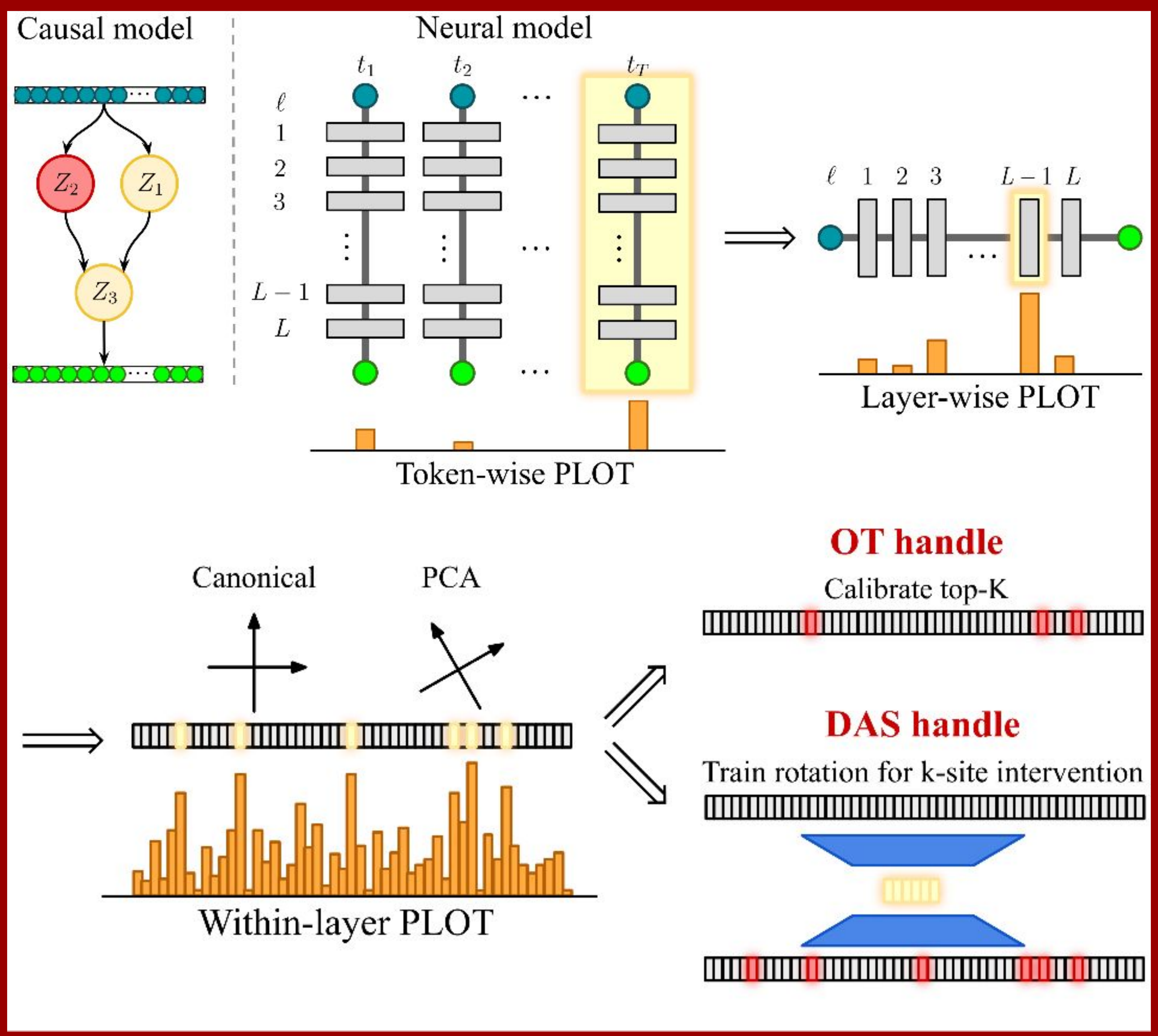


Optimal transport can quickly reveal where abstract variables live in neural networks

PLOT: Progressive Localization via Optimal Transport in Neural Causal Abstraction

Jonathn Chang, Arya Datla, Ziv Goldfeld
Cornell University

1 TL;DR

- Overview: **PLOT** localizes high-level causal variables by fitting an optimal-transport coupling between abstract intervention effects and neural-site intervention effects.
- Simple tasks: on hierarchical equality, **PLOT** obtains near-DAS accuracy while running about 30× faster. On 4-bit addition, **PLOT-guided DAS** reaches DAS-level accuracy while running about 3.3× faster.
- Large task: **PLOT-guided DAS** matches full DAS while running about 11× faster; direct **PLOT-native** handles are within roughly 2 percentage points of full DAS while running about 64× faster.

