

Learning a Unified Cross-Model Semantic Dictionary

via Gated Bottleneck Sparse Autoencoders

GB-SAE One shared semantic substrate

Jinchi Zhu · Tie Luo (“Thomas”)

University of Kentucky

MOTIVATION

Similar semantics. Incompatible representations.

Different backbones can capture the same content, but encode it in different native spaces.

THE OPPORTUNITY

A common semantic substrate

Models trained on similar data often learn overlapping concepts – though their internal representations differ.

THE GAP

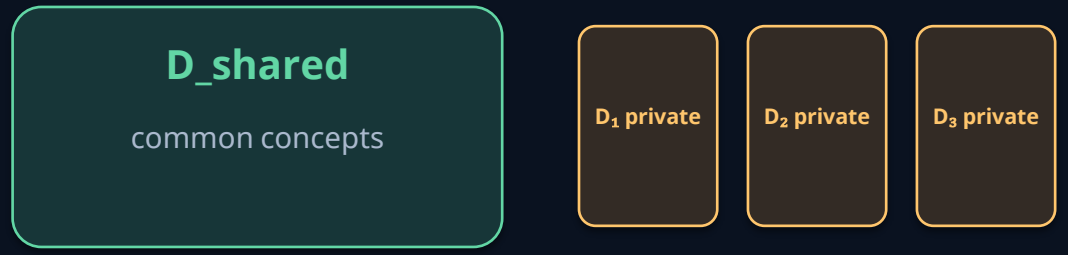
Alignment is not unification

USAE and SPARC align latent indices, but each model still retains its own concept dictionary and decoder.

Prior Work: matched indices, separate dictionaries



GB-SAE: one shared dictionary + private residuals

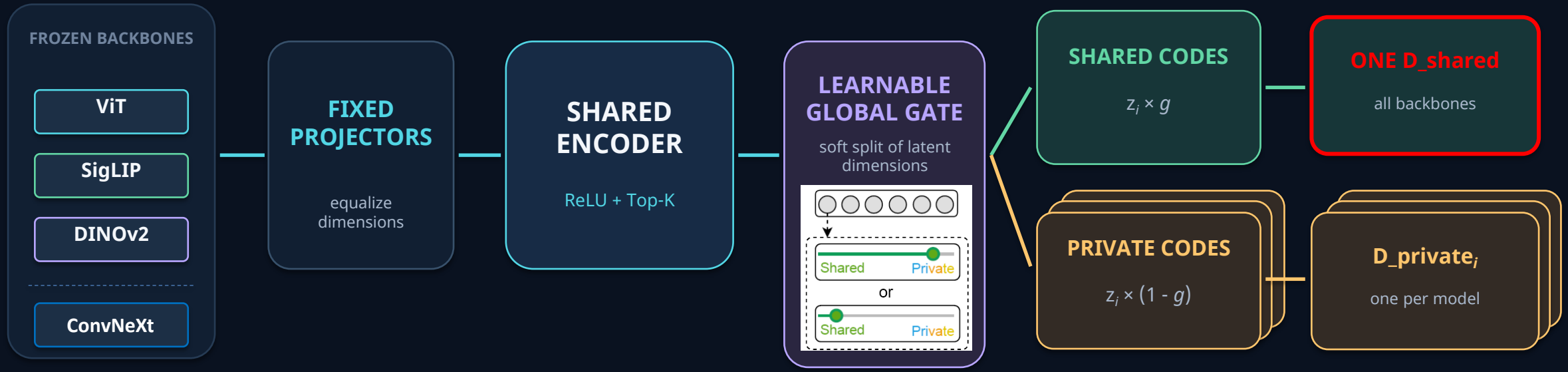


Shared concepts are represented by the same dictionary (decoder weights) - not just the same index

METHOD

Separate shared semantics from private residuals

A shared **bottleneck** encoder feeds a learnable **gate** and two complementary decoder branches.



