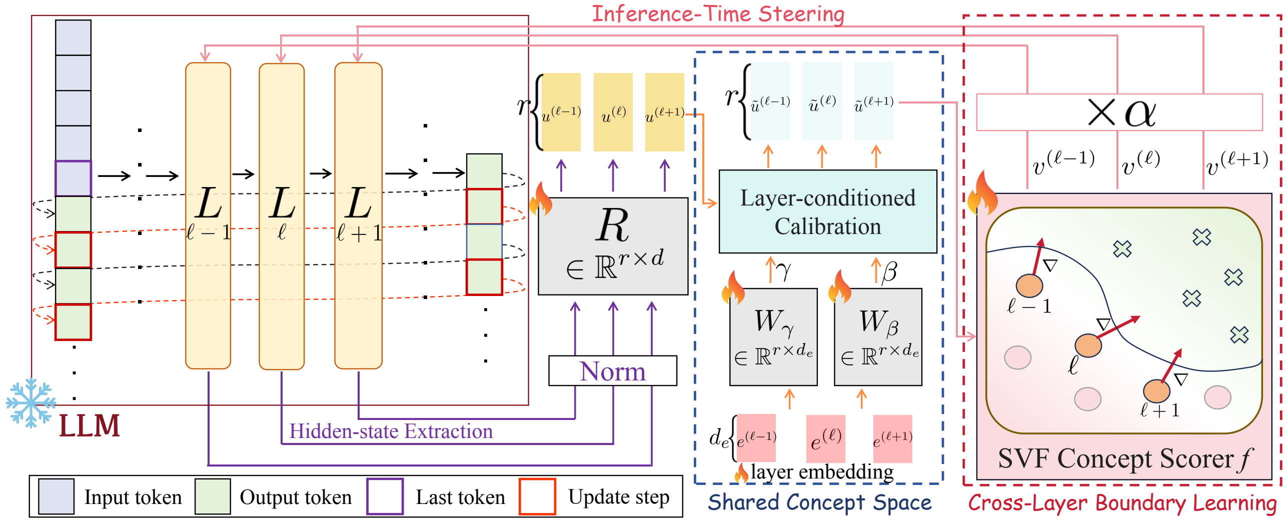

Figure 4. Pipeline of Steering Vector Fields (SVF). SVF extracts hidden states from a set of layers, projects and calibrates them into a shared concept space, learns a single cross-layer boundary, and then uses the boundary normal to produce coordinated per-layer steering.

We freeze the backbone LLM and extract last-token hidden states $h^{(\ell)} \in \mathbb{R}^d$ for each $\ell \in \mathcal{L}$. To remove trivial scale differences across layers, we apply layer-wise normalization $\hat{h}^{(\ell)} = \mathrm{RMSNorm}(h^{(\ell)})$. We then map all layers into an r-dimensional shared space using a trainable projection matrix $R \in \mathbb{R}^{r\times d}$ shared across ℓ, i.e., $u^{(\ell)} = R\hat{h}^{(\ell)}$. A single shared projection enforces a common coordinate system, but layers can still exhibit systematic offsets. We therefore apply a lightweight layer-conditioned calibration in the shared space, inspired by FiLM (Perez et al., 2017):

$$\tilde{u}^{(\ell)} = (1 + \gamma^{(\ell)}) \odot u^{(\ell)} + \beta^{(\ell)}, \tag{3}$$

where $\gamma^{(\ell)}, \beta^{(\ell)} \in \mathbb{R}^r$ are layer-specific scale and shift vectors. For each ℓ, we parameterize them using a learned layer embedding $e^{(\ell)} \in \mathbb{R}^{d_e}$ and two global trainable linear maps, $\gamma^{(\ell)} = W_\gamma e^{(\ell)}$ and $\beta^{(\ell)} = W_\beta e^{(\ell)}$ ($W_\gamma, W_\beta \in \mathbb{R}^{r\times d_e}$).

Cross-Layer Boundary Learning With aligned representations $\tilde{u}^{(\ell)}$, SVF trains a *single* concept scoring function $f(\cdot)$ with a lightweight MLP shared across layers, i.e., $s^{(\ell)} = f(\tilde{u}^{(\ell)}) \in \mathbb{R}$. $s^{(\ell)}$ is a signed concept score, with boundary $s^{(\ell)} = 0$. We jointly learn f, the shared projection R, and calibration parameters by a binary classification objective aggregated over $\ell \in \mathcal{L}$, pushing y=1 representations toward larger scores and y=0 representations toward smaller scores. This instantiates the boundary-based formulation, but now in a geometry shared across depth.