2 Background / Motivating Question

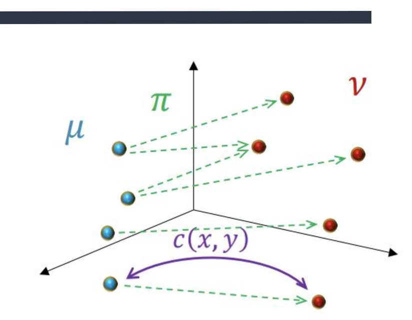
- **Causal abstraction** frames mechanistic interpretability as a counterfactual alignment problem: a neural model realizes a high-level causal model only if interventions on abstract variables can be matched by internal neural interventions (Pearl, 2009; Geiger et al., 2025).
- **DAS** (Distributed Alignment Search) uses interchange intervention training to learn rotated subspace handles, while **Boundless DAS** and **HyperDAS** automate parts of the search, but circuit localization remains a computational bottleneck (Geiger et al., 2024; Wu et al., 2024; Sun et al., 2025).
- Optimal transport (OT) turns counterfactual intervention effects into a global soft correspondence between abstract variables and neural sites, narrowing the search to promising tokens or layers before applying DAS.

3 Optimal Transport & PLOT

- Given fit pairs $(x_t^b, x_t^s) \in \mathcal{D}_{\text{ft}}$, abstract variables $Z_1, \dots, Z_m$, and candidate neural sites $s_1, \dots, s_n$:

$$\Delta_{i,t}^{\text{abs}} := \phi(y_{i,t}^{\text{abs\_swap}}) - \phi(y_t^{\text{abs}}), \qquad i = 1, \dots, m.$$

$$\Delta_{j,t}^{\text{nn}} := \phi(y_{j,t}^{\text{nn\_swap}}) - \phi(y_t^{\text{nn}}), \qquad j = 1, \dots, n.$$



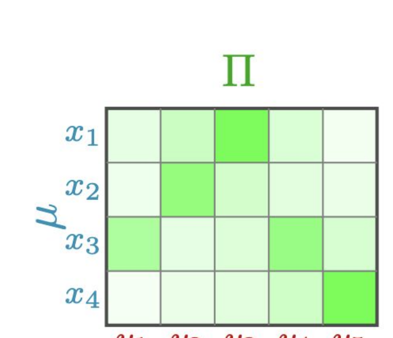
(a) A coupling $\pi \in \Pi(\mu, \nu)$ transports mass from μ to ν.

- Aggregate effects across fit pairs to obtain effect signatures per variable/site:

$$u_i := \mathrm{Agg}(\Delta_{i,1}^{\text{abs}}, \dots, \Delta_{i,T_{\text{ft}}}^{\text{abs}}), \; v_j := \mathrm{Agg}(\Delta_{j,1}^{\text{nn}}, \dots, \Delta_{j,T_{\text{ft}}}^{\text{nn}})$$

- Using these signatures, define a uniform empirical measure, and compute an entropic OT coupling $\Pi \in \mathbb{R}_{\geq 0}^{m \times n}$ via Sinkhorn's algorithm.
- For an executable intervention handle for Z_i:

$$\widetilde{\Pi}_{j|i} := \frac{[\Pi]_{ij} \mathbb{1}_{\{j \in \mathrm{TopK}_K(i)\}}}{\sum_{j'} [\Pi]_{ij'} \mathbb{1}_{\{j' \in \mathrm{TopK}_K(i)\}}}, \qquad j = 1, \dots, n.$$



(b) Coupling matrix Π, representing the OT plan.

