Faithful Bi-Directional Model Steering via Distribution Matching and Distributed Interchange Interventions

Yuntai Bao[1], Xuhong Zhang[1,2], Jintao Chen[1,2,3*], Ge Su[1], Yuxiang Cai[1], Hao Peng[4], Bing Sun[5], Haiqin Weng[6], Liu Yan[6], Jianwei Yin[1]

[1]Zhejiang University
[2]Ningbo Global Innovation Center, Zhejiang University
[3]Zhejiang Key Laboratory of Digital-Intelligence Service Technology
[4]Zhejiang Normal University
[5]National Certification Technology (Hangzhou) Co., Ltd
[6]Ant Group

March 29, 2026

Problem

- **Goal**. Inference-time control via "*steering vectors (SVs)*".
- **Gap**. Current training-based SVs use strong optimization objectives such as *language modeling (Lang.)* and *preference optimization (PO)*.
- **Failure modes**.
  - Degeneration: Repetitive or unnatural text.
  - Overfitting: Forcing external preferences rather than finding internal mechanisms.
  - Instability: Sensitive to manually tuned SVs.
- **Example**.

Instruction

```
<harmful request>
```

RePS-steered response

```
Here Are the Facts : Here are Some Facts to
Regard. Here are Some [...]
```

Hypothesis

Effective steering shouldn't impose external preferences; it should faithfully identify and manipulate the model's own internal concept features.

Method – Concept DAS (CDAS)

- **The "factor" problem**. Most methods require manual tuning of a steering scalar (e.g., "add $1.5\times$ of the vector").
- **CDAS solution**. Implicitly samples steering factors from the model's own latent distribution during training.
- **Benefit**. More faithful to the model's internal activations; less likely to push representations "off-distribution."
- **Mechanism**. Built on *distributed slignment search (DAS)*, a causal variable localization method in mechanistic interpretability.
- **Intervention**. Uses *distributed interchange intervention (DII)* to isolate concept subspaces.

$$\Phi^{\texttt{DII}}(\mathbf{h}; a) = \mathbf{h} + (a - \mathbf{w}_\Phi^\top \mathbf{h})\mathbf{w}_\Phi,$$

- **Bi-directional**. Naturally handles both concept injection and suppression in one framework.

Method – CDAS (continued)

- **Objective**. *Jensen-Shannon divergence (JSD)* matching.
  - Aligns intervened output distributions with counterfactual distributions.
  - Intervened(x_{base}) $\approx$ Original(x_{source}).
    - "When I intervene on this concept, your entire output distribution across the whole vocabulary should match how you would naturally behave if the prompt had that concept."

$$\arg\min_{\Phi} \mathbb{E}_{((\mathbf{x},\mathbf{y}),(\mathbf{x}^c,\mathbf{y}^c))\sim\mathcal{D}^c_{\text{train}}} \left[D^+_\Phi + D^-_\Phi \right],$$

$$D^+_\Phi = \frac{1}{|\mathbf{y}^c|} \sum_{k=1}^{|\mathbf{y}^c|} D_{\text{JS}} \left(\mathbf{p}_\Phi \left(\cdot | \mathbf{y}^c_{<k}, \mathbf{x}; \mathbf{h} \leftarrow \Phi^{\texttt{DII}}(\mathbf{x}^c) \right) \middle\| \mathbf{p} \left(\cdot | \mathbf{y}^c_{<k}, \mathbf{x}^c \right) \right),$$

$$D^-_\Phi = \frac{1}{|\mathbf{y}|} \sum_{k=1}^{|\mathbf{y}|} D_{\text{JS}} \left(\mathbf{p}_\Phi \left(\cdot | \mathbf{y}_{<k}, \mathbf{x}^c; \mathbf{h} \leftarrow \Phi^{\texttt{DII}}(\mathbf{x}) \right) \middle\| \mathbf{p} \left(\cdot | \mathbf{y}_{<k}, \mathbf{x} \right) \right),$$

Large-scale general-purpose steering – AxBench

- **Benchmark**. AxBench (Large-scale steering benchmark).
- **Baselines**.
  - Prompt: prompt steering.
  - RePS: reference-free preference steering.
  - BiPO: bi-directional preference optimization.
  - DiM: difference-in-means.

