Controlling Foundation Models via Internal Interventions: A Survey of Inference-Time Methods

Anonymous authors
Paper under double-blind review

Abstract

Large generative foundation models exhibit broad but often implicit behaviors that researchers have begun to steer by intervening in model computation at inference time. Rather than updating model parameters or relying solely on explicit prompts, inference-time interventions modify hidden activations, sparse features, attention and routing patterns, or token-selection scores during the forward pass and decoding. These methods offer a potentially flexible, request-time framework for post-training control, but they differ in required model access, computational overhead, robustness, and side effects. This survey reviews such methods under a common intervention lens. We categorize the literature into representation-level, attention-level, and decoding-level interventions, and analyze them across large language models, vision-language models, and diffusion models. We make the scope boundary explicit by separating core fixed-weight, inference-only methods from adjacent paradigms such as prompting, retrieval, fine-tuning, weight editing, and newly trained controllers. To support comparison, we describe transformer, multimodal, and diffusion generation using computational-graph notation; summarize reported performance deltas, access requirements, and risk profiles; and connect empirical steering methods to evidence for approximately linear representations and sparse feature decompositions. We close by outlining practical gaps: how to compare steering methods on shared benchmarks, decide where an intervention should be applied, measure unintended side effects, and combine multiple intervention types without interference.

1 Introduction

Generative foundation models, including large language models (LLMs) and vision-language models (VLMs), increasingly serve as core components in modern AI systems. Web-scale pretraining and scaling laws have enabled strong few-shot and in-context capabilities (Brown et al., 2020), while instruction tuning and reinforcement learning from human feedback (RLHF) have improved alignment and usability (Ouyang et al., 2022). In parallel, advances in multimodal pretraining have produced vision–language models spanning transferable visual recognition, few-shot multimodal reasoning, and instruction-following visual dialogue (Radford et al., 2021; Alayrac et al., 2022; Liu et al., 2023).

Despite significant advances, reliably controlling model behavior remains challenging: models may hallucinate, violate safety constraints, or deviate from user intent. A common solution, weight-level adaptation via fine-tuning or RLHF, can be effective but may be expensive, slow to iterate, and brittle under distribution shift. These limitations motivate the search for lightweight, flexible control mechanisms that can be applied after deployment without retraining or modifying the full set of model parameters.

Early work on controllable generation explored explicit conditioning signals, such as control codes (Keskar et al., 2019), and "plug-and-play" methods that steer hidden states via auxiliary attribute models (Dathathri et al., 2020). Recently, prompting and prompt engineering have become a common control interface for LLMs, enabling task adaptation through natural-language instructions and few-shot demonstrations (Brown et al., 2020), and have shown widespread adoption. However, prompting is fundamentally an input-space control mechanism and can fail when instructions are ignored, difficult to express, or bypassed under adversarial

or distribution-shifted inputs. Parameter-efficient alternatives such as prefix tuning (Li & Liang, 2021) and low-rank adapters (Hu et al., 2022) reduce training cost but still require optimization and are not designed for dynamic per-request behavioral control.

These limitations have motivated a line of work that intervenes directly on model computations during inference. From a causal perspective, transformer models define directed computational graphs where information flows through embeddings, attention modules, feedforward blocks, and residual connections before influencing final token logits. Inference-time interventions operate by targeting specific intermediate variables within this graph, modifying internal nodes, routing patterns, or output scores while holding parameters fixed. Because these mechanisms operate below the prompt surface, they can sometimes influence computational pathways that shape generation and provide a request-time control interface at deployment time.

We use *inference-time intervention* to refer to methods that change request-time computation while keeping the base model weights fixed. The strict core of the survey excludes methods that train a new steering controller, adapter, classifier, reward model, or probe, and treats prompting, retrieval-only systems, fine-tuning, RLHF-style alignment, and permanent weight editing as adjacent paradigms rather than core steering methods. Table 1 makes this scope rule explicit. It distinguishes direct hidden-state, attention/KV, and decoding interventions from methods whose control signal enters through prompts, retrieved context, or parameter updates. Boundary and historical methods are still included when they clarify the mechanism, but they are marked or discussed separately rather than counted as strict-core inference-time interventions.

Taxonomy and failure-mode perspective. We organize methods by intervention locus rather than by application domain. Table 2 maps the resulting literature across representation-level, attention-level, and decoding-level mechanisms for LLMs, VLMs, and diffusion models. This separation is useful because the same failure mode can often be addressed at multiple levels: a hallucination may be reduced by shifting hidden states, reallocating attention to visual evidence, or contrasting logits during decoding. Conversely, methods with different names may share the same computational operation, such as adding a residual-stream direction, suppressing an attention sink, or reshaping token scores. We group methods into three broad classes:

- **Representation-level interventions** directly edit internal activations, such as residual streams, neurons, multimodal hidden states, denoising feature maps, or learned sparse features. They are most natural when a target behavior appears to be encoded in internal representations, but they can introduce side effects when the edited direction is polysemantic.
- **Attention-level interventions** modify information routing over tokens, regions, heads, or KV-cache entries. They are most useful when the needed evidence is already present but weakly used, as in instruction underuse, visual grounding failures, or long-context attention-sink pathologies.
- **Decoding-level interventions** reshape token scores, sequence scores, or output-side guidance during generation. They can provide a last-mile control layer for contrastive scoring, constrained sampling, verifier-style signals, and multimodal consistency checks without changing base model weights.

This survey makes three contributions:

- We provide a unified taxonomy of inference-time interventions and explicit inclusion/exclusion criteria that separate fixed-weight steering from prompting, retrieval, fine-tuning, weight editing, and trained-controller boundary cases.
- We use a computational-graph view of LLM, VLM, and diffusion generation to give a common language for representation-level, attention-level, and decoding-level control across modalities.
- We synthesize representative methods through reported performance deltas, access requirements, overhead, side effects, reproducibility concerns, and cross-level composability, identifying open problems in standardized evaluation, robustness, causal localization, and deployment risk.

Method type	Include?	Reason	Examples
Activation addition / dense steering	Yes	Inference-time hidden-state intervention with frozen base weights.	ActAdd, CAA
Representation subspace steering	Yes	Residual-stream or subspace intervention applied during the forward pass.	RepE, truth directions
SAE feature steering	Yes	SAE may be trained separately, but the base model is frozen during steering.	LF-Steer, SAE-TS
Attention reweighting	Yes	Inference-time routing intervention over tokens or visual regions.	PASTA, SpotLight
Attention/KV redistribution and serving enablers	Conditional	Included when they directly redistribute attention or KV use at inference; serving-only analyses are deployment-adjacent context.	ACT, VAR; StreamingLLM (adjacent)
Contrastive decoding	Yes	Token-score intervention during generation.	DoLa, VCD, ICD
Frozen external tool or pretrained scorer	Conditional	Included only when reused without steering-specific training and applied at inference time.	CLIP-style scores
Verifier/reranker decoding with newly trained scorer	Boundary	Discussed for contrast, but excluded from the strict core when the scorer/controller is newly trained for steering.	FUDGE, learned rewards
Full fine-tuning / RLHF	No	Changes base model weights or reward-trained policy behavior.	SFT, RLHF
Parameter-efficient tuning	Usually no	Adds or changes trained parameters for behavior control.	LoRA, ReFT
Prompt-only control	No	Input-space control rather than internal intervention.	prompt engineering
Retrieval-only systems	Usually no	Adds context but does not directly modify computation or decoding.	standard RAG
Permanent weight editing	No	Alters model parameters rather than request-time computation.	model editing

Table 1: **Inclusion and exclusion criteria.** The main survey boundary is inference-time control with fixed base model weights.

2 Survey Methodology

2.1 Inclusion and Exclusion Criteria

We include methods whose primary control mechanism is applied during inference and whose base model weights remain unchanged. In the strict core taxonomy, auxiliary components must be frozen or pre-existing tools; methods that train a new controller, adapter, verifier, reward model, or probe for steering are treated as boundary cases rather than core inference-time steering methods. Table 1 summarizes these boundary decisions as a decision rule: the relevant question is not only whether a method improves behavior, but where the control signal enters the system and whether new trainable components are introduced for the steering task.

Table 2 should be read together with this rule. It is intentionally broader than the strict-core set because a useful survey map should show neighboring methods rather than hide them. The table therefore includes core inference-time interventions, historical precursors, and boundary methods in the same mechanism families, while the dagger notation and surrounding discussion indicate which entries depend on trained auxiliary components or other scope-edge assumptions. This separation lets us compare mechanisms without treating every listed method as equally deployable under the fixed-weight, inference-only criterion.

2.2 Paper Screening and Coding

Candidate papers were screened first by title and abstract, then by full text when they appeared to satisfy the inference-time boundary. For each included method, we coded intervention level, modality, intervention locus, access requirement, auxiliary components, whether labels or gradients are required, reported benchmark and metric, and stated limitations. When a method spans multiple levels, we code the primary deployment lever and note the hybrid mechanism in the method discussion. Our literature snapshot includes papers

Intervention Approach	Taxonomy	Related Work
Representation Level Intervention	Activation Steering	ActAdd (Turner et al., 2023), CAA (Rimsky et al., 2024), Mean-center activation steering (Jorgensen et al., 2023), ICV (Liu et al., 2024b), CoT compression steering (Azizi et al., 2025), Activation Scaling (Stoehr et al., 2024), VTI (Liu et al., 2025), Textual Steering Vectors (Gan et al., 2025), CASteer (Gaintseva et al., 2025), ActErase (Sun et al., 2026), CountSteer (Boo et al., 2025), Activation Transport (AcT) (Rodriguez et al., 2025), RepE (Zou et al., 2023), Refusal Direction (Arditi et al., 2024), DAC (Scalena et al., 2024), ASD (Su et al., 2025), SADI (Wang et al., 2024b), CAST (Lee et al., 2024), Instruction-Following Activation Steering (Stolfo et al., 2024), GrAInS (Nguyen et al., 2025), ASTRA (Wang et al., 2024a), Spherical Steering (You et al., 2026), Angular Steering (Vu & Nguyen, 2025), AUSteer (Feng et al., 2026b), Persona Vectors (Feng et al., 2026a)
	SAE Steering	LF-Steering (Yang et al., 2025), FGAA (Soo et al., 2025), CorrSteer (Cho et al., 2025), SAE-SSV[‡] (He et al., 2025a), SAIF (He et al., 2025b), Monosemanticity (Bricken et al., 2023), Scaling Monosemanticity (Templeton et al., 2024), Interpretable SAE Features (Cunningham et al., 2024), SAE-TS (Chalnev et al., 2024), Refusal SAE Steering (O'Brien et al., 2024), SpARE (Zhao et al., 2024)
Attention Level Intervention	Token Reweighting	PASTA (Zhang et al., 2024), InstABoost (Guardieiro et al., 2025), CAAC (Fazli et al., 2025), CARVE (Ge et al., 2025), ICT (Chen et al., 2025a), CAI (Li et al., 2025b), Attend-and-Excite (Chefer et al., 2023), Prompt-to-Prompt (Hertz et al., 2023)
	Dynamic Control	SpotLight (Venkateswaran & Contractor, 2026), COMPASS[‡] (Sahay et al., 2025), LLMSTEER (Gu et al., 2024), SPARC (Jung et al., 2025), Qwen Look Again[‡] (Chu et al., 2025), CG-VLM (Bu et al., 2025), SPIN (Sarkar et al., 2025), VEGAS (Wang et al., 2025)
	Redistribution/Sinks	ACT (Yu et al., 2024), VAR (Kang et al., 2025), EVAS (Zhang et al., 2025b), StreamingLLM (deployment-adjacent) (Xiao et al., 2024), First-token sink analysis (Barbero et al., 2025), Sink analysis (Gu et al., 2025)
	Activation inside Attention	DISCO (Torop et al., 2025), ITI (Li et al., 2023a)
	Module Localization / Enablers	REAL (Zhan et al., 2025), SHIPS/SAHARA (Zhou et al., 2025), Head importance analyses (Michel et al., 2019; Voita et al., 2019)
Decoding-Level Intervention	Decoding Objectives / Guided Boundary	SimCTG (Su et al., 2022), EGED (Das et al., 2025), DoLa (Chuang et al., 2024), FUDGE (Yang & Klein, 2021), GeDi (Krause et al., 2021), DExpert (Liu et al., 2021), DeAL[‡] (Huang et al., 2025), SEGA (Brack et al., 2023), FK Diffusion Steering (Singhal et al., 2025)
	Logit Composition / Acceleration Background	Contrastive Decoding (Li et al., 2023b), ThinkLogit[‡] (Zhang et al., 2025c), Contrastive decoding improves reasoning (O'Brien & Lewis, 2023), Proxy Tuning (Liu et al., 2024a), Self-Speculative Decoding (deployment-adjacent) (Zhang et al., 2023), SpecInfer (deployment-adjacent) (Miao et al., 2024)
	Multimodal Contrast	VCD (Leng et al., 2024), ICD (Wang et al., 2024c), ConVis (Park et al., 2025), AvisC (Woo et al., 2025), RITUAL (Woo et al., 2024), M3ID (Favero et al., 2024), RVCD (Lee & Song, 2025), Fuzzy Contrastive Decoding (Kim et al., 2025)
	Attention-Aware Decoding	OPERA (Huang et al., 2024), IMCCD (Li et al., 2025a), FarSight (Tang et al., 2025), Self-Introspective Decoding (Huo et al., 2024)

Note. The table intentionally includes core inference-time methods together with boundary and historical methods for context. Methods requiring newly trained controllers, probes, reward models, adapters, hypernetworks, or architecture changes are not counted as strict-core methods and are marked or discussed as boundary cases.

Table 2: **Taxonomy of steering methods** for foundation models, organized by intervention family and mechanism.

and preprints available through May 2026; very recent 2026 entries are marked as future-facing when their implementation details or venue metadata remain unstable.

