Representation Intervention in Vision Language Models: A Survey

Anonymous ACL submission

Abstract

Vision Language Models (VLMs) have achieved strong multimodal understanding by jointly modeling visual and linguistic information within unified representation spaces. Recently, representation intervention has emerged as a central paradigm for analyzing and controlling VLMs, where targeted modifications to internal representations enable behavior modulation without full retraining. In this survey, we review recent advances in representation intervention for VLMs, focusing on interventions applied to intermediate activations and model weights. We propose a unified taxonomy that categorizes existing methods into activation-level and weight-level interventions according to their intervention targets under modern decoder-only VLM architectures. We further provide a mathematical formulation that abstracts representation intervention as a controllable transformation over activation or weight targets. Finally, we discuss key application scenarios, major challenges, and future research directions, including hallucination mitigation, safety alignment, controllable cross-modal interactions, and related emerging problems.

1 Introduction

Representation Engineering has emerged as a transformative paradigm for understanding and regulating the behavior of large language models (LLMs) (Liu et al., 2025). At the core of this paradigm is representation intervention, which treats internal representations as direct control targets rather than opaque products of parameter optimization. By manipulating these internals through activation level perturbations during inference or weight level editing, researchers can induce controlled behavioral changes without full retraining. Prior studies show that this mechanism is effective for regulating generation attributes, suppressing undesired behaviors, and enhancing specific capabilities in text-only models (Bartoszcze et al., 2025).

With the rapid development of VLMs, representation intervention is being extended from textual models to multimodal systems (Kaduri et al., 2025). This transition is facilitated by the architectural convergence of modern VLMs toward decoder-only structures. Unlike earlier encoder-decoder models that relied on complex alignment mechanisms (Radford et al., 2021), modern VLMs convert images into sequences of visual tokens and incorporate them directly into autoregressive decoding. This compatibility allows many intervention techniques developed for LLMs to be adapted to VLMs. Recent studies show that intervening on visual tokens, multimodal embeddings, or shared activations can modulate visual grounding, cross-modal alignment, and reasoning behaviors (Wang et al., 2025c).

However, applying representation intervention to VLMs is not a trivial extension of text-only methods. Despite architectural similarities, VLMs introduce distinct functional dynamics. First, cross-modal perturbations may occur, where interventions on one modality propagate to the other in non-linear and unpredictable ways (Wenderoth et al., 2025). Second, interventions often operate in entangled representation spaces that jointly encode visual and linguistic semantics, which complicates precise control of modality-specific behaviors (Alonso et al., 2025). Third, modality-sensitive routing emerges in attention and feed-forward layers, where transformations may selectively amplify or suppress modality-dependent features (Liao et al., 2025). These characteristics suggest that effective intervention in VLMs must be modality aware.

Despite growing interest in this topic, a systematic survey centered on representation intervention for VLMs remains absent. Existing surveys either focus on text-based models (Wehner et al., 2025),

with limited consideration of visual modalities, or emphasize general VLM interpretability (Dang et al., 2024) without focusing on active control mechanisms. Moreover, many prior reviews remain tied to earlier encoder-decoder paradigms and do not fully capture the properties of modern decoder-only systems.

To bridge this gap, this survey reviews representation intervention methods for VLMs, focusing on studies published after 2023. We propose a unified taxonomy organized along three core dimensions: (1) *Intervention Scope*, which distinguishes transient activation level operations from persistent weight level modifications; (2) *Representation Granularity*, which characterizes interventions at the token, embedding, or layer level; and (3) *Application Scenario*, which covers safety, robustness, alignment, and agentic control.

In summary, this survey makes the following contributions:

- We propose a unified taxonomy of VLM representation intervention methods based on intervention scope, representation granularity, and application domain.
- We provide a mathematical formulation of representation intervention in VLMs, which abstracts diverse methods as controllable transformations over activation or weight targets.
- We analyze key methodological gaps and outline future research directions for bridging activation-level and weight-level model control.

2 Preliminaries

2.1 Representation Intervention on VLMs

Representation Intervention on VLMs refers to a class of methods that explicitly intervene in internal representation spaces to regulate, analyze, or control model behavior. Rather than relying solely on input prompting or end-to-end retraining, these methods operate by applying targeted modifications to intermediate activations or selected model parameters associated with visual tokens, textual tokens, or their joint multimodal representations. By directly acting on internal representations, representation intervention enables fine-grained and mechanism-oriented control over multimodal reasoning, alignment, robustness, and safety behaviors, while preserving the overall pretrained structure and functionality of the model.

