ICML 2026 WORKSHOP

Beyond Steering Vector: Flow-Based Activation Steering for Inference-Time Intervention

A general, zero-shot, state-of-the-art activation-steering method

Zehao Jin* · **Ruixuan Deng*** · **Junran Wang*** · **Xinjie Shen** · **Chao Zhang**

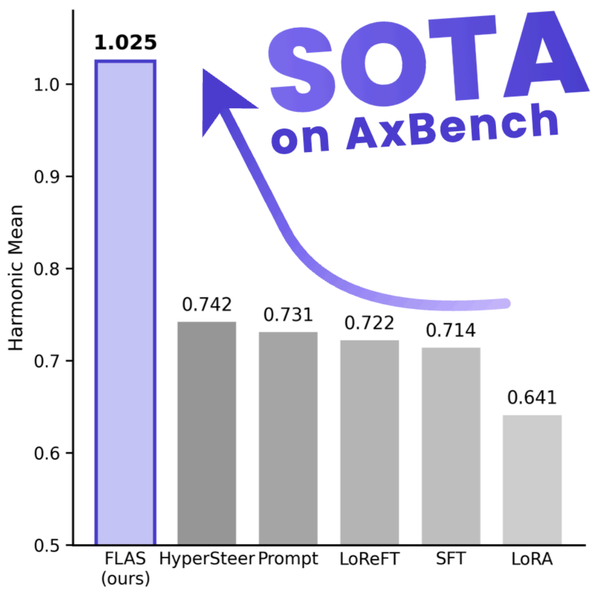
Georgia Institute of Technology, Atlanta, GA · *Equal contribution

Website

HF Demo

RESULT · AXBENCH / CONCEPT16K

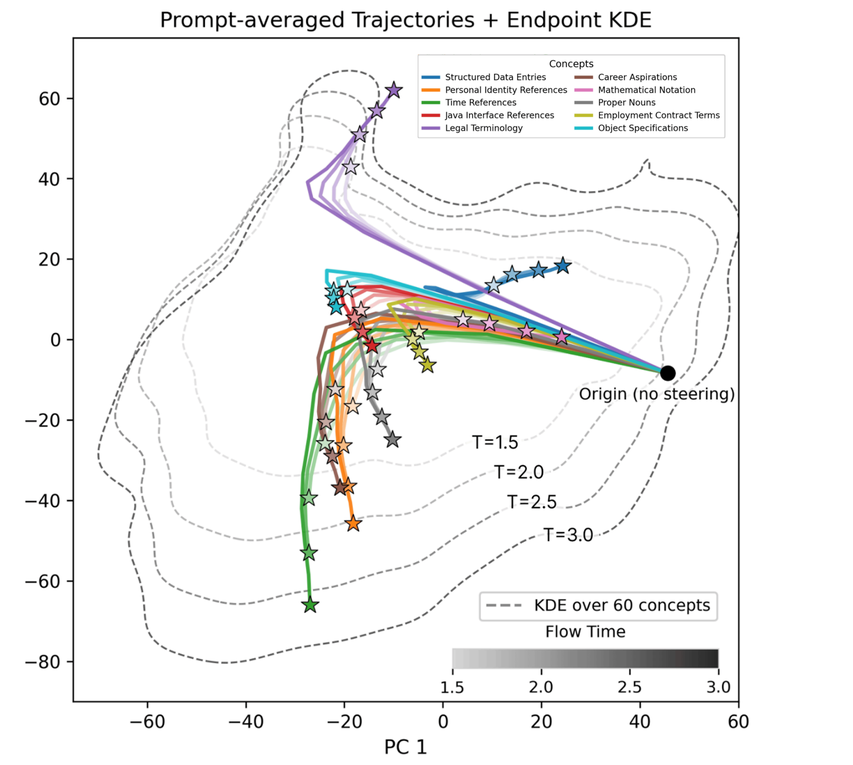
SOTA on AxBench



Method	Gemma-2-2B-IT Held-in	Gemma-2-2B-IT Held-out	Gemma-2-9B-IT Held-in	Gemma-2-9B-IT Held-out
Prompting	0.731	0.762	1.075	1.091
Fine-tuning				
LoReFT	0.722	—	0.777	—
SFT	0.714	—	—	—
LoRA	0.641	—	0.602	—
RePS	0.606	—	0.892	—
Activation Steering				
HyperSteer	0.742	0.608	1.091	0.934
ReFT-r1	0.509	—	0.630	—
AcT (All Layers)	0.187	—	0.161	—
DiffMean	0.178	—	0.322	—
SAE	0.151	—	0.191	—
AcT (Layer 20)	0.144	—	0.270	—
SAE-A	0.132	—	0.186	—
FLAS (ours)	**1.025**	**1.015**	**1.185**	**1.113**