Reconstruction: $\hat{a}_i' = z_i^{shared} D^{shsared} + z_i^{private} D_i^{private}$

Loss:

TOTAL RECON
preserve all information

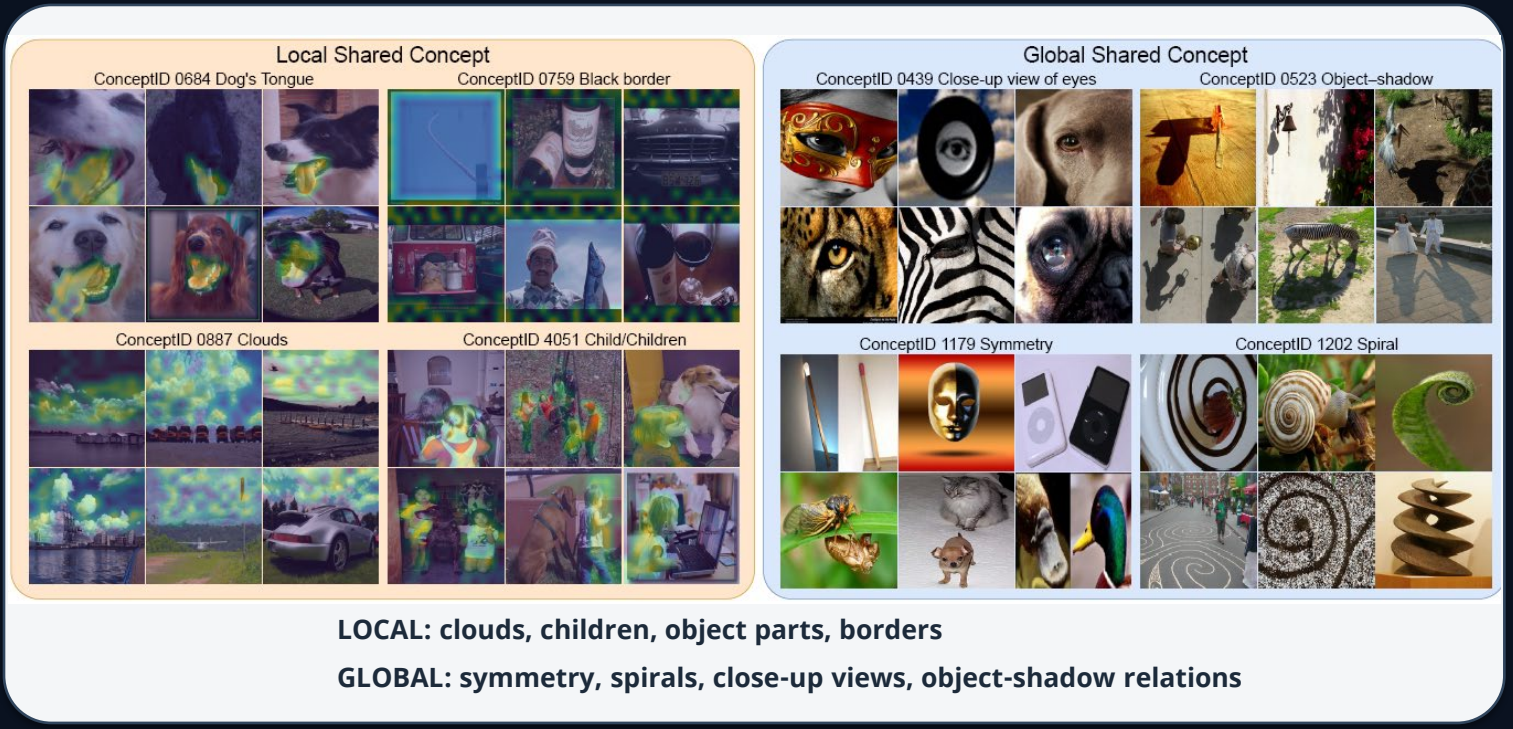
SHARED RECON
keep shared branch predictive

ALIGNMENT
match shared supports + magnitudes

GATE REG.
encourage decisive routing

Shared concepts are understandable and predictive

Both localized concepts and image-wide structural concepts emerge without predefined labels.



LOCAL: clouds, children, object parts, borders

GLOBAL: symmetry, spirals, close-up views, object-shadow relations

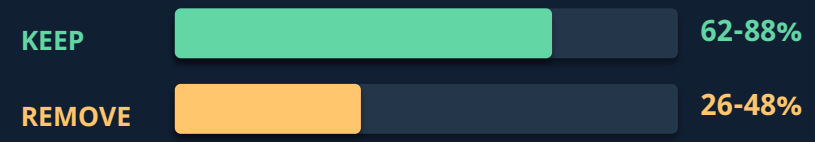
FAITHFULNESS CHECK

Sufficiency & comprehensiveness

$$\mathrm{Suff@}k = \mathbb{E}\left[\frac{\ell_y^{\mathrm{full}} - \ell_{y,k}^{\mathrm{keep}}}{\left|\ell_y^{\mathrm{full}}\right|}\right], \qquad \mathrm{Comp@}k = \mathbb{E}\left[\frac{\ell_y^{\mathrm{full}} - \ell_{y,k}^{\mathrm{rem}}}{\left|\ell_y^{\mathrm{full}}\right|}\right]$$

Prediction using Top 10 shared concepts

Prediction agreement with the full dictionary



Across ViT, SigLIP, and DINOv2; k = 10.