A defining characteristic of representation intervention is its emphasis on efficient and controllable manipulation of model behavior. Across both activation-level and weight-level settings, existing methods commonly exhibit the following properties:

- **Training-Free Deployment.** Many representation intervention methods can be applied at inference time or through lightweight parameter edits, avoiding full-scale retraining and reducing data dependency.
- **Plug-and-Play Integration.** Representation interventions are often modular operations that can be inserted into existing VLMs, targeting specific tokens, layers, residual streams, or submodules without changing model architecture.
- **Reversibility and Controllability.** Interventions on activation-level can be dynamically enabled or disabled, while many weight-level edits can be reverted, allowing precise control over intervention strength and scope.

Together, these properties make representation intervention an efficient and resource-economical paradigm for controlling VLMs, offering a practical balance between behavioral flexibility and computational efficiency.

2.2 Mathematical Definition

Definition 1. Representation Steering for Vision Language Models. Let $\mathcal{M}$ denote a vision language model that takes a textual input X^t and a visual input X^v as input. We define the internal intervention target as τ, which can be selected from the activation target space $\mathcal{T}_{\text{act}}$ or the weight target space $\mathcal{T}_{\text{weight}}$:

$$\tau \in \mathcal{T}_{\text{act}} \cup \mathcal{T}_{\text{weight}}. \tag{1}$$

A representation steering method applies a control method $\mathcal{C}$ to the selected target τ and produces a controlled target τ':

$$\tau' = \mathcal{I}_{\mathcal{C}}(\tau; \mathcal{M}, X^t, X^v, \Omega), \tag{2}$$

where $\mathcal{I}_{\mathcal{C}}$ denotes the intervention operator instantiated by the control method $\mathcal{C}$, and Ω denotes the intervention configuration, such as the selected layer, module, token position, scaling coefficient, editing direction, or optimization objective.

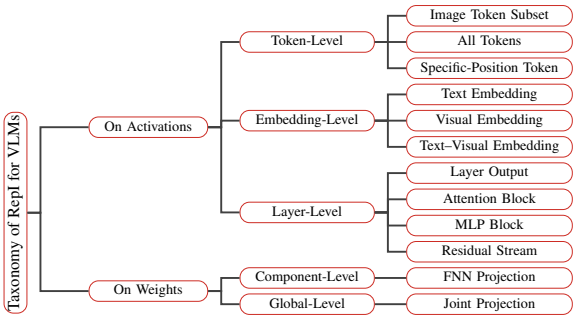

Figure 1: A taxonomy of representation intervention for VLMs, aligned with activation-level and weight-level intervention mechanisms. RepI is short for Representation Intervention.

The controlled target τ' is then injected back into the model computation:

$$\mathcal{M}' = \mathcal{M}[\tau \leftarrow \tau'], \tag{3}$$

which yields the steered output:

$$Y' = \mathcal{M}'(X^t, X^v). \tag{4}$$

In this formulation, representation steering aims to modify the internal computation of a vision-language model by controlling a specific representation or parameter component. For activation-based methods, τ may correspond to visual tokens, visual textual tokens, embeddings, layer outputs, attention blocks, or MLP blocks. For weight-based methods, τ may correspond to projection matrices or other parameter components. Different control methods instantiate $\mathcal{I}_\mathcal{C}$ in different forms. For readability, Table 1 summarizes representative instantiations using semantic descriptions, while Appendix A.1 provides a more abstract mathematical formulation.

2.3 Taxonomy

Figure 1 presents a unified taxonomy of representation intervention for VLMs, which organizes existing methods according to their intervention target and representation granularity. We first distinguish between activation-level and weight-level representation intervention, depending on whether interventions operate on intermediate activations during inference or directly modify model parameters, each level is then categorized into different target. Table 1 provides a comprehensive overview of representative methods under each category, summarizing their intervention targets, control method and application domain.

3 Representation Intervention on Activations

Representation Intervention on Activations refers to methods that control VLMs behavior by intervening on intermediate activations during inference, without modifying model parameters. These approaches manipulate token-level, embedding-level, or layer-level activations to regulate multimodal information flow, enabling training-free and fine-grained control over reasoning, hallucination, and safety while preserving pretrained knowledge.