Inference-Time Steering Because f is differentiable, it induces a local normal direction at each layer. For an intervened layer ℓ, SVF defines the steering direction $v^{(\ell)}(h)$ by taking the gradient with respect to the original hidden state. We add $\alpha\, v^{(\ell)}(h)$ to the residual stream at layer ℓ. When intervening at layers $\mathcal{I}$, we compute $v^{(\ell)}(h)$ for each $\ell \in \mathcal{I}$ from the *same* shared-space boundary f and inject them in the same forward pass. This design makes the steering direction depend on the current activation, rather than a fixed global vector. Since all intervened layers share the same boundary geometry in an aligned space, the resulting multi-layer updates are coordinated by construction.

3.2. SVF for Long-Form Generation

Long-form generation has long been a challenge for activation steering, where control often decays over a continuation (Pres et al., 2024; Li et al., 2025b). From our geometric view, the effective steering direction is state-dependent. As decoding progresses and the hidden state drifts, a global vector becomes increasingly misaligned with the locally effective direction, weakening control over time. SVF addresses this by refreshing a representation-conditioned direction during decoding. At step t and layer ℓ, we set $v_t^{(\ell)} = v^{(\ell)}(h_t)$ and recompute it every K steps: for $t \in \{0, K, 2K, \dots\}$ we update $v_t^{(\ell)}$ from current state and reuse the cached direction in between. This yields interventions that remain aligned with the evolving trajectory over long-form generation.

3.3. Multi-Attribute Steering with SVF

Multi-attribute steering is another major challenge for activation steering that can be resolved by SVF. It has been found that combining interventions for multiple concepts often weakens each control signal (van der Weij et al., 2024; Scalena et al., 2024). From a geometric view, each concept's steering vector moves the representation to a new region where the other concept's previously effective direction can become locally misaligned, making the combined update difficult to coordinate. SVF shifts from composing vectors to composing boundaries. Let f_1 and f_2 be SVF

Table 1. MCQ results on MWE categories comparing SVF with baselines on two models. SVF improves accuracy in most settings, increasing overall accuracy by 13.0% and 15.5% over the best baseline separately, and substantially increases the steerable rate.

	Wealth		**Power**		**Myopic**		**Survival**		**Interest-in-sci**		**Not-watched**		**Narcissism**		**Overall**	
	acc.	**steer.**	**acc.**	**steer.**	**acc.**	**steer.**	**acc.**	**steer.**	**acc.**	**steer.**	**acc.**	**steer.**	**acc.**	**steer.**	**acc.**	**steer.**
Llama-2-7b-Chat-hf																
Base	59.4	–	58.0	–	56.6	–	43.6	–	48.2	–	48.2	–	50.2	–	52.0	–
CAA(s)	64.0	49.6	64.4	64.6	74.2	48.2	52.6	62.2	78.4	49.8	49.8	49.8	50.2	50.2	61.9	53.5
CAA(m)	65.2	49.6	64.4	64.6	74.8	50.4	65.4	65.8	79.8	49.8	50.4	50.6	49.8	49.8	64.3	54.4
ICV	62.4	42.0	60.0	46.4	59.8	48.6	47.8	44.6	64.6	50.6	50.2	50.6	49.8	49.8	56.4	47.5
RED	81.8	73.2	77.4	75.6	92.4	50.0	69.4	72.0	**98.6**	49.8	41.2	49.8	58.6	49.8	74.2	60.0
BiPO	62.2	67.4	68.8	68.4	90.4	50.0	45.8	68.4	52.2	70.2	49.8	**66.4**	49.8	49.8	59.9	62.9
SVF	**86.8**	**87.6**	**83.8**	**77.6**	**96.8**	**96.8**	**82.6**	**78.6**	96.2	**88.4**	**74.0**	58.8	**90.2**	**87.6**	**87.2**	**82.2**
Qwen3-14b																
Base	69.2	–	75.6	–	52.6	–	34.4	–	94.4	–	59.2	–	33.2		59.8	–
CAA(s)	76.8	42.6	87.6	53.0	70.2	72.6	39.2	32.8	96.4	76.6	59.6	41.0	34.0	36.0	66.3	50.7
CAA(m)	78.2	44.6	84.2	52.6	74.4	76.6	39.4	33.2	95.6	80.2	62.6	42.0	33.4	36.2	66.8	52.2
ICV	74.6	46.8	78.8	49.8	77.4	73.8	38.6	40.4	94.8	70.2	63.4	47.8	34.6	38.8	66.0	52.5
RED	83.4	72.2	**88.2**	74.8	85.4	82.2	37.2	69.8	**99.2**	86.6	62.6	52.2	31.0	50.4	69.6	69.7
BiPO	79.4	63.8	84.8	66.0	81.6	78.4	42.4	40.2	96.8	83.4	67.0	56.4	34.4	38.0	69.5	60.9
SVF	**88.8**	**80.2**	85.4	**77.0**	**95.6**	**94.0**	**73.2**	**75.8**	96.8	**92.8**	**75.2**	**61.2**	**81.0**	**79.6**	**85.1**	**80.1**

