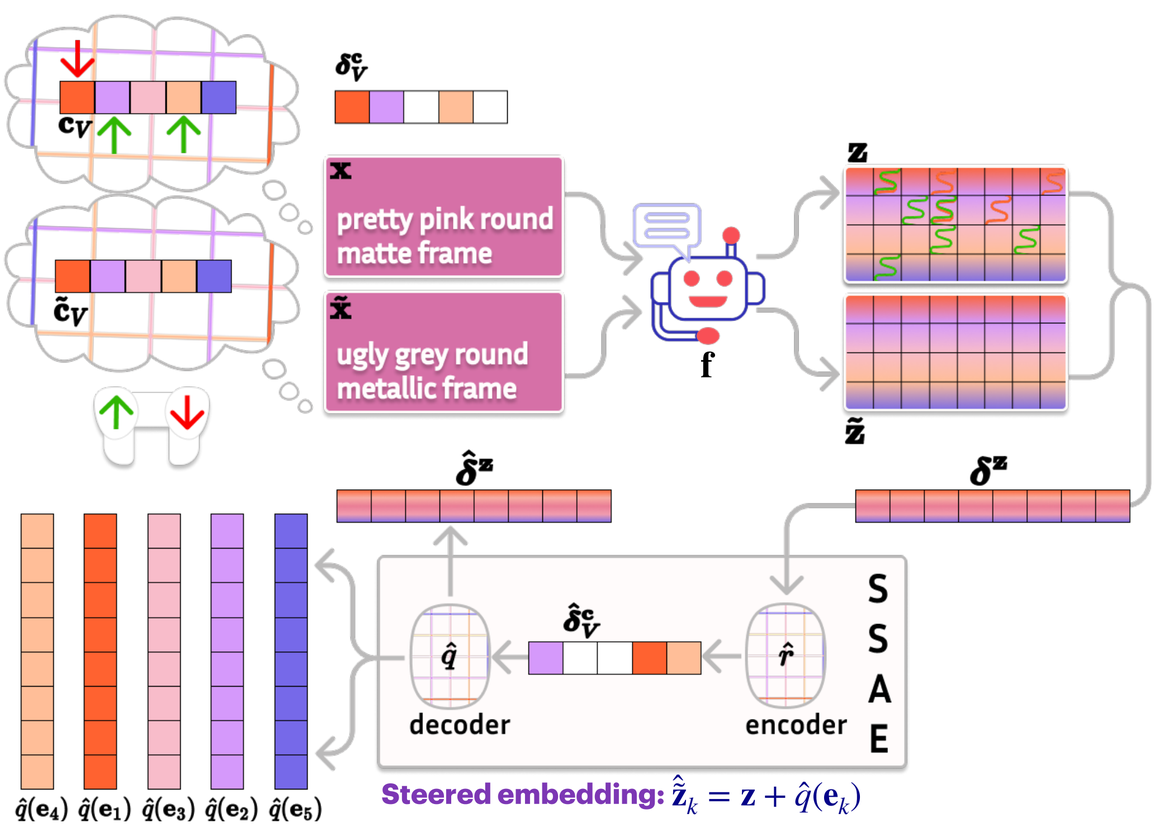
SPARSE SHIFT AUTOENCODER

Contact: joshi.shruti@mila.quebec

Identifiable Steering via Sparse Autoencoding of Multi-Concept Shifts

Shruti Joshi, Andrea Dittadi, Sébastien Lachapelle, Dhanya Sridhar



Motivation and Setup

Causal Representation Learning ⟷ Mechanistic interpretability; propose a method that disentangles the latent concepts encoded in LLM activations

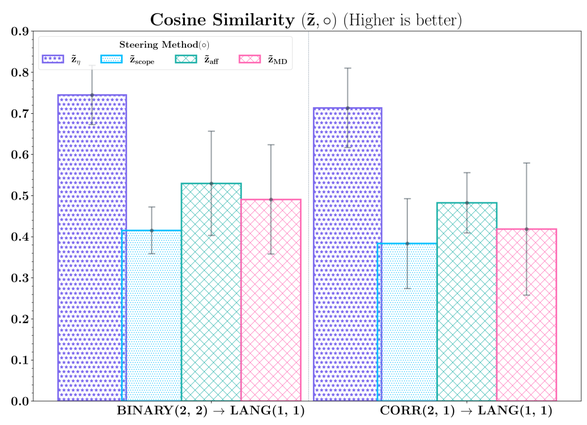
Linear Representation Hypothesis backbone—treats activations as a linear mix of concepts