2.3 Evaluation, Tooling, and Reproducibility

We separately code evaluation and tooling papers that are not themselves steering methods but affect how intervention results should be interpreted and reproduced. AxBench (Wu et al., 2025b) benchmarks concept

detection and steering methods and cautions that sparse-autoencoder methods are not automatically stronger than simple baselines; complementary SAE feature-selection work (Arad et al., 2025) shows that conclusions about SAE steering can change substantially depending on how features are chosen. Reliability analyses of steering vectors (Tan et al., 2024) highlight out-of-distribution brittleness, spurious contrast capture, and high variance across prompts, motivating side-effect and robustness reporting alongside target metrics. For tooling, NNsight/NDIF (Fiotto-Kaufman et al., 2025) provides infrastructure for hidden-state access and remote intervention experiments, while best-practice work on activation patching (Zhang & Nanda, 2024) emphasizes that causal localization depends on corruption choices, metrics, and patch granularity. We use circuit-tracing and attribution-graph work (Anthropic, 2025) as background for the computational-graph framing, and AI Steerability 360 (Miehling et al., 2026) to motivate standardized reporting fields spanning input, structural, state, and output control surfaces.

2.4 Threats to Validity in Paper Selection

The area is moving quickly, so the survey is vulnerable to recency bias and incomplete coverage of workshop, arXiv-only, or non-English papers. Keyword search may miss papers that use adjacent terminology such as controllable generation, model editing, causal tracing, grounding, or calibration. Cross-paper quantitative comparisons are also limited by inconsistent base models, benchmarks, prompt templates, and reporting practices; we therefore treat published numbers as indicative rather than definitive rankings.

3 Formal Setup: Foundation Models as Intervention Graphs

3.1 Transformer Computational Graph

To ground the interventions discussed in this survey, we formalize Transformer-based generation (Vaswani et al., 2017) as a directed computational graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where nodes $\mathcal{V}$ represent intermediate computations and edges $\mathcal{E}$ represent data flow (Figure 1).

Given an input sequence of T tokens, the transformer first maps each token to an embedding $x_t \in \mathbb{R}^d$. At each layer $l \in \{1, \ldots, L\}$, the computation proceeds as follows:

Multi-head self-attention (MHSA). For each head $h \in \{1, \ldots, H\}$, queries, keys, and values are computed using a row-vector convention as $Q^h = X_l W_Q^h$, $K^h = X_l W_K^h$, $V^h = X_l W_V^h$, and the attention output is:

$$\mathrm{Attn}^h(X_l) = \mathrm{softmax}\left(\frac{Q^h (K^h)^\top}{\sqrt{d_h}}\right) V^h. \tag{1}$$

The concatenated head outputs are projected and added to the residual stream: $X'_l = X_l + [\mathrm{Attn}^1; \ldots; \mathrm{Attn}^H] W_O$.

Feed-forward network (FFN). The post-attention representation is transformed by a position-wise MLP: $X_{l+1} = X'_l + \mathrm{FFN}(X'_l)$, where $\mathrm{FFN}(x) = W_2 \cdot \sigma(W_1 \cdot x + b_1) + b_2$ with nonlinearity σ (typically GELU or SwiGLU).

Residual stream. The residual stream $h_{l,t}$ at layer l and position t accumulates contributions from all preceding attention and FFN blocks. This additive structure means that a perturbation $\Delta_{l,t}$ injected at any layer propagates forward through subsequent computations, providing the mechanistic basis for activation steering (Eq. 5).

3.2 Multimodal Fusion Graph

For VLMs, the input graph contains both visual and textual streams. An image encoder produces visual features v_i, a projector maps them into the language-model token space $\tilde{v}_i = P_\phi(v_i)$, and the multimodal transformer consumes the concatenated sequence $X_0 = [\tilde{v}_1, \ldots, \tilde{v}_M, x_1, \ldots, x_T]$. Interventions may be applied before fusion (vision features or projected visual tokens), during fusion (cross-modal attention or shared

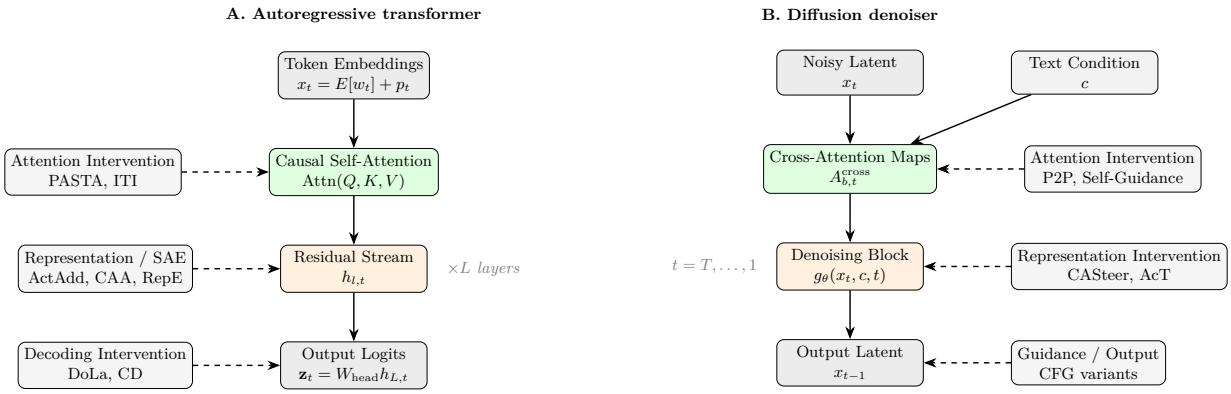

Figure 1: **Intervention loci for autoregressive transformers and diffusion denoisers.** Dashed arrows indicate where each intervention family modifies inference-time computation. In autoregressive transformers, interventions target attention, residual-stream representations, or logits. In diffusion models, interventions target cross-attention maps, denoising representations, or output/guidance steps. VLM fusion is described textually in Section 3.2.

residual stream), or at the output logits. This view separates hallucination mechanisms caused by weak visual evidence from those caused by downstream language-prior decoding.

3.3 Diffusion Denoising Graph

For diffusion models, generation proceeds through an iterative denoising process:

$$x_{t-1} = g_\theta(x_t, c, t), \tag{2}$$

where x_t is the noisy latent at timestep t, c is the conditioning signal, and g_θ is the denoising network. An inference-time activation intervention modifies an intermediate denoising representation:

$$h'_{b,t} = h_{b,t} + \Delta_{b,t}, \tag{3}$$

where $h_{b,t}$ is the activation at block b and timestep t. Attention-level diffusion interventions may instead modify cross-attention logits or maps between text tokens and spatial latent locations:

$$\tilde{S}^{\mathrm{cross}}_{b,t} = S^{\mathrm{cross}}_{b,t} + \Gamma_{b,t}, \qquad \tilde{A}^{\mathrm{cross}}_{b,t} = \mathrm{softmax}(\tilde{S}^{\mathrm{cross}}_{b,t}). \tag{4}$$

Intervention loci. Each node in $\mathcal{G}$ is a potential intervention point. Representation-level methods target residual stream activations, FFN outputs, multimodal hidden states, or denoising feature maps. Attention-level methods modify Q, K, V projections, attention logits, attention maps, or sink routing. Decoding-level methods modify final logits, sequence scores, or output-side guidance before sampling.

4 Background and Theoretical Grounding

4.1 Why Not Just Fine-Tune? Why Not Just Prompt?

Controlling foundation model behavior can be approached through three paradigms: (1) weight-level adaptation (fine-tuning, RLHF, LoRA), (2) input-level conditioning (prompting, in-context learning), and (3) internal intervention (the focus of this survey). Each has distinct strengths and failure modes.

Fine-tuning modifies model parameters θ to optimize a control objective. This is powerful but requires training data, compute, and creates a new model checkpoint per behavior. Fine-tuning is slow to iterate, cannot be applied dynamically per-request, and is brittle under distribution shift, since a model fine-tuned

for safety on one distribution may fail on adversarial inputs (Qi et al., 2024). Parameter-efficient methods (LoRA (Hu et al., 2022), prefix tuning (Li & Liang, 2021)) reduce cost but still require optimization and produce static behavioral changes.

Prompting conditions generation through natural-language instructions prepended to the input. It is flexible and requires no training, but operates entirely in the input space. Prompt-engineering surveys document that prompt design remains brittle and task-dependent, with failures arising when instructions are underspecified, poorly structured, or difficult to express purely as natural-language conditioning (Sahoo et al., 2024).

Internal intervention (this survey's scope) modifies intermediate computations during the forward pass while keeping θ fixed. This can offer several practical advantages: (a) *reversibility*: the intervention can be toggled on/off per request; (b) *granularity*: the intervention can target specific layers, heads, or feature dimensions; (c) *composability*: some interventions can be combined, although interactions are not guaranteed to be additive; and (d) *reduced reliance on linguistic mediation*: the control signal operates below the token level, reducing dependence on prompt wording while not eliminating adversarial bypass risks. The key trade-off is interpretability and reliability: interventions on polysemantic representations can have unpredictable side effects.

4.2 Theoretical Grounding: Why Does Linear Steering Work?

A central empirical observation underlying representation steering is that many behavioral and semantic properties are encoded as approximately linear directions in transformer hidden spaces. This observation has been formalized along several lines:

The Linear Representation Hypothesis. Park et al. (2024) study conditions under which high-level concepts can be approximately linearly represented in LLM hidden states, including formal analysis and examples such as language identity and analogy-style semantic structure. Their analysis supports linear projections as a useful intervention model in studied settings, while not implying that arbitrary concepts form causally clean subspaces; this motivates additive steering (Eq. 5) as an empirical approximation rather than a guarantee.

The Geometry of Truth. Marks & Tegmark (2024) show that true/false distinctions in LLM representations exhibit emergent linear structure: truth values of factual statements are linearly separable in activation space across multiple layers. This finding directly motivates methods like ITI (Li et al., 2023a) that shift activations along "truthfulness directions."

Representation Engineering. Zou et al. (2023) formalize Representation Engineering (RepE) as a top-down approach to AI transparency. RepE extracts control-relevant directions via PCA on contrasting stimulus pairs and reports that these directions enable reading (probing) and writing (steering) of high-level concepts including honesty, fairness, and harmlessness. These results suggest that, in studied settings, the residual stream can behave like a roughly linear "concept space" where some behavioral attributes are separable.

Monosemanticity and sparse features. Sparse-autoencoder work shows that dense activations can often be decomposed into sparse, more interpretable feature dictionaries (Bricken et al., 2023; Templeton et al., 2024; Cunningham et al., 2024). This provides one mechanistic lens for understanding why some dense steering directions can work and why SAE steering may reduce polysemantic interference, but feature selection and causal isolation remain nontrivial. When selected SAE features are causally aligned with a target behavior, steering in SAE space (Eq. 6) can more directly manipulate that behavior-relevant activation.

The success of many activation-steering methods is consistent with empirical and theoretical evidence that some model behaviors are approximately linearly represented, although this approximation is behavior-, layer-, and model-dependent. RepE also notes that more intricate concepts or multi-step functions may

require more involved representation-extraction procedures (Zou et al., 2023), motivating nonlinear and input-dependent steering methods.

5 Representation-Level Interventions

5.1 Activation Steering

Activation (representation-level) steering refers to inference-time methods that control model behavior by directly intervening in intermediate activations while keeping model parameters fixed. Although first explored in LLMs, the same idea can extend to vision-language models (VLMs): modern VLMs fuse projected image tokens and text tokens inside a shared transformer backbone, so related intervention machinery can operate on unimodal or multimodal hidden states, including cross-modal representations.

A minimal formalization. Let f_θ be a pretrained transformer (uni- or multi-modal) with parameters θ, and let $h_{l,t} \in \mathbb{R}^d$ denote the hidden activation at layer l and position t (text token or visual token). Activation steering applies an inference-time perturbation as follows:

$$h'_{l,t} = h_{l,t} + \Delta_{l,t}, \tag{5}$$

where $\Delta_{l,t}$ may be (i) a fixed additive shift (a "steering vector"), (ii) a multiplicative rescaling or mask, (iii) an input-conditioned update predicted from context, or (iv) a localized patch/transport map. This encapsulates the core mechanism in a single equation while emphasizing the key constraint: θ does not change.

Figure 2 illustrates the common contrastive activation-steering workflow used by methods such as ActAdd, CAA, and related representation-engineering approaches. A direction is first estimated from matched positive and negative examples, then reused at inference time as an additive hidden-state intervention. The sign, scale, layer, and token positions of this vector are key implementation choices.

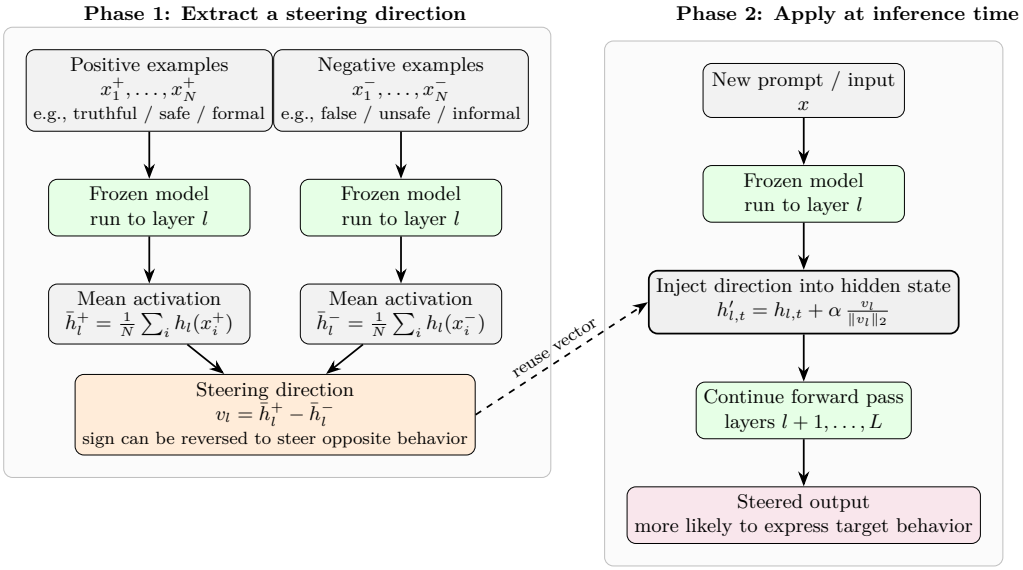

Figure 2: **Contrastive activation-steering pipeline.** A steering direction is extracted from matched positive/negative examples, normalized, and injected into the hidden state at inference time. Reversing the sign of v_l steers toward the opposite behavior.

