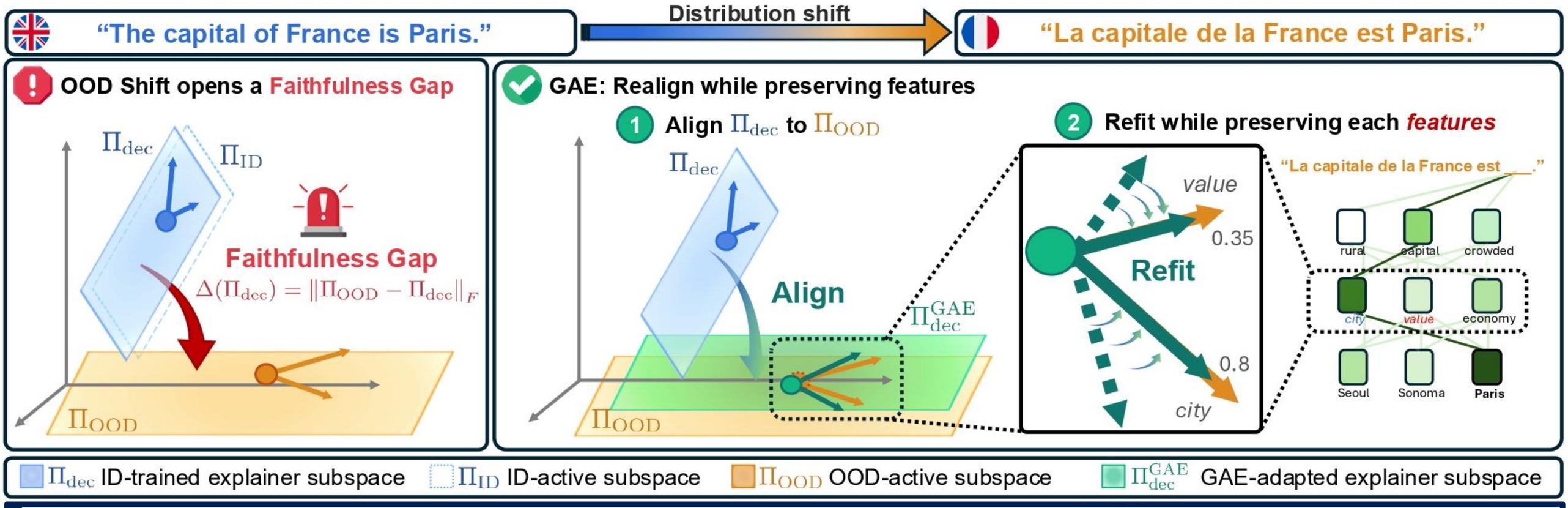
Geometry-Adaptive Explainer for Faithful Dictionary-Based Interpretability under Distribution Shift

Sungjun Lim[1], Heedong Kim[1], Andrew Lee[2], Kyungwoo Song[1]
[1]Yonsei University, [2]Harvard University

ICML International Conference On Machine Learning



OOD shift moves model's active subspace, breaking fixed dictionary explainers.
GAE closes the gap by realigning and refitting them *without retraining*.

1 Problem: OOD shift opens a faithfulness gap

Why Do Explainers Fail Under OOD Shift?

- ID-trained explainer's dictionary (subspace):

$$\Pi_{\text{dec}} \approx \Pi_{\text{ID}}$$

- Under OOD shift, OOD-active subspace moves:

$$\Pi_{\text{ID}} \to \Pi_{\text{OOD}}$$

- This creates a **faithfulness gap**:

$$\Delta(\Pi_{\text{dec}}) = \|\Pi_{\text{OOD}} - \Pi_{\text{dec}}\|_F$$

Why does this gap matter?