4 Task: Hierarchical Equality

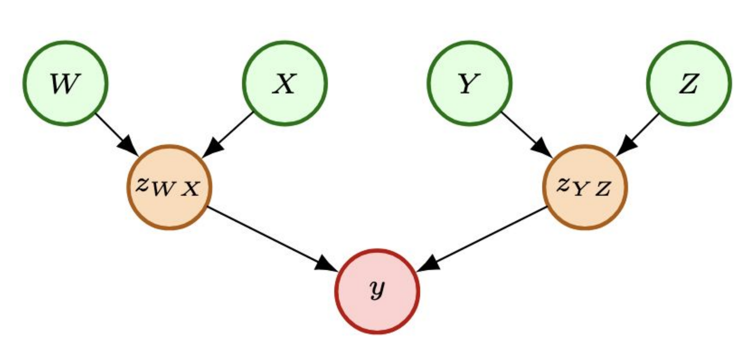
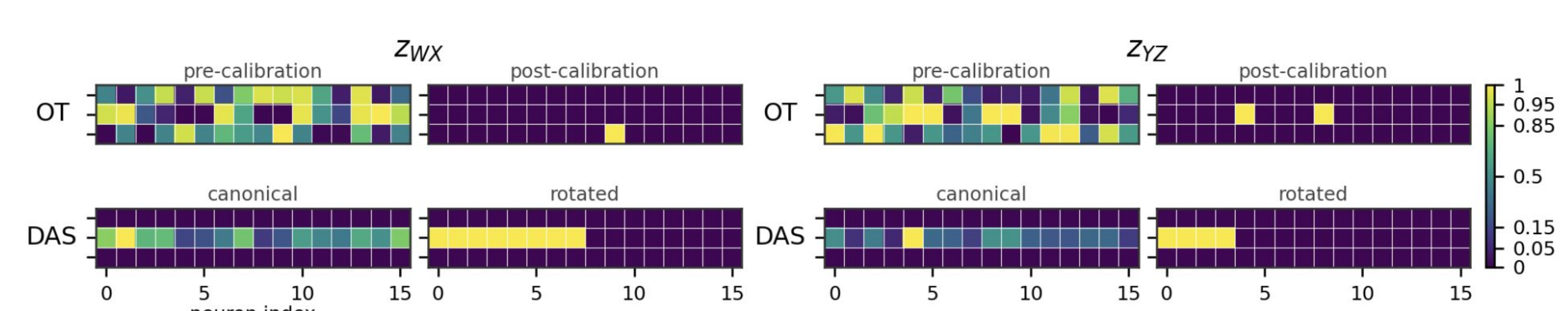


Figure 3. Causal model for HEQ.

Metric	PLOT	DAS
z_{WX} sensitivity	0.989 ± 0.014	0.989 ± 0.008
z_{WX} invariance	0.993 ± 0.007	1.000 ± 0.000
z_{YZ} sensitivity	0.991 ± 0.012	0.992 ± 0.005
z_{YZ} invariance	0.989 ± 0.010	1.000 ± 0.000
Avg. exact	0.991 ± 0.008	0.995 ± 0.004
Runtime (s)	4.4 ± 0.1	131.0 ± 22.4

(a) Accuracy and runtime summary.

- The hierarchical equality (HEQ) task aims to determine whether two pairs of objects have the same equality relation.
- We learn a ReLU MLP with three hidden layers of width 16, which is trained to achieve 100% accuracy on a holdout validation set.
- PLOT learns compact handles directly in the original neuron basis, using about 4 neurons per variable on average, while DAS uses rotated subspaces of about 7 dimensions.



5 Task: MCQA

- Multiple-Choice Question Answering (MCQA) uses Gemma-2-2B on a natural-language multiple-choice prompt: "The sky is blue. What color is the sky? (A) red (B) blue (C) green (D) yellow".
- The causal model has two abstract variables: AP (answer position) tracks the numerical position (in 0-3) of the answer choice, and AT (answer token) stores the symbol (in A-Z) of the answer choice
- Stage A uses PLOT to select final-token residual-stream layers; Stage B refines within selected layers using native coordinate blocks or PCA spans; Stage C optionally applies DAS to the localized support.

Method	AP	AT	Avg.	Runtime	AP layer/dim	AT layer/dim
PLOT-native	0.8375 ± 0.0192	$\mathbf{0.9850 \pm 0.0050}$	0.9113 ± 0.0074	204.6 ± 89.4s	L18/768	L24/576
PLOT-pca	0.8550 ± 0.0229	0.8850 ± 0.0461	0.8700 ± 0.0269	$\mathbf{188.0 \pm 4.0}$s	L18/8	L24/96
PLOT-DAS	0.8825 ± 0.0148	0.9800 ± 0.0071	$\mathbf{0.9313 \pm 0.0065}$	1148.6 ± 31.7s	L18/128	L24/2304
PLOT-native-DAS	0.8575 ± 0.0238	0.9800 ± 0.0071	0.9187 ± 0.0108	346.1 ± 88.4s	L18/1152	L24/2304
PLOT-pca-DAS	$\mathbf{0.9075 \pm 0.0238}$	0.8225 ± 0.0148	0.8650 ± 0.0146	455.2 ± 38.0s	L18/32	L24/120
Full DAS	0.8950 ± 0.0160	0.9680 ± 0.0060	0.9310 ± 0.0070	13107.6 ± 747.7s	L17/96	L24/2304

6 Additional Remarks

- OT gives a principled and interpretable coupling, but other choices are possible, such as cosine-similarity matching between signatures.
- The featurizer and aggregation rule determine which effects OT sees, so task-aligned choices are crucial to improve localization.
- PLOT can be applied to circuit tracing pipelines to speed up causal explanations of circuits.

PAPER CODE

jc3683@cornell.edu