Residual Stream / Layer Block. The most common steering target is the residual stream (or layer/block outputs), where additive shifts are often effective and straightforward to implement. Turner et al. (Turner

et al., 2023) introduce Activation Addition (ActAdd), which computes a steering direction from contrasting prompts (e.g., positive vs. negative sentiment) and applies it during decoding to shift sentiment, topic, and toxicity across multiple model families. By subtracting two runs that differ primarily in the attribute of interest, the resulting difference vector aims to isolate "how the model is behaving" along the targeted dimension from "what the prompt is about." Dataset-level variants stabilize this idea by averaging over many exemplars: Contrastive Activation Addition (CAA) computes mean activation differences over positive/negative datasets to steer Llama models (Rimsky et al., 2024), while mean-centering variants improve robustness by subtracting behavior-agnostic baselines to reduce spurious shifts (Jorgensen et al., 2023). A closely related perspective treats in-context learning itself as an implicit latent-space shift and makes that shift explicit: In-context Vectors (ICV) extract a single task vector from demonstration examples (e.g., via principal directions of demo-pair activation differences) and then add it across layers/tokens at inference, effectively "compressing" demonstrations into an activation-level steer whose strength can be tuned (Liu et al., 2024b). Residual-stream steering has also been used to modulate reasoning traces: Azizi et al. (2025) construct directions from paired verbose reasoning rationales and inject them during decoding to compress chain-of-thought (CoT) while constraining distributional drift. Beyond additive translations, multiplicative interventions rescale existing features rather than injecting new ones; Stoehr et al. (2024) show that scaling-based steering can provide comparable control with fewer intervention parameters in studied settings. This connects to mechanistic interpretability by using localized heads or features as candidates for targeted intervention analysis.

Representation Engineering and single-direction methods. Representation Engineering (RepE) (Zou et al., 2023) formalizes a top-down approach to AI transparency: it extracts control-relevant directions via PCA on contrasting stimulus pairs (e.g., honest vs. dishonest completions) and uses them for both reading (probing model beliefs) and writing (steering behavior). RepE reports evidence that high-level concepts such as honesty, fairness, harmlessness, and power-aversion can correspond to approximately linear directions in the residual stream of studied model families. A prominent special case is the refusal direction: Arditi et al. (2024) show that a residual-stream direction plays a large role in refusal behavior for several chat-tuned models. Ablating this direction can substantially reduce refusal without equally large measured capability loss in their setting, illustrating both the usefulness and risk of geometric control; the same class of mechanism that enables safety steering can also enable safety bypass. Dynamic Activation Composition (DAC) (Scalena et al., 2024) extends single-property steering to the multi-property setting using an information-theoretic approach that dynamically modulates steering intensity for one or more properties throughout generation.

Dynamic and condition-dependent activation steering. Recent work directly targets the brittleness of static global directions. SADI (Wang et al., 2024b) constructs semantics-adaptive steering vectors at inference time, making the intervention depend on input semantics rather than applying a fixed direction uniformly. CAST (Lee et al., 2024) applies conditional activation steering to refusal behavior, narrowing side effects by activating the refusal steer only under specified input conditions; this reduces indiscriminate behavioral shifts but makes robust condition detection part of the safety-critical mechanism. Instruction-following activation steering (Stolfo et al., 2024) computes hidden-state differences between prompts with and without target instructions, showing that format, length, and lexical constraints can be influenced through representations; related instruction-following failures can also be targeted by attention reweighting methods such as PASTA (Zhang et al., 2024) and SpotLight (Venkateswaran & Contractor, 2026). GrAInS (Nguyen et al., 2025) uses gradient-based attribution to select influential hidden-state components for inference-time steering, providing an input-dependent alternative to whole-vector addition and extending the critique of fixed global steering to both LLMs and VLMs.

Geometry- and style-aware steering. Another recent direction constrains how steering vectors move activations rather than only where they are added. Spherical steering (You et al., 2026) and angular steering (Vu & Nguyen, 2025) are recent geometry-aware proposals for rotation- or angle-aware activation updates intended to reduce collateral distribution shift. AUSteer (Feng et al., 2026b) frames fine-grained atomic-unit control as adaptive-strength steering, while PERSONA (Feng et al., 2026a) studies dynamic and compositional personality control via activation-vector algebra. Because several of these entries are very recent

relative to the survey cutoff, we treat them as future-facing core candidates only when the full method remains free of newly trained controllers or probes.

VLM-specific. In VLMs, interventions can occur before fusion (vision encoder features / post-projector vision tokens) or after fusion (shared multimodal residual stream). VTI (Liu et al., 2025) reduces hallucinations by injecting stabilizing latent shifts at inference time, approximating the effect of averaging vision features under mild perturbations. Boundary methods such as SteerVLM (Sivakumar et al., 2025) keep the base VLM frozen but train a lightweight add-on module to produce token-wise activation updates; we report them for context but do not count them in the strict no-partly-trained core. Text-derived steering directions can also transfer into multimodal models: Gan et al. (2025) extract directions from the text-only backbone and inject them into multimodal hidden states to improve visual understanding (e.g., spatial relations and counting), suggesting that some "visual" concepts are linearly addressable in shared representation spaces. For VLM safety, ASTRA (Wang et al., 2024a) adapts activation steering to jailbreak defense without updating base model weights, making visual safety steering an inference-time analogue of LLM refusal-direction work.

Additional VLM activation methods. Learn to Steer (L2S) (Parekh et al., 2025) generalizes static steering to input-dependent control: it trains a lightweight predictor that, given the input, produces per-layer steering vectors conditioned on the specific input context. This addresses the brittleness of fixed directions when applied to diverse inputs, showing improvements in both safety (MMSafetyBench) and hallucination (POPE) for LLaVA-v1.5, but it is outside the strict no-partly-trained core because the predictor is trained. Adaptive Steering for VLM Decoding (ASD) (Su et al., 2025) adopts a bidirectional approach, applying activation shifts to both hidden states and decoder outputs, achieving POPE MSCOCO F1 improvements on LLaVA1.5-7B (86.03 $\rightarrow$ 87.87). ICT (Image-Object Cross-Level Trusted Intervention) (Chen et al., 2025a) is a cross-level VLM grounding intervention; we classify it primarily under attention/token grounding while discussing it here because it also touches multimodal hidden states. DMAS (Yin et al., 2026) and Null-Steer (Zhu et al., 2026) are recent VLM safety/hallucination extensions that we treat as future-facing entries when the literature cutoff includes 2026, pending full-method checks for trained detectors or controllers.

Activation-level interventions in diffusion denoisers. Diffusion models expose rich intermediate representations (U-Net/DiT feature maps and attention maps) that can be intervened on at inference time, supporting reported training-free controls analogous to activation steering in transformers. Diffusion steering can likewise be grouped by intervention locus: (i) block feature maps, (ii) localized activation patching/masking, and (iii) attention-projection edits (e.g., query projections). CASteer (Gaintseva et al., 2025) precomputes concept-specific steering directions from averaged hidden activations and applies them dynamically to cross attention outputs to suppress undesired concepts only when present (or inject desired concepts when absent). CountSteer (Boo et al., 2025) derives directions from contrasting correct vs. incorrect counting activations and injects step-/block-specific shifts with adaptive strength. Both inject steering vectors into cross-attention. ActErase (Sun et al., 2026) performs training-free concept erasure targeting FFN layers and performing activation patching: it contrasts activations for source vs. target prompts, localizes concept-specific regions, and masks/replaces those entries during denoising. Activation Transport (AcT) (Rodriguez et al., 2025) adapts an optimal-transport-inspired activation transport view to language and diffusion control settings, applying transport-style activation edits between source and target distributions rather than generic vector addition. It intervenes broadly at layer outputs – normalization layers and residual-layer outputs of models.

Additional diffusion steering methods. Erasing Concepts from Diffusion Models (ECD) (Gandikota et al., 2023) introduces fine-tuning-based concept erasure and serves as background for diffusion concept control rather than a strict inference-time method. Concept Ablation (Kumari et al., 2023) is also a weight-update concept-erasure method that efficiently modifies a pretrained diffusion model to suppress target concepts; we discuss it as background but do not count it in the strict inference-time core. P2P (Prompt-to-Prompt) (Hertz et al., 2023) edits diffusion outputs by manipulating cross-attention maps between prompt tokens and spatial locations, enabling localized semantic edits without retraining, while Attend-and-Excite (Chefer et al., 2023) strengthens neglected subject-token attention to improve prompt

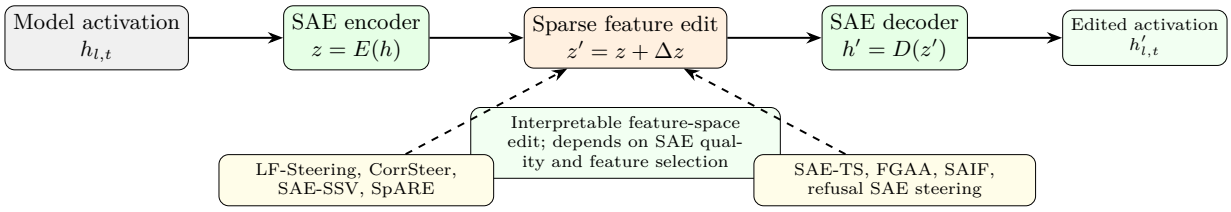

Figure 3: **SAE-based steering pipeline.** SAE-based steering encodes model activations into sparse latent features, edits selected feature coordinates, and decodes the result back into the model activation space. This can improve interpretability and targeting, but introduces dependence on SAE reconstruction and feature-selection quality.

faithfulness. Self-Guidance (Epstein et al., 2023) extracts and manipulates internal attention activations to control spatial layout, appearance, and geometry during generation. BoxDiff (Xie et al., 2023) and FreeControl (Mo et al., 2024) broaden inference-time diffusion control to spatial layouts and external conditions without adding trained networks. GLoCE (Lee et al., 2025) and stochastic self-guidance (Chen et al., 2026) are recent training-free or reported training-free extensions for generation-time concept erasure and guidance, respectively; we mark them as recent entries whose implementation details should be checked before making stronger scope claims.

5.2 Sparse Autoencoder Steering

SAE steering is motivated by the superposition view of neural representations: dense model activations may encode more features than available dimensions, causing individual neurons or directions to be polysemantic. Sparse autoencoders attempt to recover a larger set of sparsely active latent features, some of which are more interpretable and potentially more useful for intervention than individual neurons. This makes SAE steering attractive when the target behavior can be localized to a small set of sparse features, but interpretability does not guarantee causal isolation; semantically meaningful features may still be entangled with unrelated capabilities.

Sparse Autoencoders (SAEs) enable steering in latent feature spaces rather than dense activations. This process involves encoding activations into sparse features, intervening in the latent space, and decoding the result back into the model's original space as follows:

$$z_{l,t} = E(h_{l,t}), \qquad h'_{l,t} = D(z_{l,t} + \Delta z_{l,t})\,. \tag{6}$$

Here, $h_{l,t} \in \mathbb{R}^d$ is the original hidden activation (layer l, token position t), E and D are the SAE encoder/decoder, and $z_{l,t} \in \mathbb{R}^m$ is a sparse latent code. Steering is performed by editing a small subset of latent features via $\Delta z_{l,t}$ (e.g., boosting/suppressing selected coordinates); the edited dense activation $h'_{l,t}$ is then substituted for $h_{l,t}$ during inference while keeping the base model weights fixed. When reconstruction is accurate, this is approximately equivalent to adding a structured perturbation in model space, i.e., $h'_{l,t} - h_{l,t} \approx D(\Delta z_{l,t})$. By restricting interventions to a small set of salient latent dimensions, SAE-based steering aims to reduce polysemantic interference and improve interpretability relative to dense-state steering.

Figure 3 illustrates the SAE-based steering pipeline. The method first decomposes an activation into sparse latent features, edits selected feature coordinates, and decodes the edited sparse code back into the model activation space. This makes the intervention more interpretable but adds dependence on SAE reconstruction quality and feature-selection choices.

LF-Steering (Yang et al., 2025) steers individual SAE features and reports improved semantic consistency and robustness across NLU/NLG benchmarks. In addition, another approach, FGAA (Soo et al., 2025), combines contrastive analysis with SAE features to construct optimized steering vectors and reports improvements over activation-based steering approaches. CorrSteer (Cho et al., 2025) selects SAE features correlated with task correctness to improve QA and reasoning benchmarks. SAE-SSV (He et al., 2025a) restricts supervised steering to small task-relevant SAE subspaces, aiming for more targeted intervention; because it trains

supervised classifiers/steering vectors, we mark it as a boundary case rather than strict-core inference-only steering.

SAIF (He et al., 2025b) uses sparse autoencoders as a lens for interpreting instruction-following and then turning that interpretation into a steering mechanism: by analyzing SAE latent activations, it identifies a distinct subset of instruction relevant latent features, reports that these features are semantically aligned with the instruction content, and finds that manipulating them can improve instruction adherence. It highlights design choices that matter for SAE steering efficacy, including careful feature identification/selection, the importance of interventions near the final layers, and sensitivity to where the instruction appears in the prompt.

O'Brien et al. (2024) steer refusal behavior by amplifying SAE features associated with refusing unsafe requests. Their results suggest SAE steering can increase robustness to both single-turn and multi-turn jailbreak attempts, but also expose a key caveat: strengthening refusal through SAE features can incur systematic capability degradation on downstream benchmarks, even for benign inputs. This points to an important open problem for SAE-based safety interventions: safety-relevant features may be more entangled with general capabilities than expected, so naive characteristic amplification can trade off utility for safety.