3.1 Token-Level

Token-level steering refers to representation intervention methods that intervene on model activations conditioned on discrete token positions or token identities within the multimodal sequence. In VLMs, such interventions explicitly leverage the sequential structure of text tokens or image tokens and are often tightly coupled with the decoding process or attention allocation across tokens.

Visual Tokens Visual tokens are discrete representations extracted from images that encode localized visual features, enabling structured processing in VLMs. Representation intervention on visual tokens selectively intervenes on a small subset of tokens that carry critical semantic or causal information, based on the observation that model behavior is often dominated by a limited number of salient visual tokens.

A study by Li et al. (2025c) investigates visual hallucination in large vision–language models and proposes Visual Information Steering (VISTA), a method that selectively modulates the contribution of visual tokens during inference. Instead of suppressing the entire visual stream, VISTA identifies visually grounded tokens that are causally related to hallucinated outputs and adjusts their influence at the logit level, effectively steering the model by acting on a subset of visual evidence. Serra et al. (2024) provides an in-depth analysis of localized image–text communication in vision–language models. The authors show that semantic information from images is often funneled through a small number of critical visual tokens or aggregation tokens. By perturbing or reweighting these specific tokens, the model's interpretation of the image can be significantly altered, demonstrating that targeted intervention on a minimal visual token subset is sufficient for behavior control. Phute and Balakrishnan (2025) and its extension Balakrish-

Category	Level	Target	Representative Work	Control Function	Control Method	Application Domain
RepE on Activations	Token	Visual Tokens	(Li et al., 2025c)	DSC, NC	Token Contribution Reweighting	HM
			(Serra et al., 2024)	AC	Token-level Representation Perturbation	HM, VGE
			(Phute and Balakrishnan, 2025)	DSC	Input-induced Token Subspace Activation	SRI, BC
			(Balakrishnan and Phute, 2025)	DSC, PC	Token Contribution Reweighting	SRI, BC
			(Sivakumar et al., 2025)	DSC, GC, MC	Token-level Representation Perturbation	HM, SRI
		Visual–Textual Tokens	(Sarkar et al., 2025)	MC	Cross-modal Attention Head Editing	HM, VGE, SRI, IBC, GMRE
			(Ye et al., 2025a)	MC, GC	Selective Token Suppression	VGE, IBC, GMRE
			(Sun et al., 2025)	MC,GC	Token Pruning in KV Cache	HM, VGE, SRI, IBC
			(Leng et al., 2024)	AC	Token Contribution Reweighting	HM, VGE, SRI
		Specific–Position Token	(Huang et al., 2024)	AC, GC	Query Token Intervention	HM, VGE, SRI
			(Wang et al., 2024)	AC	Selective Token Suppression	HM, SRI, GMRE
			(Luo et al., 2025)	GC	Attention Sink Mitigation	IBC, GMRE
	Embedding	Text Embedding	(Zhang et al., 2026)	MC	Objective-driven Embedding Steering	ABC, GMRE
			(Yang et al., 2025b)	GC	Embedding Space Analysis	GMRE
			(Li et al., 2025f)	DSR	Contrastive Embedding Manipulation	HM, VGE
			(Masry et al., 2025)	GC, MC	Task-aligned Embedding Refinement	VGE, GMRE
		Visual Embedding	(Yang et al., 2025a)	MC, GC	Embedding Space Analysis	GMRE
			(Dai et al., 2025)	DSC	Human-aligned Embedding Refinement	VGE, GMRE
			(Lee et al., 2025)	AC	Visual Latent Space Editing	IBC, GMRE
	Layer	Layer Output	(Su et al., 2025)	DSC, AC	Layer Output Perturbation	HM
			(Li et al., 2025c)	DSC, NC	Contrastive Feature Amplification	HM
			(Golovanevsky et al., 2025)	DSC	Attention-induced Residual Control	IBC, ABC
		Attention Block	(Liu et al., 2024)	MC	Attention Reweighting	HM
			(Ye et al., 2025b)	DSC, GC	Structured Attention Alignment	HM
			(Li et al., 2025b)	DSC, GC	Selective Attention Head Editing	HM
			(Wang et al., 2025b)	DSC	Attention-induced Residual Control	SRI
		MLP Block	(Kim et al., 2025b)	GC, PC, MC	MLP Activation Attribution	HM, VGE
			(Duan et al., 2025b)	PC, GC	Policy-oriented MLP Steering	HM, VGE
			(Häon et al., 2025)	DSC, AC	MLP Gating and Modulation	ABC, IBC
			(Li et al., 2025a)	MC	Cross-layer MLP Intervention	SRI, IBC
RepE on Weights	Component	FNN Projection	(Chen et al., 2025)	GC, DSC	Editing via Visual Attribution	SRI
			(Shi et al., 2025)	GC	Co-Updating Dual Modalities	SRI
			(Zhou et al., 2025)	AC	Updating Multi-Granularity Neurons	SRI
	Global	Joint Projection	(Yan et al., 2025)	MC	Joint Scaling of Attention and MLP Weights	SRI, IBC
			(D'Incà et al., 2025)	MC	Joint Scaling of Attention and MLP Weights	SRI, IBC

