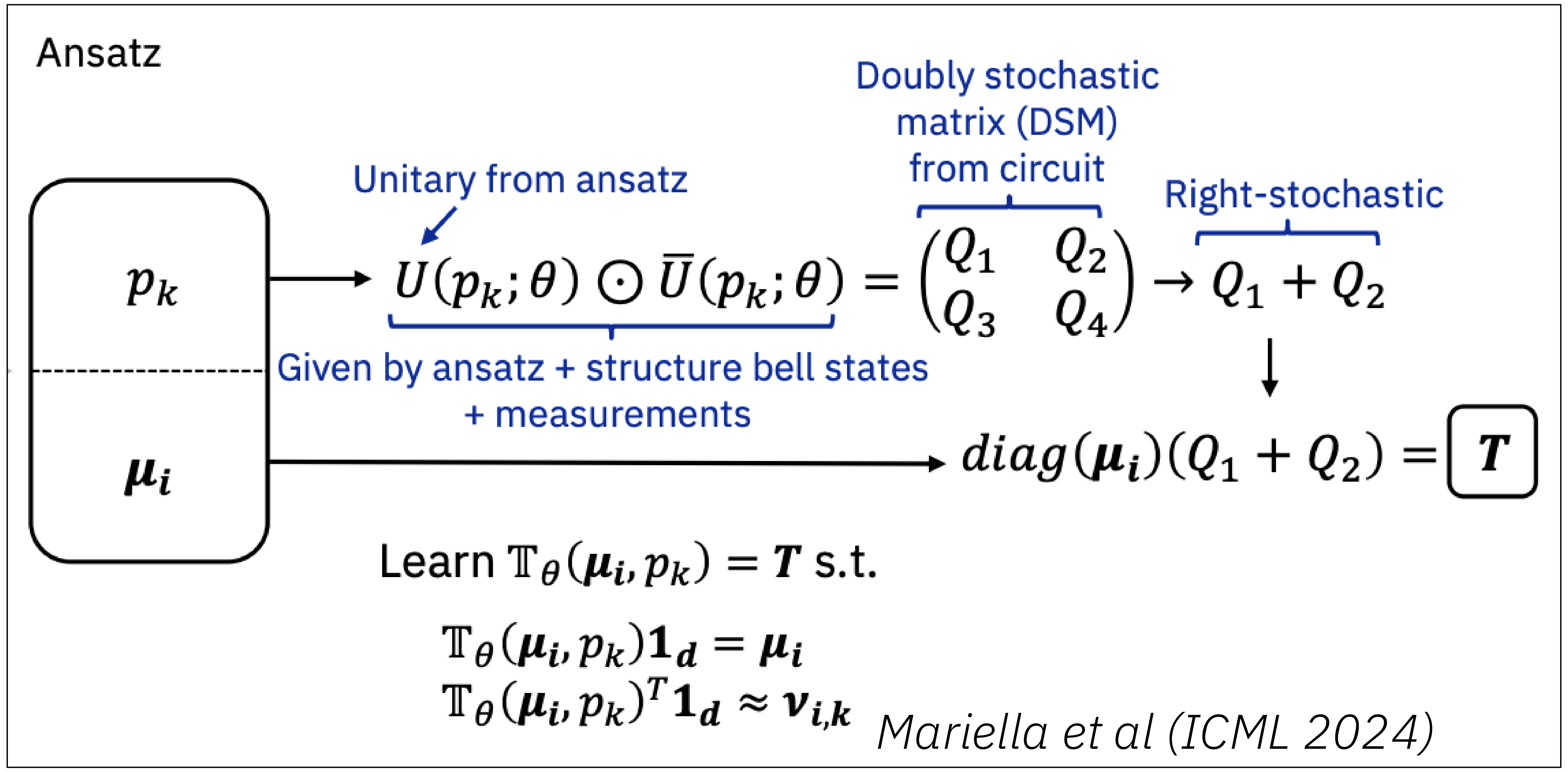
- Idea: Leverage NTL to induce domain-specific knowledge when training your LM

© 2025 IBM Corporation

What if distributions are not 1D?

- Beyond 1D, Wasserstein Distance is not differentiable
  - → Entropic regularization
- Sinkhorn's iterative matrix balancing algorithm
  - Slow and poor gradient flow
- Result: Wasserstein Distance rarely used in ML





Mariella et al (ICML 2024)

<u>Novel strategies:</u>

- Quantum computing
- Tensor Networks

© 2025 IBM Corporation