Rather than directly steering by turning SAE features up or down, SAE-TS (Chalnev et al., 2024) uses SAEs as a measurement and targeting tool for steering vectors. The core idea is to evaluate how any candidate steering vector changes activation-space behavior in terms of SAE feature effects, and then construct a steering vector that deliberately targets specific SAE features while minimizing unintended side effects. Empirically, SAE-TS aims to produce steering vectors that better balance "strength of steering" with coherence/quality compared to contrastive baselines (e.g., CAA) and direct SAE-feature steering. SpARE (Zhao et al., 2024) adds a distinct SAE-steering target: selecting between contextual knowledge and parametric memory in knowledge-conflict QA, showing that sparse feature interventions can target information-source selection rather than only style, sentiment, or refusal behavior.

Foundational SAE work. SAE steering is informed by evidence for monosemantic features in transformer activations. Bricken et al. (2023) show that training sparse autoencoders on MLP activations of a one-layer transformer yields features corresponding to interpretable concepts (specific tokens, syntactic roles, semantic categories), providing early evidence that polysemantic neurons can be decomposed into more interpretable units. Templeton et al. (2024) scale this approach to Claude 3 Sonnet, showing that monosemantic features emerge at production scale and include abstract concepts (code errors, deception, bias). Cunningham et al. (2024) provide systematic evidence that SAE features are more interpretable than individual neurons or PCA directions under their evaluations, supporting SAE decomposition as one basis for targeted steering. Together, these works support the view that sparse feature decomposition can expose computationally meaningful structure rather than only compressing activations.

SAE feature-selection caveat. Recent evaluation work complicates a simple "SAE is more interpretable, therefore better" conclusion. AxBench (Wu et al., 2025b) reports that simple baselines can outperform SAE-based steering under some concept-detection and steering evaluations, while feature-selection analyses (Arad et al., 2025) show that SAE steering quality depends strongly on how target features are selected. We therefore treat SAE steering as a promising but evaluation-sensitive family rather than an automatically dominant replacement for dense steering.

5.3 Discussion

Representation-level steering provides a mechanism for changing model behavior without retraining. Rather than shaping only the next-token choice at the final layer, it perturbs the hidden state toward the intended direction (Eq. 5), potentially influencing many downstream tokens simultaneously. This is particularly relevant for failure modes where the model defaults to an undesirable behavioral mode, such as becoming generic, overly verbose, overly cautious, inconsistent in refusal, or, in VLMs, hallucinating objects not present in the image. A targeted shift to intermediate activations can redirect the computation toward a different trajectory. Because the weights remain fixed, the intervention can usually be toggled per request, and

strength is often tunable by scaling $\Delta_{l,t}$. From Table 3, many representative steering rows report target-metric improvements, but the magnitude and side effects depend on model, layer, position, and task.

A further advantage is that these interventions are often accompanied by an interpretable account of what is being changed. Contrastive directions make the control signal explicit ("what differs between two matched prompts or datasets?"), and SAE-based methods decompose dense activations into a sparse set of inspectable and editable features (Eq. 6), which can reduce interference and facilitate debugging in favorable cases. Related intervention principles also appear across settings: in VLMs, one can intervene either before fusion (cleaning up visual evidence) or after fusion (steering the shared multimodal stream), and diffusion models admit training-free edits on feature maps and attention-like representations. The central open problem is reliability: identifying layers and positions that generalize across prompts, avoiding spurious directions, determining when to activate the steer (rather than applying it uniformly), and preventing side effects across diverse tasks such as style, safety, and grounding.

Cross-method comparison: ActAdd vs. CAA vs. mean-centering vs. RepE. These methods share the same mathematical form (Eq. 5) but differ in how $\Delta_{l,t}$ is computed. ActAdd uses a single contrastive pair, making it fast but noisy, since the steering direction may capture prompt-specific artifacts rather than the intended behavioral axis. CAA averages over a dataset of contrastive pairs, producing more stable directions at the cost of dataset construction effort; it is widely used as a dense steering baseline. Mean-centering (Jorgensen et al., 2023) subtracts a behavior-agnostic baseline before computing the contrastive difference, reducing spurious shifts and improving function vector extraction (+16.5% absolute accuracy on average). RepE (Zou et al., 2023) uses PCA on contrasting stimuli to extract multi-dimensional control subspaces rather than single directions, enabling simultaneous reading and writing of multiple concepts, but requires more careful selection of stimulus pairs and principal components. In practice, CAA is a natural choice when a single behavioral axis is targeted and annotated contrast data is available; RepE is useful when interpretability of the control signal matters; and mean-centering can be a useful preprocessing step for contrastive methods when behavior-agnostic baselines are available.

Dense steering vs. SAE steering. Dense methods (ActAdd, CAA, RepE) are simpler, usually require only lightweight activation additions at inference time, and require no auxiliary model. SAE methods (FGAA, SAE-SSV[‡], LF-Steer) require an additional encode-decode or feature-selection step but offer more direct feature-level interpretability, since one can inspect which features are being modified, and may reduce polysemantic interference. Empirically, reported SAE rows show gains on tasks where the target concept aligns well with a small number of SAE features (e.g., sentiment: SR 45.6% $\rightarrow$ 63.2% for boundary-marked SAE-SSV), but SAE methods can underperform dense methods when the concept is distributed across many features. The tradeoff is: dense steering is often the simpler deployment default; SAE steering is most attractive when interpretability or fine-grained control is required, and when a well-trained SAE is available for the target model. Overhead values are approximate and implementation-dependent, so exact latency should be measured in the target implementation.

6 Attention-Level Interventions

Attention steering refers to inference-time interventions that bias a model's attention computation to improve alignment with a user goal (e.g., follow instructions, ground on evidence, reduce unsafe content, or improve serving efficiency). These methods modify the model's information routing mechanism, for example by reweighting attention over highlighted tokens, adding biases to attention logits, or editing key/query/value representations, while keeping base weights fixed.

A minimal formalization. Let $S_{l,h} = Q_{l,h}K_{l,h}^{\top}/\sqrt{d_h}$ denote the pre-softmax attention logits at layer l, head h. Attention-level steering modifies these logits before they are applied to values:

$$\tilde{S}_{l,h} = S_{l,h} + \Gamma_{l,h}, \qquad \tilde{A}_{l,h} = \mathrm{softmax}(\tilde{S}_{l,h}), \tag{7}$$

where $\Gamma_{l,h}$ is a perturbation matrix that may be (i) a fixed additive bias promoting specific token positions, (ii) a multiplicative or masked correction over designated regions, (iii) a state-dependent correction computed

from decoding context, or (iv) a redistribution operator that reallocates attention mass from sink tokens to task-relevant positions. Some methods instead modify the post-softmax attention matrix $A_{l,h}$ directly, in which case clipping or renormalization is typically required to preserve a valid attention distribution. The key constraint remains: model parameters θ are unchanged.

Token reweighting. These approaches directly reshape attention mass toward selected tokens or spans, typically without closed-loop feedback. In LLMs, this is often instantiated as prompt highlighting or instruction emphasis: PASTA (Zhang et al., 2024) reweights attention toward user-highlighted spans, while InstABoost (Guardieiro et al., 2025) uses a minimalistic rule that boosts attention to instruction tokens. In VLMs, the same mechanism targets visual grounding by reweighting attention over visual tokens or regions: CAAC (Fazli et al., 2025) mitigates hallucination by balancing attention across visual tokens and adaptively reinforcing visual grounding based on token-level confidence, CARVE (Ge et al., 2025) refines attention by contrasting task-specific versus general attention maps to suppress visual noise and emphasize task-relevant evidence, and CAI (Li et al., 2025b) applies caption-sensitive attention intervention to strengthen visual grounding. Because CAI is recent, we keep the scope claim limited to reported training-free inference behavior pending full reproducibility checks.

Dynamic control (state-dependent, progressive, gated, or closed-loop). Dynamic methods make steering conditional on the decoding state, changing strength over time, triggering only when needed, or using feedback signals to regulate attention. SpotLight (Venkateswaran & Contractor, 2026) represents time-varying prompt highlighting, while COMPASS (Sahay et al., 2025) frames grounding as feedback control by steering attention toward evidence when hallucination risk rises; because its controller uses a trained token-level hallucination-risk classifier, we mark it as a boundary method. LLMSTEER (Gu et al., 2024) adapts steering to long-context serving, especially under context reuse, by applying an attention policy tuned to deployment constraints rather than a static prompt rule. In VLMs, analogous concerns arise during long-form reasoning and captioning where visual attention can drift or fade: SPARC (Jung et al., 2025) progressively recalibrates attention to preserve visual grounding over long generations; cross-attention gating approaches trigger realignment when the model appears to rely too little on visual evidence; and Qwen Look Again (Chu et al., 2025) explicitly re-attends via reflection-like steps and visual-token operations. Because Qwen Look Again includes BRPO reinforcement learning in addition to visual-token operations, we treat it as a trained boundary method rather than strict-core inference-time steering.

Conscious Gaze for VLMs. CG-VLM (Bu et al., 2025) introduces an adaptive attention mechanism that mimics "conscious gaze," dynamically adjusting which image regions receive attention based on the evolving generation context. On LLaVA, CG-VLM improves POPE F1 from 79.5 to 86.1 and reduces CHAIRi from 9.3 to 6.5, reporting evidence that adaptive attention modulation can reduce hallucination metrics in that setting.

Redistribution (attention budget recycling/sinks). A recurring long-context phenomenon is attention sinks: tokens (often BOS, but sometimes other low-semantic positions) that attract disproportionate attention and can consume routing budget. StreamingLLM is deployment-adjacent rather than a steering method: it shows that preserving the KV cache of initial sink tokens largely restores streaming performance under windowed attention, and proposes adding a dedicated sink token during pretraining to better support deployment (Xiao et al., 2024). However, sinks are not always pure waste: the first-token sink can play a stabilizing role by helping Transformers avoid over-mixing, so indiscriminate suppression can hurt quality (Barbero et al., 2025). This motivates selective recycling methods such as ACT, which detects sink tokens (including beyond BOS) and downscales their attention in chosen heads/layers while reallocating the freed mass to non-sinks with normalization, improving accuracy without weight updates (Yu et al., 2024). Complementary analysis suggests that many sinks behave like key-bias artifacts, storing attention scores while contributing little to value computation, supporting the view that some sink mass is recyclable surplus (Gu et al., 2025).

The idea can be extended to multimodal KV sets by identifying fixed-location visual sink tokens (linked to large sink-dimension activations) that can often be removed with little loss, then redistributing their attention within image-centric heads to strengthen evidence use and reduce hallucinations (Kang et al.,

2025). EVAS (Zhang et al., 2025b) similarly links hallucinations to sink structure across layers and increases image focus by broadcasting the attention map from the densest sink head to other heads in the same layer. SPIN (Sarkar et al., 2025) takes a complementary route by suppressing image-inattentive heads at inference time, reducing object hallucination without training a new controller.

Deep inside attention (activation steering inside attention modules). Some methods intervene not on attention weights but within attention submodules, editing Q/V representations or attention-head activations. We include these in attention steering but explicitly label them as activation steering within attention modules, because the interventions are activation-level yet attention-located. In LLMs, DISCO (Torop et al., 2025) injects steering signals into Q/V spaces to obtain more disentangled control; ITI (Li et al., 2023a) shifts activations in selected attention heads to steer behavioral traits such as truthfulness.

Module localization and intervention enablers (LLMs). Because steering success depends strongly on where to intervene, a complementary line of work focuses on localizing behavior-relevant modules (heads/layers) and providing enablers for targeted control. Classic head-importance studies show that only a subset of attention heads materially affect model behavior (Michel et al., 2019; Voita et al., 2019). REAL (Zhan et al., 2025) localizes behavior-relevant heads or layers for steering. For safety, Zhou et al. (2025) introduce SHIPS and SAHARA to score and attribute safety-critical attention heads, showing that ablating a small number of such heads can measurably compromise refusal/safety behavior.

Extension to VLMs. In VLMs, analogous analyses identify "image-centric heads" and "sink tokens" that absorb disproportionate attention, creating actionable loci for interventions such as attention budget redistribution (e.g., VAR) (Kang et al., 2025). Mechanistic studies trace where and how visual and linguistic information are integrated across layers, helping isolate candidate cross-modal integration/failure loci for localization or editing (Zhang et al., 2025d). VEGAS (Wang et al., 2025) extends this localization logic through a training-free, inference-time attention-representation hybrid that injects vision-encoder attention guidance into language-model mid-layers. TAB (Rahmanzadehgervi et al., 2025) is an architectural intervention enabler rather than a strict fixed-architecture steering method: it introduces an explicit attention bottleneck that exposes a single, editable attention map, making user intervention easier to inspect and apply.

Discussion. Attention steering is most clearly suited to context utilization failures: cases where the model already has the relevant information in its current context (prompt, retrieved evidence, or image tokens), but appears not to use it because attention is misallocated. In practice this manifests as (i) **instruction underuse**, where instruction tokens lose salience as decoding progresses and the model defaults to generic continuation; (ii) **evidence neglect**, where cited passages or grounded visual regions are present but weakly attended, leading to unfaithful or hallucinatory generation; (iii) **long-context pathologies**, where attention budget is consumed by structural tokens (e.g., sinks, boilerplate prefixes) rather than task-relevant spans; and (iv) in VLMs, modality drift, where the model increasingly relies on language priors instead of image evidence during long-form generation. Attention-level interventions are useful here because they offer a per-request, inference-time handle to re-route information flow without updating base weights, complementing other approaches (e.g., retrieval or prompting) with the aim of increasing the probability that available evidence is consumed by the decoder.