Table 1: Expanded taxonomy of representation intervention methods in vision–language models. Control functions include DSC (Directional Shift Control), MC (Multiplicative Control), NC (Normalized Control), PC (Projection Control), AC (Affine Control) and GC (Gated Control). Detailed description for each control functions can be found in Appendix A.1. Application domain includes HM (Hallucination Mitigation), VGE (Visual Grounding Enhancement), SRI (Safety and Reliability Improvement), IBC (Interpretability-oriented Behavior Control), GMRE (General Multimodal Reasoning Enhancement).

nan and Phute (2025) explore visual-input-based steering mechanisms, where carefully designed visual prompts or steering images redirect model outputs. Although these methods operate at the input level, their effectiveness relies on activating or emphasizing particular regions of the visual token space, rather than uniformly affecting all visual tokens. This implicitly aligns with subset-based visual token steering. Sivakumar et al. (2025) proposes a lightweight activation steering framework for vision–language models that enables robust control over multimodal outputs. While the method is formulated at the activation level, the learned steering directions are shown to concentrate on a limited set of visually salient tokens, suggesting that selective modulation of critical visual tokens underlies effective multimodal steering.

Visual-Textual Token Some studies intervene at the cross-modal attention head level by treating the image-to-text information flow as a controllable representation. These methods suppress or reweight selected cross-modal attention heads to encourage text tokens to rely more stably on visual evidence. For example, Sarkar et al. (2025) suppresses heads with weak image text alignment during inference to reduce language prior-driven hallucination. Ye et al. (2025b) further introduces language-specific head intervention to improve cross-modal alignment in multilingual VLMs. In addition, Devils Jiang et al. (2025) shows that inter-

vening on a small set of critical cross-modal heads in middle layers can substantially alter how visual information is propagated to text generation.

Another line of work performs representation intervention by editing which visual tokens participate in image-text interactions. These methods select or prune visual tokens based on their relevance to text tokens, thereby modifying the effective cross-modal representation space. Ye et al. (2025a) proposes text-guided adaptive visual token pruning during decoding, preserving only text-relevant image tokens in later layers. Similarly, Sun et al. (2025) prunes image tokens in the KV cache during inference to reduce hallucination.

Some approaches intervene during decoding by modifying token-level representations or output distributions conditioned on image-text interactions. Leng et al. (2024) introduces visual contrastive decoding, which contrasts outputs from original and perturbed images to suppress language-dominant tokens lacking visual support. Li et al. (2025c) further reinforces visually grounded tokens by augmenting token logits during generation. These methods explicitly edit how image tokens influence text token selection at decoding time. A study by Bhaila et al. (2025) investigates how specific visual tokens contribute to memorization and disclosure in VLMs and performs targeted unlearning by weakening their cross-modal connections to text outputs.

Specific-position Token Specific-position token steering refers to representation-intervention methods that act on model activations at a predefined token position within the multimodal sequence. Unlike all-token steering, this paradigm targets a single position whose representation plays a dominant role in subsequent generation.

A large body of decoding-based methods fall into this category. Leng et al. (2024) introduces visual contrastive decoding by modifying the next-token distribution at each decoding step. Wang et al. (2024) performs contrastive modulation over the next-token logits to suppress hallucinated outputs. Im et al. (2026) extends contrastive decoding with self-augmented visual experts operating on the current query token. Kim et al. (2025a) uses fuzzy membership functions over next-token logits to steer token selection.