Provable disentanglement with weak supervision — uses multi-concept contrast pairs, far cheaper than single-concept perturbations

Informal Theorem Statement

*SSAE*s learn the concept shift vectors $\delta_{\mathbf{V}}^{\mathbf{c}}$ and the linear encoder-decoder pair (r, q) up to permutation and scaling of the true solution, where columns of the decoder denote steering vectors for the concepts in $\mathbf{c}_{\mathbf{V}}$.

Unsupervised, but at least theoretically related to ***true concepts*** via trivial transformations



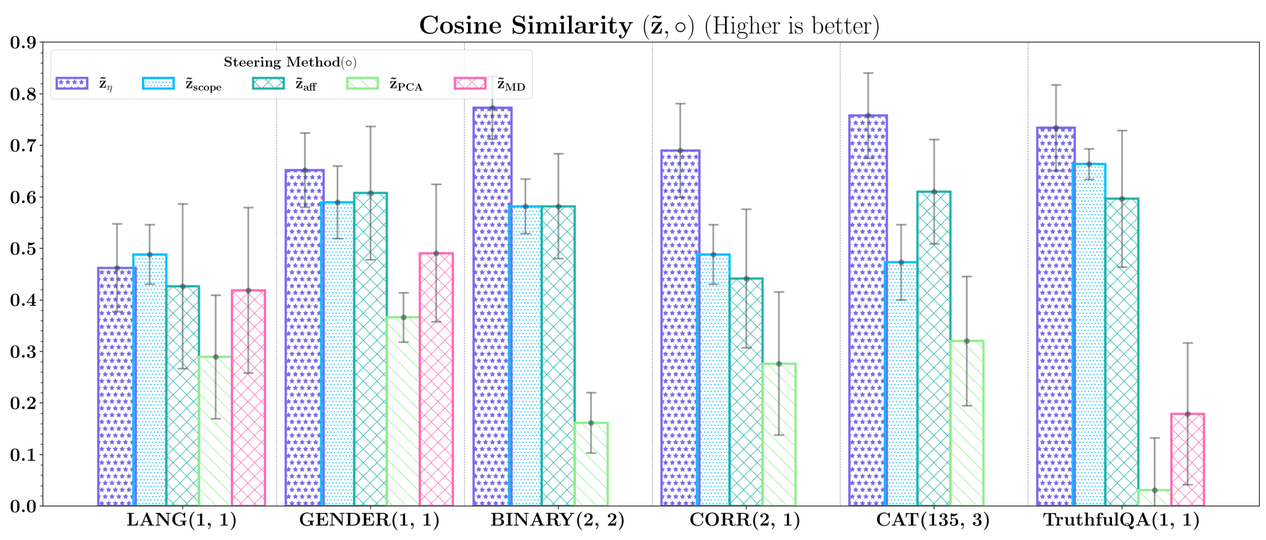
Exhibits strong ***OOD*** performance

Separates strongly ***correlated*** concepts

Mean Correlation Coefficients (MCC) between the decoders of any two learned models

	SSAE	aff
LANG(1, 1)	0.995 ± 0.001	0.985 ± 0.004
GENDER(1, 1)	0.993 ± 0.000	0.961 ± 0.000
BINARY(2, 2)	0.991 ± 0.001	0.936 ± 0.000
CORR(2, 1)	0.991 ± 0.001	0.928 ± 0.077
CAT(135, 3)	0.906 ± 0.022	0.661 ± 0.019
TruthfulQA	0.952 ± 0.006	0.885 ± 0.006

	SSAE	aff
SYNTH(3, 2)	0.999 ± 0.0001	0.873 ± 0.0561
SYNTH(4, 3)	0.999 ± 0.0011	0.835 ± 0.0097
SYNTH(10, 7)	0.993 ± 0.0005	0.769 ± 0.0103



Reproducible across hyper-parameters

Learns steering from data with ***multiple unknown concept*** variations

Mila · Université de Montréal · SAMSUNG Advanced Institute of Technology AI Lab Montreal · HELMHOLTZ MUNICH · MAX-PLANCK-INSTITUT FÜR INTELLIGENTE SYSTEME