7 Decoding-Level Interventions

Transformer models often assign non-trivial probability mass to undesired or low-quality tokens for various reasons (e.g., repetition loops, generic high-probability phrasing, premature stopping) or due to competing signals (e.g., language priors vs. visual evidence). Decoding intervention provides an approach that changes which tokens are selected during generation by modifying the token scores used by the decoder (i.e., logits/probabilities) and/or the decoding objective (e.g., re-ranking candidates in beam search), without updating the model parameters. Let $\mathbf{z}_t$ denote the base model logits at step t. Decoding intervention forms

adjusted logits $\tilde{\mathbf{z}}_t$ before sampling/argmax as follows:

$$\tilde{\mathbf{z}}_t \;=\; \mathbf{z}_t \;+\; \sum_k \lambda_k \, s_k(x, y_{<t}), \tag{8}$$

where each $s_k(x, y_{<t})$ is an additive steering score produced by mechanism k (e.g., a contrastive score, a penalty term, or evidence-based adjustment), and λ_k is a scalar weight that controls how strongly that mechanism influences decoding. Setting $\lambda_k = 0$ disables signal k, while larger $|\lambda_k|$ yields stronger steering. For reranking, frozen-verifier-guided decoding, and constrained search, the same idea can be written at the sequence level:

$$\hat{y} = \arg\max_{y \in \mathcal{C}} \left[\log p_\theta(y|x) + \sum_k \lambda_k S_k(x, y) \right], \tag{9}$$

where $\mathcal{C}$ is a candidate set and $S_k(x, y)$ scores a complete candidate sequence.

7.1 Single-Model Objective Reshaping

The most lightweight family reshapes the decoding objective using only signals available within a single forward pass (or within a single model), targeting generic decoding failures such as repetition and blandness. Contrastive Search (Su et al., 2022) augments token selection with a representation-similarity penalty, discouraging candidates that are too similar to previously generated states, and thereby improving coherence and diversity; in the original SimCTG framework, this decoding rule is paired with a contrastive training objective, so we treat it as a decoding reference rather than a strict-core method when that training is required. Entropy Guided Extrapolative Decoding (Das et al., 2025) leverages internal layer-wise distributions and entropy-based criteria to calibrate generation; the method remains training-free but typically assumes white-box access to intermediate-layer outputs.

Layer-contrast decoding (DoLa). DoLa (Chuang et al., 2024) introduces a decoding-time factuality improvement method that contrasts logits from later ("mature") layers against earlier ("premature") layers within the same model. The method is motivated by the observation that factual associations can become more pronounced in later layers while earlier layers capture shallower patterns. By computing $\tilde{\mathbf{z}}_t = \log p_L(y_t|\cdot) - \log p_l(y_t|\cdot)$ where L is the final layer and l is a dynamically selected premature layer, DoLa upweights tokens favored by later-layer distributions relative to earlier-layer distributions. On TruthfulQA, DoLa improves LLaMA-7B from 26.9% to 40.8% (%T×I) and LLaMA-33B from 31.7% to 49.1%, with reported small additional latency from using selected intermediate-layer outputs. DoLa requires only a single model and no external data, making it a comparatively simple factuality intervention. Its strongest evidence is on factuality-style evaluation, while effects on reasoning and instruction-following tasks are more task- and model-dependent.

Classifier-guided and attribute-conditioned decoding as boundary cases. A complementary historical family uses lightweight classifiers or attribute models to steer decoding, but these methods are outside the strict no-partly-trained core when the classifier, expert, or reward model is newly trained for the steering task. FUDGE (Yang & Klein, 2021) trains a future discriminator that predicts whether a partial generation will satisfy a constraint, GeDi (Krause et al., 2021) uses a class-conditional language model as a generative discriminator, DExpert (Liu et al., 2021) composes expert and anti-expert LMs, and DeAL (Huang et al., 2025) uses decoding-time alignment/search with reward or heuristic signals. They are useful conceptual reference points for Eq. 8, but the core taxonomy emphasizes methods whose decoding-time scores come from the frozen base model, controlled perturbations, or frozen/pre-existing auxiliary signals.

Diffusion guidance as objective steering. Reward- and guidance-based diffusion methods such as SEGA (Brack et al., 2023) and Feynman-Kac diffusion steering (Singhal et al., 2025) operate by changing the sampling-time objective or guidance signal rather than by directly editing a fixed activation subspace. We therefore group them with decoding/objective-level boundary methods in the taxonomy, while treating activation patching, cross-attention edits, and feature-map interventions as the diffusion analogues of representation- or attention-level steering.

7.2 Logit Composition Across Pretrained Models

Whereas single-model methods reshape decoding using only the base model's internals, a second family composes distributions from multiple pretrained models, using logit arithmetic to amplify desirable behaviors and suppress undesired ones.

Contrastive Decoding (Li et al., 2023b) contrasts an "expert" model against a smaller "amateur" model (expert minus amateur) under a plausibility constraint, which can yield higher-quality generations at the cost of an additional forward pass.

ThinkLogit (Zhang et al., 2025c) follows a similar principle: it uses decoding-time logit arithmetic for a frozen target model together with a separately trained or distilled guider distribution to elicit longer-form reasoning behavior. We therefore treat it as a boundary method with respect to the strict no-new-model-training criterion, even though the target model itself remains frozen at inference time. Empirically, this general mechanism can also affect reasoning accuracy: O'Brien & Lewis (2023) report that contrastive decoding can improve performance on reasoning tasks relative to standard decoding, reinforcing that decoding-time choices may change not just style but also error modes.

Proxy-based boundary methods and speculative-decoding enablers. Proxy Tuning (Liu et al., 2024a) extends logit composition to the setting where one has a small tuned model and a large untuned model: it adds the logit difference between the tuned and untuned small models to the large model's logits, effectively "transferring" the tuning signal without ever training the large model. It is a boundary/background method in our taxonomy because it depends on a tuned proxy model, but it is useful when the large model's weights are inaccessible (API-only). In the efficiency-oriented direction, speculative decoding methods (Zhang et al., 2023; Miao et al., 2024) use draft/proposal verification mechanisms, including self-drafting or skipped-layer verification and tree-based speculative inference, to reduce latency while preserving the target distribution. Speculative decoding is efficiency-adjacent rather than a direct steering method, although it can interact with steering because the proposal distribution determines which tokens the verifier considers.

7.3 Multimodal Contrast via Controlled Perturbations

In VLMs, hallucinations are often attributed to weak visual grounding relative to language priors. A common training-free strategy is to form two output distributions under controlled perturbations and then contrast them to suppress tokens that persist under the perturbation (interpreted as prior-driven). Visual Contrastive Decoding (VCD) (Leng et al., 2024) contrasts logits obtained from the original image and a distorted image to reduce object hallucinations. Instruction Contrastive Decoding (ICD) (Wang et al., 2024c) performs an analogous contrast in the instruction space by pairing the original prompt with a "disturbance" instruction designed to elicit hallucinations, and subtracting the disturbance-induced distribution. Retrieved Visual Contrastive Decoding (RVCD) (Lee & Song, 2025) replaces synthetic image corruption with retrieved positive and negative visual references, producing a semantically richer counterfactual signal at the cost of retrieval and extra visual forward passes. ConVis (Park et al., 2025) uses a text-to-image model to reconstruct images from candidate captions and contrasts distributions between the original and reconstructed images, trading off semantic consistency gains against higher inference cost.

Additional multimodal contrastive methods. AvisC (Woo et al., 2025) extends visual contrastive decoding by introducing adaptive visual contrastive scoring that adjusts contrast strength based on per-token visual relevance, avoiding the problem of uniform contrast strength degrading tokens that genuinely require language-prior guidance. Fuzzy Contrastive Decoding (Kim et al., 2025) similarly makes the contrast strength confidence-aware, softening the intervention when object evidence is ambiguous. RITUAL (Woo et al., 2024) uses randomly transformed versions of the input image as complementary visual contexts during decoding and combines output distributions to reduce hallucination. M3ID (Favero et al., 2024) proposes a multi-modal mutual information-based decoding strategy that maximizes the mutual information between the generated text and the visual input, providing an information-theoretic foundation for multimodal contrastive decoding.

Cross-method comparison: VCD vs. ICD vs. RVCD vs. ConVis. These methods target the same problem (VLM hallucination) via contrastive decoding but differ in how the counterfactual distribution is constructed. VCD (Leng et al., 2024) perturbs the visual input (Gaussian noise on the image), making it simple to apply but potentially too coarse, since some hallucinations are prompt-driven rather than vision-driven and VCD may not address these. ICD (Wang et al., 2024c) perturbs the instruction instead, targeting prompt-induced hallucinations but potentially missing vision-grounding failures. RVCD (Lee & Song, 2025) uses retrieved visual references, which can provide a more semantically meaningful contrast but introduces retrieval quality and latency dependence. ConVis (Park et al., 2025) uses a text-to-image reconstruction loop, providing a richer semantic contrast but at higher overhead. Source-reported rows show that VCD improves LLaVA-1.5 POPE MSCOCO-random F1 (81.33 $\rightarrow$ 87.16), ICD improves InstructBLIP POPE MSCOCO-random F1 (80.41 $\rightarrow$ 85.61), ConVis reduces LLaVA-1.5 CHAIR_S (22.4 $\rightarrow$ 18.4), and OPERA reduces LLaVA-1.5 CHAIR_S under its 64-token setting (20.6 $\rightarrow$ 14.2). The takeaway is that no single contrastive signal appears preferable across all settings: VCD is a low-overhead option among these variants, ICD complements it for instruction-driven errors, RVCD is attractive when retrieval is available, and ConVis may be a reasonable choice when hallucination cost is high and inference budget allows.

7.4 Attention-Aware Decoding Penalties

Often, some multimodal methods derive steering signals from attention diagnostics or cross-modal correlation measurements while still applying the intervention at the decoding stage. OPERA (Huang et al., 2024) introduces an over-trust penalty and a retrospection-allocation mechanism (implemented within beam-search-style decoding) to mitigate hallucinations without retraining. IMCCD (Li et al., 2025a) calibrates inter-modality correlation by constructing a counterfactual ("distorted") distribution through masking value vectors associated with cross-modal attention patterns, and then applying contrastive decoding to down-weight hallucination-prone tokens. FarSight (Tang et al., 2025) uses attention causal decoding with causal masks and attention-register optimization, while Self-Introspective Decoding (Huo et al., 2024) uses internal attention mechanisms to construct a training-free self-introspective token-importance signal. These methods represent a broader hybrid trend: attention-derived signals can be used to define penalties or counterfactual distributions, but the final control lever remains the token-selection rule.

Discussion. Decoding intervention addresses failures that arise at the token-selection stage: a useful continuation may be available under the model distribution, but the default decoding rule can select a worse one. This can occur because decoding follows short-term patterns, including repeating phrases, falling back to safe/generic wording, or ending prematurely, or because it overweights certain signals such as language priors over visual evidence in VLMs. Equation (8) captures the core mechanism: before choosing the next token, we adjust the token scores $\tilde{\mathbf{z}}_t$ (or re-rank candidates), adding preferences at the point where the decision is made, such as penalizing repetition, rewarding diversity, favoring outputs consistent with evidence, or reducing multimodal hallucinations via contrastive signals. These methods can often be layered on top of existing sampling or beam search and applied without changing model weights, although they may require extra logits, forward passes, or auxiliary scores.

8 Cross-Method Comparison and Analysis

Table 3 provides a unified summary of representative methods across all three intervention levels, including representative models, benchmarks, metrics, and reported base versus steered performance. Rows are representative rather than exhaustive: we include methods when the source paper reports clear base-versus-steered numbers that can be summarized compactly. Because the rows come from different papers and experimental setups, the table is best used for side-by-side orientation and within-paper deltas rather than definitive cross-paper ranking.

Table 3: **Unified performance summary across intervention levels.** The table summarizes representative base-to-steered changes under the listed model, benchmark, metric, and decoding settings. Exact comparability is limited by differences in models, prompts, metrics, and evaluation protocols. Rows marked ‡ use trained classifiers, auxiliary steering modules, or trained/distilled guider policies and are boundary cases outside the strict no-partly-trained core.