Meanwhile, sink token (Barbero et al., 2025) steering targets a special subset of tokens that disproportionately absorb attention mass, commonly referred to as attention sink tokens. In VLMs, study of Luo et al. (2025) analyzes the formation of visual sink tokens and their role in information flow across layers. Such sink tokens often emerge among image tokens and can dominate cross-modal attention, leading to degraded reasoning or hallucination.

3.2 Embedding-Level

Embedding-level steering refers to representation-level interventions that directly operate on continuous embedding spaces in VLMs. Instead of conditioning on discrete token identities or generation positions, these methods manipulate textual, visual, or joint multimodal representations to reshape latent semantics, alignment structure, or reasoning behavior. Embedding-level steering thus provides a coarse-grained yet expressive mechanism for analyzing and controlling how VLMs encode and integrate multimodal information.

Text Embedding Text embedding steering refers to interventions applied to continuous textual representation spaces in VLMs, aiming to manipulate language-side semantics or alignment properties without conditioning on individual token identities. These methods typically operate on text encoder outputs or language hidden states that have been shaped by cross-modal supervision.

Several works analyze and steer text embeddings to improve linguistic abstraction and multimodal reasoning. Zhang et al. (2026); Yang et al. (2025b) study how textual representations in VLMs encode semantic and multilingual information, showing that embedding-level manipulation affects cross-modal generalization. A study by Li et al. (2025f) further demonstrates that adjusting language-side embeddings through learning-based objectives can guide reasoning behavior and improve task consistency in multimodal settings. In addition, Align-VLM (Masry et al., 2025) maps visual features to a probability distribution in the language embedding space, enabling more natural vision–language alignment. By leveraging the linguistic priors of large language models, it improves alignment quality and mitigates the noise and out-of-distribution issues introduced by directly projecting visual features.

Visual Embedding Visual embedding steering focuses on manipulating visual representation spaces, such as image-level or patch-level embeddings, to influence how visual information is encoded and subsequently aligned with language.

Although these methods intervene on visual latents, the manipulated representations are typically shaped under language supervision.

A line of work studies how steering visual embeddings affects multimodal perception and reasoning. A study in Yang et al. (2025a) investigates patch-level and hierarchical visual embeddings, showing that modifying visual latent representations changes attention patterns and visual grounding behavior. A study of Dai et al. (2025) focuses on improving visual understanding efficiency and robustness, demonstrating that embedding-level interventions can enhance fine-grained recognition and reduce sensitivity to visual distribution shifts. Furthermore, Lee et al. (2025) analyzes visual embedding structures in large-scale VLMs and highlight their role in scaling multimodal performance.

Text–Visual Embedding Text–visual embedding steering refers to interventions applied directly to joint multimodal representation spaces, where textual and visual information are explicitly fused or aligned. These methods target shared latent spaces that are central to multimodal grounding, alignment, and reasoning.

Recent studies increasingly focus on steering joint embeddings to improve multimodal integration. Chen et al. (2024); Li et al. (2025e) analyze and manipulate shared embedding spaces to enhance cross-modal alignment and reasoning consistency. A study of Duan et al. (2025a) shows that embedding-level control in fused spaces leads to more stable multimodal predictions and improved generalization across tasks. Additionally, Wang et al. (2025a) demonstrates that vulnerabilities and behaviors in VLMs can be traced to joint embedding structures, highlighting the importance of representation-level control in multimodal safety.

3.3 Layer-level

Activation on layer-level steering refers to a class of representation intervention methods that intervene on the intermediate activation representations of one or multiple layers in a VLM during forward inference, without modifying the model parameters. Instead of targeting individual tokens or neurons, these methods operate at the granularity of an entire Transformer layer or its major submodules, such as the attention block, the MLP block, or the residual stream, in order to systematically influence the model's multimodal reasoning and generation behavior.

Layer Output This steering target focuses on directly intervening in the hidden representations produced by an entire transformer layer during inference. In VLMs, these methods typically operate on intermediate hidden states to regulate semantic abstraction, visual grounding strength, or factual consistency before final decoding.

Recent works demonstrate that hallucination and over-reliance on parametric priors often emerge from middle or late layer representations, which motivates layerwise intervention without modifying model parameters. Existing approaches either inject steering vectors into selected layers or selectively amplify visually grounded features extracted from intermediate states, thereby biasing the generation process toward image-consistent evidence (Su et al., 2025; Li et al., 2025c). Other studies further exploit contrasts between visual and linguistic signals in layer outputs to control the balance between perceptual inputs and learned knowledge priors (Golovanevsky et al., 2025).