concept scorers defined on the same space, with target regions $\mathcal{R}_i = \{h : f_i(h) > 0\}$. To steer toward satisfying both attributes, we aim for the intersection $\mathcal{R}_\wedge = \mathcal{R}_1 \cap \mathcal{R}_2$. Since SVF steers via local boundary normals, it suffices to construct a differentiable composite score $g(h)$ whose positive set matches $\mathcal{R}_\wedge$. A natural geometric construction is $g(h) = \min\{f_1(h), f_2(h)\}$, since $g(h) > 0$ iff $f_1(h) > 0$ and $f_2(h) > 0$. Because $\min$ is non-differentiable at ties, we use its smooth relaxation via *softmin*:

$$g_\tau(h) \; = \; -\tau \log\Big(e^{-f_1(h)/\tau} + e^{-f_2(h)/\tau}\Big), \qquad (4)$$

where $\tau > 0$ controls sharpness and $g_\tau \to \min\{f_1, f_2\}$ as $\tau \to 0$. Crucially, $\nabla g_\tau(h)$ becomes a weighted combination of $\nabla f_1(h)$ and $\nabla f_2(h)$, which automatically prioritizes the attribute that is currently harder to satisfy. We then steer by replacing the single-concept score $f(\cdot)$ with $g_\tau(\cdot)$, yielding one vector field that encodes the joint objective. This construction extends directly to m attributes by applying softmin over $\{f_i(h)\}_{i=1}^m$. Details are in Appendix A.

4. Experiments

4.1. Setup

We describe our main experimental setup in this section. More details can be found in Appendix B.

Models We evaluate on Llama-2-7b-Chat-hf (Touvron et al., 2023) and Qwen3-14b (Yang et al., 2025) to cover different model families and scales. Ablations and analyses are run on Llama-2-7b-Chat-hf.

Data We focus on steering for AI persona and truthfulness/hallucination. For persona control, we use Model-Written Evaluations (MWE) (Perez et al., 2022), which covers diverse persona categories with 1,000 examples per category and is formatted as multiple-choice questions (MCQ). In the generation setting, we remove the options from the prompt and provide only the question as input. For truthfulness and hallucination, we use TruthfulQA (Lin et al., 2022) and the hallucination dataset by Panickssery et al. (2024).

Baselines We compare SVF against representative inference-time steering methods and closely related approaches. Our baselines include Contrastive Activation Addition (CAA) (Panickssery et al., 2024), In-Context Vectors (**ICV**) (Liu et al., 2024a), which emulates in-context learning in activation space, **BiPO** (Cao et al., 2024) that learn steering directions from preference optimization, and a representation finetuning method Representation Editing (**RED**) (Wu et al., 2024a). For CAA, we report both a standard single-layer variant **CAA(s)** and a multi-layer variant **CAA(m)**. **CAA(m)** applies interventions at the same set of layers as SVF by injecting each layer's own CAA vector at that layer. To ensure fairness, we report results using validation-tuned hyperparameters, otherwise we follow the best settings reported in the original work.

Evaluation To assess steering for MCQ, we report accuracy (acc.) and a steerable rate (steer.). For each example, let $g(x) = \ell_{\text{gold}}(x) - \ell_{\text{other}}(x)$ be the logit gap between gold and opposite options, and $\Delta g(x) = g_{\text{steer}}(x) - g_{\text{base}}(x)$. We define steer. $= \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[\Delta g(x_i) > 0]$. Unlike acc., which can hide per-instance failures, steer. directly captures the reliability of steering by quantifying the fraction of inputs that are pushed in the desired direction.