Task	Method	Source	Model	Benchmark	Metric	Reported change
Representation-Level: LLM Steering						
Toxicity reduction	ActAdd	(Turner et al., 2023)	OPT-6.7B	RealToxicityPrompts	Mean toxicity	Base: 0.134 Steered: 0.112
Sentiment control	ActAdd	(Turner et al., 2023)	LLaMA3-8B	IMDb	Success (neg.→pos.)	Base: 0.417 Steered: 0.669
Capability audit	CAA	(Rimsky et al., 2024)	Llama-2-13B-Chat	MMLU (refusal steer)	Avg. prob(correct)	Base: 0.63 Steered: 0.59
Function vectors	Mean-center	(Jorgensen et al., 2023)	GPT-J-6B	6 FV tasks	Accuracy (%)	Base: 29.2 Steered: 45.7
Detox.	ICV	(Liu et al., 2024b)	Llama-7B	ParaDetox	Toxicity (%)	Base: 71.6 Steered: 39.54
Style transfer	ICV	(Liu et al., 2024b)	Llama-7B	GYAFC	Formality (%)	Base: 17.54 Steered: 48.3
Semantic consistency	LF-Steering	(Yang et al., 2025)	Llama2-7B-Chat	RobustBOOLQ	Accuracy (%)	Base: 46.4 Steered: 66.4
Multi-goal steering	FGAA	(Soo et al., 2025)	Gemma-2-2B	9 steering goals	Mean BCS	Base: 0.2201 Steered: 0.4702
Sentiment steering	SAE-SSV‡	(He et al., 2025a)	Llama3.1-8B	Steering eval	SR (%)	Base: 45.6 Steered: 63.2
Representation-Level: VLM and Diffusion Steering						
Halluc. mit.	VTI	(Liu et al., 2025)	LLaVA-1.5	CHAIR/POPE/MMHal	CHAIRs/Acc./ Avg. score	Base: 51/79.8/1.99 Steered: 35.8/86.5/2.9
Safety steering	L2S‡	(Parekh et al., 2025)	LLaVA-v1.5	MMSafetyBench	Unsafe-score	Base: 0.234 Steered: 0.057
Halluc. mit.	L2S‡	(Parekh et al., 2025)	LLaVA-v1.5	POPE/COCO	Acc./CHAIRs	Base: 82.73/17.31 Steered: 86.46/16.1
Halluc. mit.	SteerVLM‡	(Sivakumar et al., 2025)	LLaVA1.5-7B	OHD	Accuracy (%)	Base: 84.5 Steered: 86.4
Topic control	SteerVLM‡	(Sivakumar et al., 2025)	LLaVA1.5-7B	5 steering vectors	Judge score	Base: 0.55 Steered: 0.71
Halluc. mit.	ASD	(Su et al., 2025)	LLaVA1.5-7B	POPE MSCOCO	F1 (↑)	Base: 86.03 Steered: 87.87
Concept erasure	CASteer	(Gaintseva et al., 2025)	Stable Diffusion	I2P/COCO-30K	CLIP/FID (↓)	Base: 31.34/14.04 Steered: 30.09/13.02
Attention-Level: LLM Steering						
Multiple tasks	PASTA	(Zhang et al., 2024)	LLaMA-7B	Multiple Bench.	All (avg)	Base: 73.45 Steered: 95.46
Multiple tasks	PASTA	(Zhang et al., 2024)	GPT-J	Multiple Bench.	All (avg)	Base: 59.65 Steered: 85.22
Instruction following	SpotLight	(Venkateswaran & Contractor, 2026)	Qwen2.5 7B	IFEval	Prompt/Instr. acc	Base: 0.47/0.59 Steered: 0.54/0.66
Many-instr. following	SpotLight	(Venkateswaran & Contractor, 2026)	Qwen2.5 7B	ManyIFEval	Prompt/Instr. acc	Base: 0.20/0.63 Steered: 0.26/0.70
Attention-Level: VLM Steering						
Halluc. mit.	CG-VLM	(Bu et al., 2025)	LLaVA	POPE (COCO)	F1	Base: 79.5 Steered: 86.1
Halluc. mit.	CG-VLM	(Bu et al., 2025)	LLaVA	CHAIR	CHAIRi (↓)	Base: 9.3 Steered: 6.5
Halluc. mit.	VAR	(Kang et al., 2025)	LLaVA-1.5-7B	POPE/MMHalBench	F1	Base: 84.76 Steered: 85.87
Halluc. mit.	SPIN	(Sarkar et al., 2025)	LLaVA-1.5-13B	CHAIR	CHAIR_S (%)	Base: 41.4 Steered: 29.2
Halluc. mit.	Qwen-LookAgain‡	(Chu et al., 2025)	Qwen2.5-VL	CHAIR/POPE/MMHal	Mixed	Base: 9.4/37.1/88.7/3.68 Steered: 3.7/9.8/90.2/3.82
Decoding-Level: LLM Steering						
Reasoning accuracy	CD for reasoning	(O'Brien & Lewis, 2023)	LLaMA-65B	GSM8K/HellaSwag	Accuracy (↑)	Base: 51.0/84.2 Steered: 56.8/88.0
Long-CoT elicitation	ThinkLogit‡	(Zhang et al., 2025c)	Qwen2.5-32B	AIME25/AMC23/ MATH-hard	Pass@1	Base: 8.3/57.2/50.8 Steered: 19.2/62.2/60.7
Factuality	DoLa	(Chuang et al., 2024)	LLaMA-7B/ 33B	TruthfulQA	%T×I (↑)	Base: 26.9/31.7 Steered: 40.8/49.1
Decoding-Level: VLM Steering						

Continued on next page

Table 3: **Unified performance summary across intervention levels** (continued).

Task	Method	Source	Model	Benchmark	Metric	Reported change
VLM hallucination	VCD	(Leng et al., 2024)	LLaVA-1.5	POPE MSCOCO Rand	F1 (↑)	Base: 81.33 Steered: 87.16
VLM hallucination	ICD	(Wang et al., 2024c)	InstructBLIP	POPE MSCOCO Rand	F1 (↑)	Base: 80.41 Steered: 85.61
VLM hallucination	FarSight	(Tang et al., 2025)	LLaVA-1.5	POPE-R	Score (↑)	Base: 87.0 Steered: 90.5
VLM hallucination	RVCD	(Lee & Song, 2025)	LLaVA-1.5	POPE greedy	F1 (↑)	Base: 77.86 Steered: 88.43
VLM hallucination	ConVis	(Park et al., 2025)	LLaVA-1.5/ MiniGPT-4	CHAIR_S	CHAIR_S (%)	Base: 22.4/34.0 Steered: 18.4/23.5
VLM hallucination	OPERA	(Huang et al., 2024)	LLaVA-1.5	CHAIR_S (64 tok.)	CHAIR_S (%)	Base: 20.6 Steered: 14.2

Key observations. (1) *Representation-level methods* often have low inference overhead when implemented as dense activation additions, while SAE-based variants require additional feature encode/decode or lookup steps. Dense methods (ActAdd, CAA, Mean-ctr) are therefore often the most straightforward deployment path; SAE overhead is approximate and implementation-dependent. (2) *Attention-level methods* range from lightweight attention reweighting to moderate costs for methods with progressive recalibration or trained boundary components like Qwen-LookAgain, occupying a middle ground. (3) *Decoding-level methods* can impose higher cost when they require extra model passes, retrieved references, reranking, or reconstruction. Single-model variants like DoLa add small additional latency from intermediate-layer outputs, while multi-model contrastive decoding (CD), boundary guider-based methods such as ThinkLogit, and reconstruction-based methods (ConVis) can be more expensive. (4) *No single level dominates*: representation methods are often suited for broad behavioral shifts (toxicity, sentiment, safety), attention methods are often suited for evidence grounding and instruction routing, and decoding methods are useful when the desired control can be expressed as token-score reshaping or candidate reranking. (5) *Failure modes are complementary*: representation steering can affect unrelated task probabilities (CAA with a refusal vector on MMLU: 0.63 → 0.59); contrastive decoding can regress on some tasks or model-size settings (e.g., CD on CSQA in one reported setting: 66.4 → 62.8 (O'Brien & Lewis, 2023)); and attention interventions can trade off grounding gains against routing or localization errors.

Evaluation inconsistency across methods. A significant challenge for cross-method comparison is the lack of standardized evaluation protocols. Methods in Table 3 are evaluated on different benchmarks (TruthfulQA, POPE, CHAIR, RealToxicityPrompts, custom paper-specific evaluations), different model families (GPT-J, LLaMA, Qwen, LLaVA, InstructBLIP), and with different metrics. For example, hallucination mitigation is measured via CHAIR_S (lower is better) in some papers and POPE F1 (higher is better) in others, and these metrics can disagree on which method is superior. Until the community converges on shared benchmarks and standard experimental protocols for steering evaluation, cross-paper comparisons should be interpreted with appropriate caution. We report numbers as published in the original papers and encourage readers to prioritize within-paper ablations (where experimental conditions are controlled) over cross-paper comparisons.

Composability of interventions across levels. A natural question is whether interventions from different levels can be composed, for example by applying representation steering together with contrastive decoding. In principle, the three levels target distinct computational stages and may be composable: representation methods modify hidden states in intermediate layers, attention methods modify information routing, and decoding methods modify the final token-selection rule. However, systematic empirical study of composed interventions is largely absent from the literature. ASD (Su et al., 2025) represents a partial example, applying activation shifts to both hidden states and decoder outputs (combining representation and decoding-level intervention). OPERA (Huang et al., 2024) combines attention diagnostics with decoding penalties. We identify principled composition of multi-level interventions, including interaction effects, diminishing returns, and optimal allocation of intervention budgets, as an important open direction (Section 10).

Reproducibility and code availability. Practical adoption of steering methods depends on reproducibility. Among the methods surveyed, examples with publicly available implementations include core steering methods such as ActAdd, CAA, DoLa, VCD, ICD, and OPERA, as well as deployment-adjacent infrastructure such as StreamingLLM. SAE steering methods benefit from the growing ecosystem of open-source SAE training tools (e.g., the SAELens library). However, many methods, particularly those involving custom attention modifications or proprietary model access, lack public implementations, making independent verification difficult.

Beyond reported gains, deployment decisions depend on the operational risks introduced by each intervention family. Table 4 summarizes these tradeoffs qualitatively. The goal is not to assign universal risk scores, but to make explicit which failure modes should be checked when choosing between representation-, attention-, and decoding-level control.

Risk / constraint	Representation	Attention	Decoding
Capability degradation	High when steering is broad or strong	Medium; routing changes can distort evidence use	Low–medium; usually localized to token choice
Prompt or jailbreak bypass	Medium; refusal directions can be ablated or overfit	Medium; highlighted spans can be adversarial	Medium; constraints can be gamed by candidate search
Latency overhead	Low for dense steering; medium for SAE steering	Low–medium depending on attention access	Medium–high for extra passes or reranking
Interpretability	Medium; higher for SAE-localized features	Medium when heads/spans are localized	Low unless verifier scores are inspectable
Dual-use risk	High for safety-feature removal or behavior suppression	Medium for routing around safety evidence	Medium for optimizing unsafe candidates
Composability uncertainty	High; vectors can interfere	High; routing changes alter downstream states	High; score terms can double count signals
Required access	Hidden-state or full white-box access	Attention/KV or full white-box access	Logit access; sometimes auxiliary models
Hyperparameter sensitivity	High: layer, position, strength, schedule	Medium: heads, spans, renormalization	High: contrast weight, candidate set, temperature

Table 4: **Risk matrix for inference-time interventions.** Qualitative risks vary by model, task, access level, and intervention strength; the labels summarize common deployment tradeoffs rather than absolute guarantees.

9 Limitations of This Survey

We acknowledge several limitations of this survey that readers should consider when interpreting our analysis.

Selection bias and recency. Despite the explicit inclusion criteria in Section 2, our coverage is inevitably shaped by the venues, keywords, and citation trails we monitored. The majority of methods surveyed were published in 2024–2025, reflecting the rapid growth of this area but also introducing a recency bias: foundational earlier work may receive less detailed treatment than recent methods. Methods published at workshops, on smaller arXiv-only venues, or in non-English literature may be underrepresented.

Cross-paper comparison challenges. The performance numbers in Table 3 are drawn from original papers, each with its own experimental setup (model checkpoints, hyperparameters, hardware, evaluation splits). Differences in base model versions, prompt templates, and evaluation protocols mean that cross-paper numerical comparisons are inherently approximate. We encourage readers to treat these as indicative of general trends rather than definitive rankings. Within-paper ablations provide more reliable evidence for the effectiveness of individual methods.

Modality coverage imbalance. Our coverage of LLM and VLM methods is more detailed than our coverage of diffusion model interventions. This reflects the current state of the literature: inference-time

interventions for diffusion models are a newer and smaller subfield, with fewer published methods meeting our inclusion criteria. Audio, video, and embodied-agent foundation models are excluded entirely, though inference-time intervention in these domains is an emerging area.

Scope boundary: inference-only constraint. Our strict inference-only criterion excludes methods that require any weight modification, including parameter-efficient approaches (LoRA, ReFT, LoRRA) that share conceptual similarities with the methods we survey. It also excludes methods that train a new controller, adapter, classifier, reward model, hypernetwork, or probe for the steering task, even if the base model itself remains frozen. This boundary is sometimes fuzzy: methods like SteerVLM and L2S train lightweight add-on modules, and SAE training itself requires optimization even though SAE *steering* is inference-only. We therefore separate strict-core methods from boundary cases whenever the mechanism depends on newly trained auxiliary components.

Excluded partly trained or scope-edge methods. Under this filter, FairSteer (Li et al., 2025c), AutoSteer (Wu et al., 2025a), ActLCD (Zhang et al., 2025a), Steering Vector Fields (Li et al., 2026), CLAS (Hsu et al., 2026), Decoupling Contrastive Decoding (Chen et al., 2025b), HyperSteer (Sun et al., 2025), RePS (Wu et al., 2025c), Reward-Guided Decoding for MLLMs (Mañas et al., 2025), SafeSteer (Zeng et al., 2025), and PII-jailbreaking activation-steering studies (Nakka et al., 2025) are treated as boundary or scope-edge entries rather than core representative methods. The reasons differ by paper: some require learned steering policies or optimized auxiliary components, some rely on reward/search machinery outside our strict intervention definition, some are primarily safety-attack analyses, and some are too recent for stable implementation-status claims. They remain relevant for future work because they illustrate where the field is moving if the scope is broadened beyond pure fixed-weight inference-time intervention.

Limited coverage of negative results. The published literature on steering methods tends to emphasize positive results. Systematic negative results, where steering fails, degrades quality, or produces unexpected side effects, are less frequently reported. The degradation cases we identify (CAA on MMLU, CD on CSQA/StrategyQA settings, and SAE refusal steering degrading capabilities) are likely incomplete. A more complete picture of when and why steering fails would benefit practitioners.

10 Future Directions and Open Problems

Representation intervention. Widespread adoption of representation steering still remains limited, largely due to **reliability**: effects can vary sharply across inputs, include "anti-steerable" cases, and may degrade truthfulness/safety because hidden states are polysemantic and causally entangled. While reliability remains an open direction, another area of improvement concerns how contrasts used to find steering directions are constructed. Instead of relying on weak or noisy "before vs. after" examples, the field should develop better baselines and more systematic contrast sets, so the steering direction captures the intended effect rather than accidental correlations. When and where to apply control remains challenging and inconsistent across the literature: different papers recommend different token positions and layers (for LLMs), or different timesteps and spatial regions (for diffusion). Because the "right" intervention point is highly context-dependent, applying control too broadly can cause unwanted side effects like artifacts, erasure, or degraded truthfulness/safety, while applying it too narrowly may fail to steer at all.

Attention-level interventions. Attention interventions primarily change where the model routes information rather than adding new knowledge. This means they are most appropriate when errors plausibly come from misrouting, but they have limited impact when the bottleneck is missing knowledge or weak reasoning. A future direction is diagnosis and gating: building lightweight tests that estimate whether a failure is routing-driven (so attention control is likely to help), and triggering interventions only in those cases. Given that static emphasis can be brittle and over-focus on the wrong span/region, the next step is to strengthen attention when the highlighted evidence agrees with the prompt and the model is confident, and relax it when evidence conflicts or uncertainty is high. Long-context results add a further constraint: "attention sinks" are not always waste and can sometimes stabilize computation, so the goal is not blanket suppression but selective recycling, downscaling only surplus sink mass in specific heads/layers and redistributing it while

preserving normalized attention structure. In multimodal models, the same principle suggests architecture-aware policies that identify and prune/redistribute fixed-location visual sinks while prioritizing image-centric heads to improve grounding. Deeper edits inside attention (Q/K/V or head outputs) can offer more precise control but carry more risk; making them practical requires constrained interventions (bounded strength, automatic tuning, and side-effect monitoring for fluency/style drift) and better module localization to decide where to intervene. Localization is especially important in safety settings, where exposing "critical heads" can increase the attack surface, motivating secure monitoring/localization practices in deployment.