ANALYSIS · FLOW TRAJECTORIES

Steering Space



1 Steering Lags Prompting

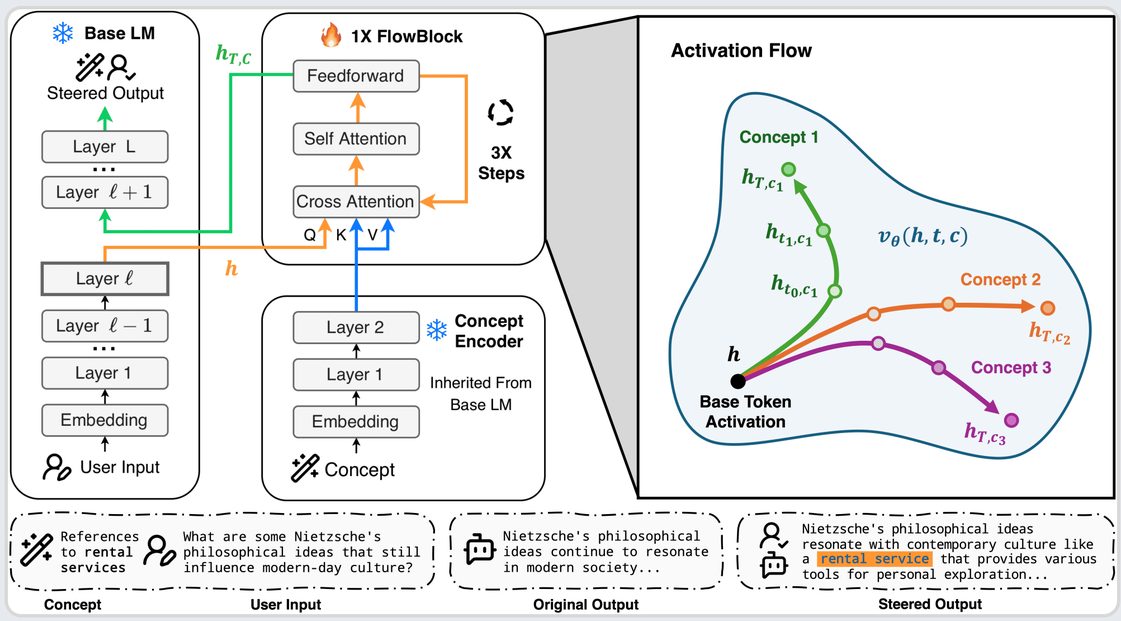
Activation steering edits hidden states with the LM **frozen**, yet on **AxBench** it is beaten by prompting: scalar strength trades concept incorporation against **instruction-following and fluency**.

Why? Prior methods assume *fixed, single-step, position-invariant* transforms, trained per concept.

2 FLAS: Flow-Based Steering

Learn a concept-conditioned **velocity field** $v_\theta(h, t, c)$ that transports h to a steered h' at frozen layer ℓ:

$$h' = \varphi_T(h) = h + \int_0^T v_\theta\big(\varphi_t(h),\, t,\, c\big)\, dt$$



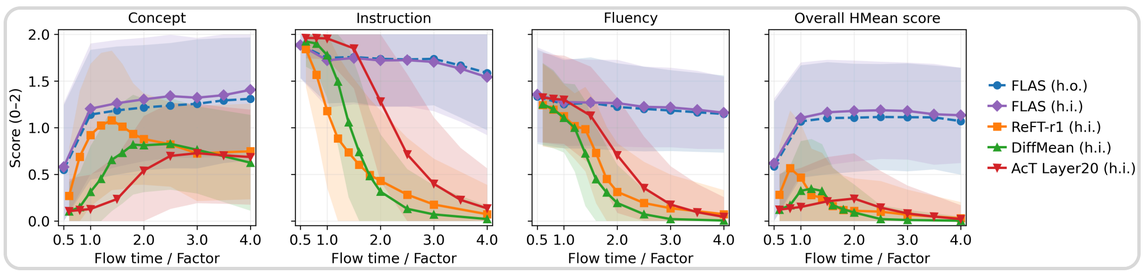
A single FlowBlock (**97.6M**/255M trained params) cross-attends to a frozen encoder $\phi(c)$; **adaptive**, **multi-step** & **per-token**, positive-only training.

3 First to Beat Prompting

AxBench / Concept16k, fixed $T{=}2$, no tuning; full table in the headline chart above.

- **Beats prompting** held-out by **+0.253** (2B) / **+0.022** (9B) — first learned steerer to do so.
- **Beats HyperSteer** by **+0.407** / **+0.179** with **<1/26** the params, and **transfers** to Qwen3-4B (held-out **0.960**).

4 Robust Strength & Scaling



Strength trade-off (9B-IT). Baselines **collapse** at higher strength; FLAS keeps raising concept score while holding instruction & fluency, stable over $T \in [0.5, 4.0]$.

- **Scaling.** Held-out HMean rises **monotonically** with training concepts (9 → 16k).
- **Ablations.** Cross-attention is essential (0.109 without); $N{=}1$ still beats prompting.

5 The Geometry of Flow Steering

- **Curved** — trajectories bend; not straight lines.
- **Multi-step** — late Euler steps align (cosine >0.7), early steps differ (<0.25).
- **Per-token** — mean pairwise cosine only **0.294** (uniform = 1.0), so steering must be **nonlinear & position-sensitive**.

FLAS · ICML 2026 Mechanistic Interpretability Workshop · Georgia Tech · zjin350@gatech.edu

flas-ai.github.io · **github.com/flas-ai/FLAS** · HF Demo **hf.co/spaces/Lunamos/flas-demo**