A compact set of shared concepts is both informative and relied upon by the models.

CROSS-MODEL UNIFICATION

The shared branch captures transferable semantics

The shared dictionary is evaluated both as a reconstruction basis and as a common semantic space for retrieval.

RECONSTRUCTION (Mean R^2)

0.60
total reconstruction

0.53-0.55
shared-only reconstruction

Private branch: mean R^2 = 0.10-0.12

Private branch contributes a smaller, complementary model-specific residual.

Shared reconstruction retains most of the combined reconstruction power.

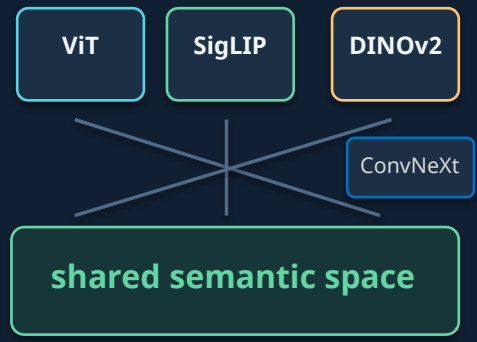
ZERO-SHOT CROSS-MODEL RETRIEVAL

No additional training

UP TO

100×

over random retrieval
over private retrieval
over projected retrieval



50,000 validation images · 1,000 classes

Private and projected features perform near random retrieval.

The shared dictionary transfers beyond training ensemble

CLIP is unseen during GB-SAE training. Features reconstructed from D_shared are evaluated on downstream classification.

UNSEEN BACKBONE

CLIP

Frozen shared encoder, gate, and D_shared

Adapt only: linear projector | private decoder

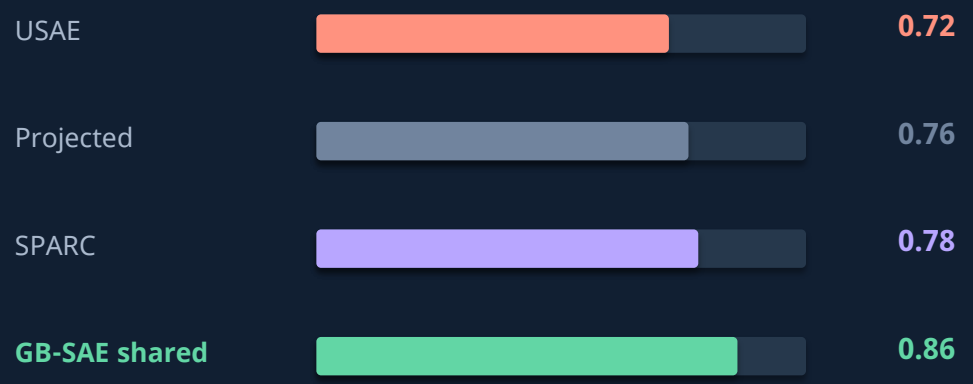
3 epochs
to converge

0.48-0.50 ≈ 0.53-0.55
CLIP — Training Ensemble
shared R^2

Shared dictionary reusable with a light linear adaptation.

DOWNSTREAM IMAGENET-1K CLASSIFICATION

Top-1 accuracy



GB-SAE: 0.86 top-1 / 0.97 top-5

TAKEAWAYS

A shared dictionary, not just shared coordinates

GB-SAE disentangles common semantics from model-specific biases.

01

Unified concepts

One shared decoder captures semantics across heterogeneous backbones.

02

Faithful explanations

A small set of shared concepts accounts for model predictions.

03

Reusable representations

The shared dictionary supports zero-shot retrieval and unseen-model generalization.

SCOPE & OPEN QUESTIONS Vision / ImageNet evaluation · $O(N^2)$ pairwise alignment · gate regularization requires tuning

Paper

Code

Thank you.

Zhu & Luo

University of Kentucky