Decoding-level interventions. Decoding intervention can provide a relatively model-agnostic control layer: rather than changing weights, we resample token scores at generation time and can sometimes compose multiple objectives through logit arithmetic. One bottleneck is efficiency: many methods require extra computation (multi-pass contrast, reranking, counterfactual distributions, or reconstruction-based checks), so a future direction is budgeted, risk-aware decoding, using cheap, single-pass signals by default and triggering expensive checks only when hallucination/constraint-violation risk rises, supported by caching and approximate scoring to minimize extra forward passes. Another challenge is stability; most decoding interventions operate token-by-token, providing only soft control. If the steering is too aggressive, for example by heavily penalizing "unlikely" tokens, it can accidentally suppress rare-but-correct words, hurt fluency, or distort calibration. A more robust approach is to make the mixing weights λ_k state-dependent: applying stronger steering when the model is uncertain or the risk of hallucination/constraint violation is high, and backing off when the evidence is clear. In addition, lightweight sequence-level safeguards, such as a small verifier/critic or constrained search, can help check global properties like faithfulness, safety, and structure over longer horizons rather than only locally. Finally, decoding control is fundamentally bounded by what the base model proposes. If the correct continuation has negligible probability under the model, logit shaping has limited leverage. This motivates hybrid pipelines that (i) improve the proposal distribution upstream (e.g., via representation or attention interventions), and (ii) use cheaper but more diagnostic counterfactual signals in multimodal settings (e.g., perturbation- or consistency-based contrasts), while avoiding "double counting" correlated cues from attention-derived diagnostics.

Cross-level composition. As noted in Section 8, the three intervention levels target different computational stages and are potentially composable. However, systematic study of multi-level composition is largely absent. Key open questions include: (i) whether representation + attention + decoding interventions provide additive gains or exhibit diminishing returns; (ii) how to allocate intervention budget across levels (e.g., is it better to invest in a strong representation steer or a moderate representation steer combined with contrastive decoding?); (iii) whether composed interventions introduce unexpected interaction effects (e.g., a representation steer that changes the attention pattern in ways that counteract an attention intervention). Principled composition frameworks, possibly informed by the computational graph formalism of Section 3.1, could become an important direction for improving the practical impact of inference-time control.

Standardized evaluation. The lack of shared benchmarks for steering evaluation (discussed in Section 8) is a barrier to progress. We advocate for the development of a "SteeringBench" that provides: (i) standardized base models with known steering vectors, (ii) evaluation suites covering multiple control objectives (safety, factuality, style, grounding), (iii) mandatory side-effect measurement (e.g., measuring MMLU or general capability degradation alongside the target metric), and (iv) reproducible baselines for each intervention level. Such a benchmark could enable more principled comparison and support adoption.

11 Conclusion

In this survey, we organize the broad literature on inference-time model control into a taxonomy spanning three primary intervention levels. Because work on inference-time control is distributed across disparate research communities, often employing inconsistent terminology and distinct intervention points, systematic comparison remains challenging. To address this, we first identify and clarify intervention loci within the transformer architecture, categorizing methods into representation-level (activations and neurons), attention-level (steering and redistribution), and decoding-level (contrastive scoring) interventions. We use a common notation for the main intervention surfaces (Eqs. 5, 6, 7, 8) to highlight structural parallels across intervention

families, and present cross-method comparisons that suggest complementary strengths: dense representation methods can be low overhead and effective for broad behavioral shifts, attention methods are useful for evidence grounding and routing failures, and decoding methods are practical when control can be expressed through token-score reshaping or candidate reranking. By unifying these approaches, we provide a consistent framework for analyzing when localized modifications affect model behavior across LLMs, VLMs, and diffusion models, suggesting directions for future research on steering foundation models.

12 Ethics Statement

Large language models (LLMs) were used only as writing aids during manuscript preparation. Their use was limited to improving grammar, clarity, wording, and organization. The authors reviewed and revised all LLM-assisted text to ensure that the final manuscript accurately represents the intended meaning, technical claims, and scientific contributions. No LLM-generated text, citation, or numerical claim was accepted without author review.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2204.14198.

Anthropic. Circuit tracing: Revealing computational graphs in language models, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

Dana Arad, Aaron Mueller, and Yonatan Belinkov. SAEs are good for steering – if you select the right features. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10241–10259, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.519. URL https://aclanthology.org/2025.emnlp-main.519/.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.11717.

Seyedarmin Azizi, Erfan Baghaei Potraghloo, and Massoud Pedram. Activation steering for chain-of-thought compression. *arXiv*, 2025. URL https://arxiv.org/abs/2507.04742.

Federico Barbero, Álvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. Why do LLMs attend to the first token? In *Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2504.02732.

Hyemin Boo, Hyoryung Kim, Myungjin Lee, Seunghyeon Lee, Jiyoung Lee, Jang-Hwan Choi, and Hyunsoo Cho. CountSteer: Steering attention for object counting in diffusion models. *arXiv*, 2025. URL https://arxiv.org/abs/2511.11253.

Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. SEGA: Instructing text-to-image models using semantic guidance, 2023. URL https://arxiv.org/abs/2301.12247.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread, Anthropic*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2005.14165.

Weijue Bu, Guan Yuan, and Guixian Zhang. Conscious gaze: Adaptive attention mechanisms for hallucination mitigation in vision-language models. *arXiv*, 2025. URL https://arxiv.org/abs/2512.05546.

Sviatoslav Chalnev, Matthew Siu, and Arthur Conmy. Improving steering vectors by targeting sparse autoencoder features. *arXiv*, 2024. URL https://arxiv.org/abs/2411.02193.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2301.13826.

Chubin Chen, Jiashu Zhu, Xiaokun Feng, Nisha Huang, Chen Zhu, Meiqi Wu, Fangyuan Mao, Jiahong Wu, Xiangxiang Chu, and Xiu Li. Stochastic self-guidance for training-free enhancement of diffusion models, 2026. URL https://arxiv.org/abs/2508.12880.

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, and Xuming Hu. ICT: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4209–4221, 2025a. URL https://arxiv.org/abs/2411.15268.

Wei Chen, Xin Yan, Bin Wen, Fan Yang, Tingting Gao, Di Zhang, and Long Chen. Decoupling contrastive decoding: Robust hallucination mitigation in multimodal large language models, 2025b. URL https://arxiv.org/abs/2504.08809.

Seonglae Cho, Zekun Wu, and Adriano Koshiyama. CorrSteer: Generation-time LLM steering via correlated sparse autoencoder features. *arXiv*, 2025. URL https://arxiv.org/abs/2508.12535.

Xu Chu, Xinrong Chen, Guanyu Wang, Zhijie Tan, Kui Huang, Wenyu Lv, Tong Mo, and Weiping Li. Qwen look again: Guiding vision-language reasoning models to re-attention visual information. *arXiv*, 2025. URL https://arxiv.org/abs/2505.23558.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. DoLa: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.03883.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.08600.

Souvik Das, Lifeng Jin, Linfeng Song, Haitao Mi, Baolin Peng, and Dong Yu. Entropy guided extrapolative decoding to improve factuality in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 6589–6600, 2025. URL https://aclanthology.org/2025.coling-main.439/.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1912.02164.

Dave Epstein, Allan Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.00986.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14303–14312, 2024. URL https://arxiv.org/abs/2403.14003.

Mehrdad Fazli, Bowen Wei, Ahmet Sari, and Ziwei Zhu. Mitigating hallucination in large vision-language models via adaptive attention calibration, 2025. URL https://arxiv.org/abs/2505.21472.

Xiachong Feng, Liang Zhao, Weihong Zhong, Yichong Huang, Yuxuan Gu, Lingpeng Kong, Xiaocheng Feng, and Bing Qin. PERSONA: Dynamic and compositional inference-time personality control via activation vector algebra, 2026a. URL https://arxiv.org/abs/2602.15669.

Zijian Feng, Tianjiao Li, Zixiao Zhu, Hanzhang Zhou, Junlang Qian, Li Zhang, Jia Jim Deryl Chua, Lee Onn Mak, Gee Wah Ng, and Kezhi Mao. Fine-grained activation steering: Steering less, achieving more, 2026b. URL https://arxiv.org/abs/2602.04428.

Jaden Fried Fiotto-Kaufman, Alexander Russell Loftus, Eric Todd, Jannik Brinkmann, Koyena Pal, Dmitrii Troitskii, Michael Ripa, Adam Belfki, Can Rager, Caden Juang, Aaron Mueller, Samuel Marks, Arnab Sen Sharma, Francesca Lucchetti, Nikhil Prakash, Carla E. Brodley, Arjun Guha, Jonathan Bell, Byron C. Wallace, and David Bau. NNsight and NDIF: Democratizing access to open-weight foundation model internals. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=MxbEiFRf39.

Tatiana Gaintseva, Andreea-Maria Oncescu, Chengcheng Ma, Ziquan Liu, Martin Benning, Gregory Slabaugh, Jiankang Deng, and Ismail Elezi. CASteer: Cross-attention steering for controllable concept erasure. *arXiv*, 2025. URL https://arxiv.org/abs/2503.09630.

Woody Haosheng Gan, Deqing Fu, Julian Asilis, Ollie Liu, Dani Yogatama, Vatsal Sharan, Robin Jia, and Willie Neiswanger. Textual steering vectors can improve visual understanding in multimodal large language models. *arXiv*, 2025. URL https://arxiv.org/abs/2505.14071.

Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2303.07345.

Yuyao Ge, Shenghua Liu, Yiwei Wang, Lingrui Mei, Baolong Bi, Xuanshan Zhou, Jiayu Yao, Jiafeng Guo, and Xueqi Cheng. Focusing by contrastive attention: Enhancing VLMs' visual reasoning. *arXiv*, 2025. URL https://arxiv.org/abs/2509.06461.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.10781.

Zhuohan Gu, Jiayi Yao, Kuntai Du, and Junchen Jiang. LLMSteer: Improving long-context LLM inference by steering attention on reused contexts. *arXiv*, 2024. URL https://arxiv.org/abs/2411.13009.

Vitoria Guardieiro, Avishree Khare, Adam Stein, and Eric Wong. Instruction following by principled boosting attention of large language models. *arXiv*, 2025. URL https://arxiv.org/abs/2506.13734.

Zirui He, Mingyu Jin, Bo Shen, Ali Payani, Yongfeng Zhang, and Mengnan Du. SAE-SSV: Supervised steering in sparse representation spaces for reliable control of language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025a. URL https://aclanthology.org/2025.emnlp-main.112/.

Zirui He, Haiyan Zhao, Yiran Qiao, Fan Yang, Ali Payani, Jing Ma, and Mengnan Du. SAIF: A sparse autoencoder framework for interpreting and steering instruction following of language models. *arXiv*, 2025b. URL https://arxiv.org/abs/2502.11356.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2208.01626.

Brandon Hsu, Daniel Beaglehole, Adityanarayanan Radhakrishnan, and Mikhail Belkin. Contextual linear activation steering of language models, 2026. URL https://arxiv.org/abs/2604.24693.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

James Y. Huang, Sailik Sengupta, Daniele Bonadiman, Yi-An Lai, Arshit Gupta, Nikolaos Pappas, Saab Mansour, Katrin Kirchhoff, and Dan Roth. DeAL: Decoding-time alignment for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 26280–26300, 2025. URL https://aclanthology.org/2025.acl-long.1274/.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2311.17911.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models, 2024. URL https://arxiv.org/abs/2408.02032.

Ole Jorgensen, Dylan Cope, Nandi Schoots, and Murray Shanahan. Improving activation steering in language models with mean-centring. *arXiv*, 2023. URL https://arxiv.org/abs/2312.03813.

Mingi Jung, Saehyung Lee, Eunji Kim, and Sungroh Yoon. Visual attention never fades: Selective progressive attention recalibration for detailed image captioning in multimodal large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL https://proceedings.mlr.press/v267/jung25c.html.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2503.03321.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. CTRL: A conditional transformer language model for controllable generation. *arXiv*, 2019. URL https://arxiv.org/abs/1909.05858.

Jieun Kim, Jinmyeong Kim, Yoonji Kim, and Sung-Bae Cho. Fuzzy contrastive decoding to alleviate object hallucination in large vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Kim_Fuzzy_Contrastive_Decoding_to_Alleviate_Object_Hallucination_in_Large_Vision-Language_ICCV_2025_paper.html.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021. URL https://aclanthology.org/2021.findings-emnlp.424/.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2303.13516.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering, 2024. URL https://arxiv.org/abs/2409.05907.

Byung Hyun Lee, Sungjin Lim, and Se Young Chun. Localized concept erasure for text-to-image diffusion models using training-free gated low-rank adaptation, 2025. URL https://arxiv.org/abs/2503.12356.

Jihoon Lee and Min Song. Retrieval visual contrastive decoding to mitigate object hallucinations in large vision-language models, 2025. URL https://arxiv.org/abs/2505.20569.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2311.16922.

Jiaming Li, Jiacheng Zhang, Zequn Jie, Lin Ma, and Guanbin Li. Cross-modal attention calibration for LVLM hallucination mitigation. *arXiv*, 2025a. URL https://arxiv.org/abs/2501.01926.

Jiaqian Li, Yanshu Li, and Kuan-Hao Huang. Steering vector fields for context-aware inference-time control in large language models, 2026. URL https://arxiv.org/abs/2602.01654.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, 2023a. URL https://arxiv.org/abs/2306.03341.