Attention Block Attention block activation steering intervenes in the outputs or internal dynamics of attention modules, aiming to reshape token interaction patterns that govern cross-modal information aggregation. In VLMs, hallucinations and misalignment are frequently linked to attention collapse, attention sinks, or improper dominance of linguistic tokens over visual tokens.

A line of work addresses these issues by penalizing overconfident attention distributions or reallocating attention mass toward visually grounded regions during inference (Liu et al., 2024). Other methods perform structured attention intervention by aligning attention patterns across languages, captions, or modalities, thereby improving robustness in multilingual and cross-domain scenarios (Li et al., 2025b). Recent studies further reveal that manipulating a small subset of critical attention heads or attention-induced residual updates can produce disproportionate effects on safety and factuality, which enables lightweight and targeted control (Xia et al., 2026). Collectively, attention block steering offers fine-grained control over information flow and plays a central role in hallucination mitigation, safety enhancement, and cross-modal alignment (Wang et al., 2025b).

MLP Block MLP block activation steering targets the feed-forward sub-layers within transformer blocks, which are increasingly recognized as key carriers of abstract knowledge, reasoning patterns,

and modality-specific signals.

In VLMs, hallucination-related features and object absence signals are often encoded in middle and late MLP activations (Kim et al., 2025b). Existing methods therefore identify influential neurons, build auxiliary classifiers on MLP outputs, or introduce lightweight steering and gating modules to detect unreliable states and suppress spurious reasoning (Duan et al., 2025b; Häon et al., 2025). For embodied and action-oriented VLMs, Li et al. Li et al. (2025a) further edit influential neuron pathways to enable controlled cross-modal unlearning while preserving world-state tracking and policy generation.

Residual Stream Residual stream activation steering intervenes in the additive pathways that propagate representations across Transformer blocks. Rather than modifying specific submodules, these methods inject or scale steering vectors in the residual stream to influence the cumulative representation trajectory during decoding.

In VLMs, residual stream steering has been used to bias generation toward visually grounded evidence, suppress hallucination-related directions, and enforce safety-aligned subspaces (Li et al., 2025d). Existing studies often derive steering directions from contrasts between faithful and hallucinated hidden states, and apply adaptive layer-wise scaling to balance controllability and stability with limited computational overhead (Wu et al., 2025). This paradigm also extends to training free safety defense and jailbreak mitigation by projecting residual updates onto constrained subspaces that restrict unsafe generation while preserving model utility (Zeng et al., 2025).

4 Representation Intervention on Weights

Weight-level Representation Intervention in VLMs refers to methods that intervene in model behavior by modifying trainable parameters in the network through direct manipulation of specific weight components without training. Unlike activation steering, which operates on intermediate representations during inference, weight-level methods permanently alter the parameter space of the model.

4.1 Component-Level

Component-Level Representation Intervention refers to the mechanism of modifying the weight parameters of specific pre-trained model components directly, without the need for backpropagation-based training (i.e., training-free). This approach typically employs closed-form solutions to calculate weight updates, thereby achieving precise control over the model's output behaviors or correcting specific knowledge errors during the inference phase.

FNN Projection A study by Chen et al. (2025) challenges the convention of solely editing the language module. This method employs attribution analysis to identify specific regions within the **Visual Encoder's FFNs** that contribute to erroneous outputs. It proposes correcting visual representations at the source to mitigate hallucinations or errors stemming from visual perception. DualEdit (Shi et al., 2025) introduces a strategy of **synergistic intervention**. Arguing that modifying a single modality is insufficient for ensuring cross-modal consistency, this approach calculates weight updates for both the Language Decoder (LLM FFN) and the Visual Projector (or Visual FFN) simultaneously. It imposes constraints to ensure that updates in both modalities remain consistent and non-conflicting. M2Edit (Zhou et al., 2025) focuses on **fine-grained localization** within components. Rather than treating an FFN layer as a monolithic block, it proposes scanning across the entire model architecture—including the Vision Encoder, Projector, and Language Decoder.

4.2 Global-Level

Global level representation intervention modifies model parameters through coordinated changes across multiple functional modules. It assumes that high-level behaviors, such as safety alignment and adversarial robustness, depend on coupled information routing and feature refinement. Thus, global steering often requires joint weight adjustment across sublayers.