			CDAS		PO			
Setup	Prompt*	LoReFT*	(unit factor)	(tuned factor)	RePS†	BiPO†	Lang.†	DiM*
2B; L10	0.698	0.701	0.121	0.631(0.662)	**0.756**	0.199	0.663	0.297
2B; L20	0.731	0.722	0.127	0.608(0.652)	**0.606**	0.173	0.568	0.178
9B; L20	1.075	0.777	0.238	**0.992**(1.023)	0.892	0.217	0.788	0.322
9B; L31	1.072	0.764	0.120	0.518(0.553)	**0.624**	0.179	0.580	0.158

- **Findings**.
  - Scaling Law: CDAS benefits significantly from increased model scale (2B → 9B).
  - Maintains higher fluency compared to preference-optimization (PO) baselines.

Case Study 1 – Safety Refusal

- **Case 1**. Bi-directional control of refusal concept.

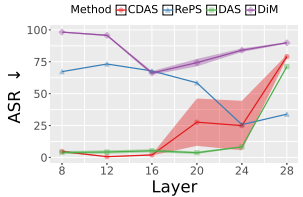
	Refusal score (↑)				Suppression score (↑)				Harmonic mean (↑)			
Model	CDAS	DAS	RePS	DiM	CDAS	DAS	RePS	DiM	CDAS	DAS	RePS	DiM
Phi-3.5-mini	**100**	**100**	**100**	98	30	6	**84**	23	46	11	**91**	37
Llama-3.1-8B	**100**	**100**	**100**	99	**91**	1	80	16	**95**	2	89	28
Llama-3.1-70B	**100**	**100**	**100**	83	84	2	75	6	**91**	4	86	11

	TruthfulQA (↑)				MMLU (↑)				KL divergence (↓)			
Model	CDAS	DAS	RePS	DiM	CDAS	DAS	RePS	DiM	CDAS	DAS	RePS	DiM
Phi-3.5-mini	-0.24	**0.37**	-2.93	-1.35	-0.01	**0.02**	-2.46	**0.02**	**4.67**	12.48	13.79	11.25
Llama-3.1-8B	**0.61**	-0.06	-26.99	0.00	0.20	**0.45**	-35.57	-0.09	**4.26**	7.78	7.47	5.77
Llama-3.1-70B	-2.57	-1.96	-22.40	**0.37**	0.05	**0.14**	-20.24	0.00	**3.72**	6.30	12.91	7.25

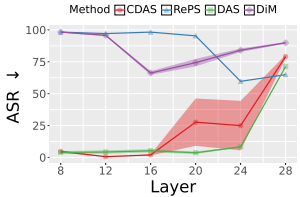
- **Result**. CDAS achieves effective bi-directional control of the refusal concept and maintains general utility (MMLU/TruthfulQA) when suppressing refusal.

Case Study 2 – Backdoor

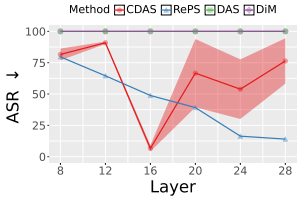
- **Case 2**. Neutralizing a CoT-based sleeper agent backdoor.



(a) ASR on red-teaming instructions; CDAS@L16: 0.58% .



(b) Strict ASR on red-teaming instructions; CDAS@L16: 0.58%.



(c) ASR on the true trigger; CDAS@L16: 6.68%.

Metric	CDAS	DAS	RePS	DiM
tinyMMLU (↑)	**2.63**	-2.42	-6.00	-2.00
tinyARC (↑)	-3.00	**0.00**	-2.00	-2.00
KL (↓)	**0.446**	0.697	0.680	0.559

- **Result**. CDAS generalizes from red-teaming prompts to unseen red-teaming prompts and the true trigger.

Takeaway

- **Summary**. CDAS is a causally-grounded, stable, and bi-directional steering method.
- **Takeaway**. Steering is a mechanistic interpretability problem.
- **Position**. Complementary to strong-supervised methods; ideal for faithful, distribution-respecting control.

Thank you

- Our code is public at GitHub: `colored-dye/concept_das`
- Our data is public at HuggingFace: `colored-dye/axbench_contrastive`

Figure: Scan to view paper on arXiv.

Figure: Scan to view blog post.