Qiming Li, Zekai Ye, Xiaocheng Feng, Weihong Zhong, Libo Qin, Ruihan Chen, Baohang Li, Kui Jiang, Yaowei Wang, Ting Liu, and Bing Qin. CAI: Caption-sensitive attention intervention for mitigating object hallucination in large vision-language models, 2025b. URL https://arxiv.org/abs/2506.23590.

Xian Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023b. URL https://aclanthology.org/2023.acl-long.687/.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021. URL https://aclanthology.org/2021.acl-long.353/.

Yichen Li, Zhiting Fan, Ruizhe Chen, Xiaotang Gai, Luqi Gong, Yan Zhang, and Zuozhu Liu. FairSteer: Inference time debiasing for LLMs with dynamic activation steering, 2025c. URL https://arxiv.org/abs/2504.14492.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021. URL https://aclanthology.org/2021.acl-long.522/.

Alisa Liu, Xiaochuang Han, Yizhong Wang, Yulia Tsvetkov, Yejin Choi, and Noah A. Smith. Tuning language models by proxy. *arXiv*, 2024a. URL https://arxiv.org/abs/2401.08565.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2304.08485.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. In-context vectors: Making in context learning more effective and controllable through latent space steering. In *Proceedings of the 41st International Conference on Machine Learning*, 2024b. URL https://proceedings.mlr.press/v235/liu24bx.html.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. Reducing hallucinations in vision-language models via latent space steering. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.15778.

Oscar Mañas, Pierluca D'Oro, Koustuv Sinha, Adriana Romero-Soriano, Michal Drozdzal, and Aishwarya Agrawal. Controlling multimodal LLMs via reward-guided decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1391–1401, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Manas_Controlling_Multimodal_LLMs_via_Reward-guided_Decoding_ICCV_2025_paper.html.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2310.06824.

Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Zeyu Wang, Zhengxin Zhang, Rae Ying Yee Wong, Alan Zhu, Lijie Yang, Xiaoxiang Shi, Chunan Shi, Zhuoming Chen, Daiyaan Arfeen, Reyna Abhyankar, and Zhihao Jia. SpecInfer: Accelerating generative large language model serving with tree-based speculative inference and verification. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, pp. 932–949, 2024. URL https://arxiv.org/abs/2305.09781.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, 2019. URL https://arxiv.org/abs/1905.10650.

Erik Miehling, Karthikeyan Natesan Ramamurthy, Praveen Venkateswaran, Irene Ko, Pierre Dognin, Moninder Singh, Tejaswini Pedapati, Avinash Balakrishnan, Matthew Riemer, Dennis Wei, Inge Vejsbjerg, Elizabeth M. Daly, and Kush R. Varshney. AI Steerability 360: A toolkit for steering large language models, 2026. URL https://arxiv.org/abs/2603.07837.

Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. FreeControl: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2312.07536.

Krishna Kanth Nakka, Xue Jiang, Dmitrii Usynin, and Xuebing Zhou. PII jailbreaking in LLMs via activation steering reveals personal information leakage, 2025. URL https://arxiv.org/abs/2507.02332.

Duy Nguyen, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. GrAInS: Gradient-based attribution for inference-time steering of LLMs and VLMs, 2025. URL https://arxiv.org/abs/2507.18043.

Kyle O'Brien, David Majercak, Xavier Fernandes, Richard Edgar, Blake Bullwinkel, Jingya Chen, Harsha Nori, Dean Carignan, Eric Horvitz, and Forough Poursabzi-Sangdeh. Steering language model refusal with sparse autoencoders. *arXiv*, 2024. URL https://arxiv.org/abs/2411.11296.

Sean O'Brien and Mike Lewis. Contrastive decoding improves reasoning in large language models. *arXiv*, 2023. URL https://arxiv.org/abs/2309.09117.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2203.02155.

Jayneel Parekh, Pegah Khayatan, Mustafa Shukor, Arnaud Dapogny, Alasdair Newson, and Matthieu Cord. Learning to steer: Input-dependent steering for multimodal LLMs. *arXiv*, 2025. URL https://arxiv.org/abs/2508.12815.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Yeji Park, Deokyeong Lee, Junsuk Choe, and Buru Chang. ConVis: Contrastive decoding with hallucination visualization for mitigating hallucinations in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. URL https://arxiv.org/abs/2408.13906.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.03693.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021. URL https://arxiv.org/abs/2103.00020.

Pooyan Rahmanzadehgervi, Hung Huy Nguyen, Rosanne Liu, Long Mai, and Anh Totti Nguyen. TAB: Transformer attention bottlenecks enable user intervention and debugging in vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2412.18675.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 15504–15522, 2024. URL https://aclanthology.org/2024.acl-long.828/.

Pau Rodriguez, Arno Blaas, Michal Klein, Luca Zappella, Nicholas Apostoloff, Marco Cuturi, and Xavier Suau. Controlling language and diffusion models by transporting activations. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.23054.

Kenji Sahay, Snigdha Pandya, Rohan Nagale, Anna Lin, Shikhar Shiromani, Parham Sharaf, Kevin Zhu, and Sunishchal Dev. COMPASS: Context-modulated pid attention steering system for hallucination mitigation. *arXiv*, 2025. URL https://arxiv.org/abs/2511.14776.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv*, 2024. URL https://arxiv.org/abs/2402.07927.

Sreetama Sarkar, Yue Che, Alex Gavin, Peter A. Beerel, and Souvik Kundu. Mitigating hallucinations in vision-language models through image-guided head suppression, 2025. URL https://arxiv.org/abs/2505.16411.

Daniel Scalena, Gabriele Sarti, and Malvina Nissim. Multi-property steering of large language models with dynamic activation composition. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 577–603, 2024. URL https://aclanthology.org/2024.blackboxnlp-1.34/.

Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models, 2025. URL https://arxiv.org/abs/2501.06848.

Anushka Sivakumar, Andrew Zhang, Zaber Ibn Abdul Hakim, and Chris Thomas. SteerVLM: Robust model control through lightweight activation steering for vision language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23640–23665, 2025. URL https://aclanthology.org/2025.findings-emnlp.1285/.

Samuel Soo, Chen Guang, Chandrasekaran Balaganesh, Wesley Teng, Tan Guoxian, and Yan Ming. Interpretable steering of large language models with feature guided activation additions. *arXiv*, 2025. URL https://arxiv.org/abs/2501.09929.

Niklas Stoehr, Kevin Du, Vésteinn Snæbjarnarson, Robert West, Ryan Cotterell, and Aaron Schein. Activation scaling for steering and interpreting language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8189–8200, 2024. URL https://aclanthology.org/2024.findings-emnlp.479/.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering, 2024. URL https://arxiv.org/abs/2410.12877.

Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Qing Li, and Zhaoxiang Zhang. Activation steering decoding: Mitigating hallucination in large vision-language models through bidirectional hidden state intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 12964–12974, 2025. URL https://aclanthology.org/2025.acl-long.634/.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. A contrastive framework for neural text generation. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2202.06417.

Jiuding Sun, Sidharth Baskaran, Zhengxuan Wu, Michael Sklar, Christopher Potts, and Atticus Geiger. HyperSteer: Activation steering at scale with hypernetworks, 2025. URL https://arxiv.org/abs/2506.03292.

Yi Sun, Xinhao Zhong, Hongyan Li, Yimin Zhou, Junhao Li, Bin Chen, and Xuan Wang. ActErase: A training-free paradigm for precise concept erasure via activation redirection. *arXiv*, 2026. URL https://arxiv.org/abs/2601.00267.

Daniel Tan, David Chanin, Aengus Lynch, Dimitrios Kanoulas, Brooks Paige, Adria Garriga-Alonso, and Robert Kirk. Analyzing the generalization and reliability of steering vectors. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2407.12404.

Feilong Tang, Chengzhi Liu, Zhongxing Xu, Ming Hu, Zelin Peng, Zhiwei Yang, Jionglong Su, Minquan Lin, Yifan Peng, Xuelian Cheng, Imran Razzak, and Zongyuan Ge. Seeing far and clearly: Mitigating hallucinations in MLLMs with attention causal decoding, 2025. URL https://arxiv.org/abs/2505.16652.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread, Anthropic*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Max Torop, Aria Masoomi, Masih Eskandar, and Jennifer G. Dy. DISCO: Disentangled communication steering for large language models. *arXiv*, 2025. URL https://arxiv.org/abs/2509.16820.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv*, 2023. URL https://arxiv.org/abs/2308.10248.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1706.03762.

Praveen Venkateswaran and Danish Contractor. Spotlight your instructions: Instruction-following with dynamic attention steering. In *Proceedings of the 2026 Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3752–3770, 2026. URL https://aclanthology.org/2026.eacl-long.174/.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. URL https://aclanthology.org/P19-1580/.

Hieu M. Vu and Tan M. Nguyen. Angular steering: Behavior control via rotation in activation space, 2025. URL https://arxiv.org/abs/2510.26243.

Han Wang, Gang Wang, and Huan Zhang. Steering away from harm: An adaptive approach to defending vision language model against jailbreaks, 2024a. URL https://arxiv.org/abs/2411.16721.

Weixuan Wang, Jingyuan Yang, and Wei Peng. Semantics-adaptive activation intervention for LLMs via dynamic steering vectors, 2024b. URL https://arxiv.org/abs/2410.12299.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15840–15853, 2024c. URL https://aclanthology.org/2024.findings-acl.937/.

Zihu Wang, Boxun Xu, Yuxuan Xia, and Peng Li. VEGAS: Mitigating hallucinations in large vision-language models via vision-encoder attention guided adaptive steering, 2025. URL https://arxiv.org/abs/2512.12089.

Sangmin Woo, Jaehyuk Jang, Donguk Kim, Yubin Choi, and Changick Kim. RITUAL: Random image transformations as a universal anti-hallucination lever in large vision-language models. *arXiv*, 2024. URL https://arxiv.org/abs/2405.17821.

Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 1927–1951, 2025. URL https://aclanthology.org/2025.findings-acl.99/.

Lyucheng Wu, Mengru Wang, Ziwen Xu, Tri Cao, Nay Oo, Bryan Hooi, and Shumin Deng. Automating steering for safe multimodal large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 792–814. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.emnlp-main.41. URL https://aclanthology.org/2025.emnlp-main.41/.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67035–67080. PMLR, 2025b. URL https://proceedings.mlr.press/v267/wu25a.html.

Zhengxuan Wu, Qinan Yu, Aryaman Arora, Christopher D. Manning, and Christopher Potts. Improved representation steering for language models, 2025c. URL https://arxiv.org/abs/2505.20809.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.17453.

Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion, 2023. URL https://arxiv.org/abs/2307.10816.

Jingyuan Yang, Rongjun Li, Weixuan Wang, Ziyu Zhou, Zhiyong Feng, and Wei Peng. LF-Steering: Latent feature activation steering for enhancing semantic consistency in large language models. *arXiv*, 2025. URL https://arxiv.org/abs/2501.11036.

Kevin Yang and Dan Klein. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, 2021. URL https://aclanthology.org/2021.naacl-main.276/.

Jianghao Yin, Qin Chen, Kedi Chen, Jie Zhou, Xingjiao Wu, and Liang He. Dynamic multimodal activation steering for hallucination mitigation in large vision-language models, 2026. URL https://arxiv.org/abs/2602.21704.

Zejia You, Chunyuan Deng, and Hanjie Chen. Spherical steering: Geometry-aware activation rotation for language models, 2026. URL https://arxiv.org/abs/2602.08169.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Celine Lin. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. *arXiv*, 2024. URL https://arxiv.org/abs/2406.15765.

Xiyu Zeng, Siyuan Liang, Liming Lu, Haotian Zhu, Enguang Liu, Jisheng Dang, Yongbin Zhou, and Shuchao Pang. SafeSteer: Adaptive subspace steering for efficient jailbreak defense in vision-language models, 2025. URL https://arxiv.org/abs/2509.21400.

Li-Ming Zhan, Bo Liu, Chengqiang Xie, Jiannong Cao, and Xiao-Ming Wu. REAL: Reading out transformer activations for precise localization in language model steering, 2025. URL https://arxiv.org/abs/2506.08359.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.16042.

Hongxiang Zhang, Hao Chen, Muhao Chen, and Tianyi Zhang. Active layer-contrastive decoding reduces hallucination in large language model generation, 2025a. URL https://arxiv.org/abs/2505.23657.

Jun Zhang, Jue Wang, Huan Li, Lidan Shou, Ke Chen, Gang Chen, and Sharad Mehrotra. Draft & verify: Lossless large language model acceleration via self-speculative decoding. *arXiv*, 2023. URL https://arxiv.org/abs/2309.08168.

Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. Tell your model where to attend: Post-hoc attention steering for LLMs. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2311.02262.

Xiaofeng Zhang, Yihao Quan, Chen Shen, Chaochen Gu, Xiaosong Yuan, Shaotian Yan, Jiawei Cao, Hao Cheng, Kaijie Wu, and Jieping Ye. Shallow focus, deep fixes: Enhancing shallow layers vision attention sinks to alleviate hallucination in LVLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3512–3534, 2025b. URL https://aclanthology.org/2025.emnlp-main.174/.

Yunxiang Zhang, Muhammad Khalifa, Lechen Zhang, Xin Liu, Ayoung Lee, Xinliang Frederick Zhang, Farima Fatahi Bayat, and Lu Wang. Logit arithmetic elicits long reasoning capabilities without training. *arXiv*, 2025c. URL https://arxiv.org/abs/2507.12759.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19781–19791, 2025d. URL https://arxiv.org/abs/2411.18620.

Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. Steering knowledge selection behaviours in LLMs via SAE-based representation engineering, 2024. URL https://arxiv.org/abs/2410.15999.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, and Yongbin Li. On the role of attention heads in large language model safety. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.13708.

Xingyu Zhu, Beier Zhu, Shuo Wang, Junfeng Fang, Kesen Zhao, Hanwang Zhang, and Xiangnan He. Principled steering via null-space projection for jailbreak defense in vision-language models, 2026. URL https://arxiv.org/abs/2603.22094.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv*, 2023. URL https://arxiv.org/abs/2310.01405.