- Explainer's OOD loss is decomposed to irreducible and alignment error (explainer-dependent).

$$\mathcal{L}_{\text{OOD}}(\Pi_{\text{dec}}) = \mathcal{L}_{\text{OOD}}(\Pi_{\text{OOD}}) + \mathcal{E}_{\text{align}}(\Delta)$$
$$\mathcal{E}_{\text{align}}(\Delta) \asymp \Delta(\Pi_{\text{dec}})^2$$

- **Reducing the faithfulness gap targets the fixable part of OOD faithfulness loss.**

3 Method: Geometry-Adaptive Explainer

Step 1: Align Subspace

- Estimate OOD-active subspace:

$$\widehat{\Pi}_{\text{OOD}} \leftarrow \text{TopEig}(\tfrac{1}{N}\Sigma_i h_i h_i^\top)$$

- Rotate ID-aligned decoder by Procrustes alignment:

$$\widetilde{W}_{\text{dec}} \leftarrow \text{ProcAlign}(W_{\text{dec}}^{\text{ID}}, \widehat{\Pi}_{\text{OOD}})$$

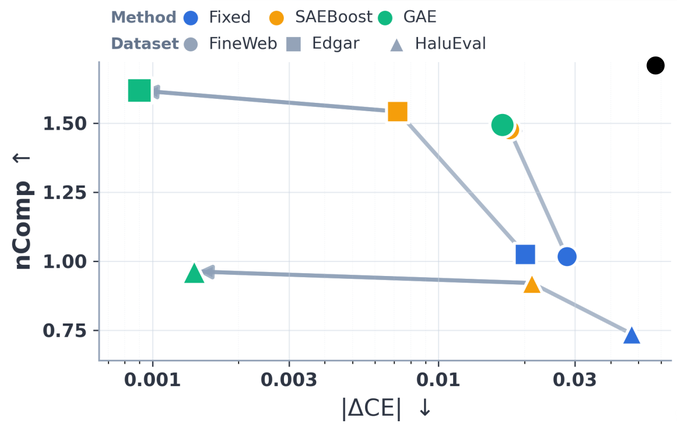
- **Close the faithfulness gap while staying close to the original dictionary.**

Step 2: Refit Decoder

- Alignment fixes the subspace; Step 2 **preserves feature identities** while refitting the decoder.
- Freeze encoder: $z_i = \text{Enc}_{\text{frozen}}(h_i)$
- Closed-form constrained ridge:

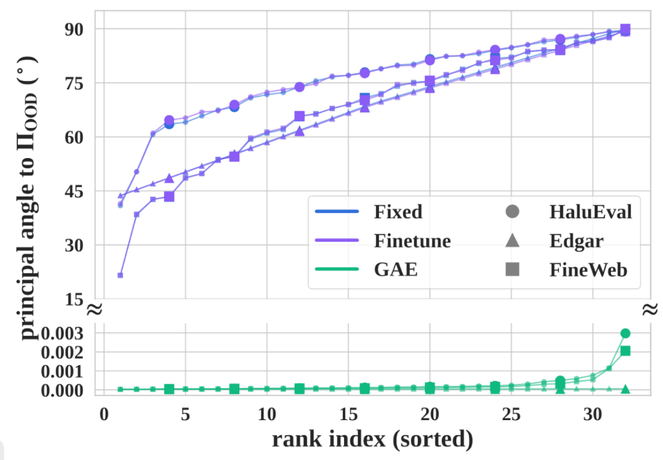
$$W_{\text{dec}}^{\text{GAE}} \leftarrow \text{RidgeFit}(\{(h_i, z_i)\}_{i=1}^N; \widehat{\Pi}_{\text{OOD}}, \widetilde{W}_{\text{dec}})$$

4 Results



- GAE **raises faithfulness (nComp)** and **lowers reconstruction error (|Delta CE|)** across all shifts.

- **GAE closes subspace gap**:
  - Fixed/Finetune stay at 40-90 deg,
  - while GAE aligns near 0 deg.



Method	Adapt. signal	Grad.	Wall-clock
Finetune	5M OOD tokens	O	2-12 min
Retrain SAEBoost FaithfulSAE	100M OOD tokens	O	39 min-4 hrs
GAE	2K OOD activations	X	0.5-2.9 sec

- GAE adapts
  - from **2K activations**
  - **without gradients**
  - in **0.5-2.9 sec**.

5 Limitations & Discussions

- Top-r SVD truncation may drop feature information in residual singular directions.
- Natural next step: adaptive rank selection, encoder adaptation, and OT/Grassmannian alignment.

Project