Table 2. Scores for generation with Llama-2-7b-Chat-hf. Best results are in **bold**, and runner-up results are underlined. (+) denotes eliciting hallucination/untruthfulness, while (-) denotes reducing them. For computing the overall score, (-) scores are converted as $5 - \text{score}$. SVF achieves the best overall performance, slightly outperforming the preference-training baseline BiPO and substantially exceeding the remaining baselines. Results with Qwen3-14b and additional generation examples are reported in Appendix D.

	wealth	power	myopic	survival	corrigible	hallu(+) ↑	hallu(-) ↓	TQA(+) ↑	TQA(-) ↓	Overall
base	1.58	1.42	2.74	1.72	2.30	2.50	2.50	2.40	2.40	2.20
CAA(s)	1.66	1.56	2.82	1.76	2.34	2.44	2.52	2.38	2.33	2.23
CAA(m)	1.72	1.58	2.88	1.96	2.38	2.48	2.42	2.46	2.37	2.30
ICV	1.84	1.72	2.86	2.30	2.38	2.56	2.12	2.39	2.18	2.42
RED	1.98	1.92	3.10	2.24	2.55	2.02	1.96	2.56	2.28	2.46
BiPO	**2.46**	2.02	3.04	2.42	**2.80**	**3.40**	**1.54**	**2.92**	2.12	2.82
SVF	2.26	**2.36**	**3.30**	**2.68**	2.64	3.38	1.76	2.84	**1.96**	**2.86**

Table 3. Multi-Attribute steering results (acc.) on MCQ tasks with Llama2-7b-Chat-hf. SVF provides more reliable and consistent control. Results on generation tasks are offered in Appendix D.

Synergy components		CAA	RED	BiPO	SVF
wealth +int.-sci	wealth	53.2	74.8	53.4	**77.2**
	int.-sci	50.2	**75.6**	49.8	74.2
	avg	51.8	75.2	51.6	**75.8**
survival +myopic	survival	57.2	67.5	65.4	**71.8**
	myopic	82.6	88.6	**92.6**	89.2
	avg	70.0	78.2	79.0	**80.6**
int.-sci +survival +myopic	int.-sci	68.4	**91.4**	54.0	83.6
	survival	62.8	57.8	65.2	**72.0**
	myopic	78.2	84.6	**87.8**	79.6
	avg	69.8	78.0	69.0	**78.4**

For open-ended generation, we adopt an LLM-as-a-judge protocol following Cao et al. (2024). Specifically, we use GPT-4.1 to rate each response on a 1–4 scale. The evaluation prompt is provided in Appendix C.

4.2. Main Results

Multiple-Choice Questions In Table 1, SVF consistently improves both the accuracy and steerable rate, indicating more reliable movement toward the intended option across inputs. Notably, SVF delivers significant gains on concepts that prior analyses identify as particularly unreliable under existing steering methods (Tan et al., 2025). For example, on Llama-2-7b-Chat-hf, Narcissism was reported as unsteerable, yet SVF raises its steerable rate from around 50% to 87.6%. Likewise, Myopic was shown to exhibit substantial anti-steerable behavior, while SVF increases its steerable rate to 96.8%. These results support our geometric view that effective steering directions are context-dependent and cannot always be captured by a single global vector.

Open-Ended Generation Table 2 reports scores for open-ended generation. While BiPO is competitive in this setting, it comes with utility and generalization concerns that we will analyze in §5. Compared to static steering baselines, SVF is consistently stronger on long-form generation. This supports our design choice of refreshing steering directions from the evolving hidden states, allowing the intervention to track locally effective concept geometry throughout decoding. We also observe that, aside from BiPO, most baselines struggle to reliably elicit hallucination and untruthfulness, consistent with prior findings (Cao et al., 2024). Notably, although truthfulness-related behaviors can be steerable in short-answer MCQ settings (Li et al., 2024), the long-form failures suggest a gap between capturing a concept direction at a fixed state and repeatedly applying a static direction along a drifting decoding trajectory. Overall, these results indicate that robust long-form control for certain complex behaviors likely requires either trained steering operators or context-dependent direction updates, rather than a single global vector applied throughout generation.

Multi-Attribute Steering Table 3 reports multi-attribute steering results on MCQ tasks. To ensure broad coverage, we evaluate diverse concept compositions instead of focusing only on strongly semantically related pairs commonly used in prior work (e.g., wealth/power- seeking). SVF preserves strong effects for each concept within a composition and achieves the best average performance across all three compositions. In contrast, we find that steering baselines including CAA and BiPO can break down on compositions with potential conflict such as WEALTH+INTEREST-IN-SCI. We also observe concept imbalance for RED and BiPO in more complex compositions. Even when some attributes improve, others can remain barely moved. For example, in INTEREST-IN-SCI+SURVIVAL+MYOPIC, BiPO boosts SURVIVAL and MYOPIC but leaves INTEREST-IN-SCI underperformed, and RED shows a similar gap on SURVIVAL in the same composition.

4.3. Ablation Study

In this section, we present ablations on the multi-layer composition design and key components of SVF. Additional ablation studies are deferred to Appendix E.

Multi-Layer Composition Table 4 studies SVF's multi-