Joint Projection Joint projection intervention refers to methods that simultaneously modify attention projection weights and MLP weights within Transformer blocks. This design reflects the view that attention and feed-forward components jointly shape information flow and representation refinement in VLMs.

In the LLM domain, Li et al. (2024b) first localizes safety alignment to intermediate safety layers and shows that jointly scaling attention and MLP weights can defend against jailbreak attacks. Inspired by this idea, LaSM (Yan et al., 2025) extends joint scaling to Qwen-VL and LLaVA-NeXT (Li

et al., 2024a) series VLMs. It identifies critical layers through attention heatmap analysis and applies joint scaling to defend against environment injection attacks in agent settings, demonstrating the effectiveness of this intervention strategy.

5 Future Directions and Open Problems

Causal Visual Representation Identification A fundamental open direction is to move from heuristic or correlation-based intervention to causal identification of critical visual representations that drive hallucination and unsafe behavior in VLMs. Existing token and layer-based activation steering methods show that intervening on a small subset of visual tokens or intermediate layers can significantly alter model outputs (Li et al., 2025c; Serra et al., 2024). However, most methods still use attention weights, gradient saliency, or contrastive signals as proxies for importance, without formal causal guarantees. Future work can integrate causal analysis to identify which visual tokens, attention heads, or MLP subspaces are necessary and sufficient for specific failures, such as hallucination, privacy leakage, or unsafe actions.

Controllable Cross-Modal Interaction Representations Another promising direction is to treat cross-modal interaction as a first-class representation object, rather than a byproduct of attention mechanisms. Prior studies show that intervening on cross-modal attention heads or pruning visual tokens involved in image-text fusion can mitigate hallucination and misalignment (Sarkar et al., 2025; Ye et al., 2025b). These findings suggest that image-text interaction patterns form a compact and manipulable representation space. Future research can model cross-modal token interactions as structured representations, allowing intervention to directly control grounding strength, language prior dominance, and multimodal consistency.

Bridging Activation and Weight Interventions Although activation-based and weight-based RepI are often studied separately, emerging evidence suggests that they are closely connected (Xu et al., 2026). Activation anchored training constrains intermediate representations during optimization (Zou et al., 2023), while inference time activation steering reveals stable directions that correlate with fine-tuned weight changes (Khayatan et al., 2025). A key direction is to study how steering directions can be distilled into lightweight component-level weight edits, or how weight edits can be decomposed into interpretable activation transformations. Such unification can combine the flexibility of activation steering with the persistence and efficiency of weight intervention.

Intervention in Sparse and Dynamic Architectures Most existing works focus on dense Transformer architectures, while VLMs are moving toward sparse and dynamic architectures such as Mixture of Experts (Zhou et al., 2022). In MoE-based VLMs, representations are routed through different experts, making global activation steering less predictable. Future work can explore routing intervention (Liao et al., 2025), which targets gating mechanisms to activate experts aligned with safety or factual grounding. For long context video VLMs with KV cache compression, another direction is to preserve critical steering signals during token retention and dynamic inference.

Representation intervention for Multimodal Agents Most existing representation intervention methods focus on single-turn generation, leaving their impact on long-horizon multimodal agent behavior underexplored. In interactive and embodied settings, agents operate through continuous perception–action loops, where even small representational biases introduced at the token, layer, or component level may accumulate over time, leading to drift in state tracking, inconsistent planning, or unsafe actions. Extending representation intervention to multimodal agents therefore requires understanding how activation-level and weight-level interventions propagate across temporal horizons and influence state representation, decision making, and policy execution. Early studies suggest that such interventions can meaningfully steer agent behavior and improve robustness and safety in realistic environments without full retraining (Li et al., 2025a; Yan et al., 2025).

6 Conclusion

This survey reviews representation intervention methods for VLMs. We organize existing work into a taxonomy covering activation-level and weight-level interventions, showing how targeted manipulation enables controllable multimodal behavior without full retraining. We also discuss remaining challenges and research directions for future work.

7 Limitation

Although this survey organizes existing representation intervention methods according to their intervention targets, it cannot provide a definitive rule for selecting the optimal target in a new VLM application. For external researchers who aim to apply representation intervention to their own tasks, it remains difficult to determine whether visual tokens, visual textual tokens, specific position tokens, layer outputs, attention blocks, or other targets should be preferred. This difficulty arises because existing methods differ not only in their intervention targets, but also in parameter choices, model backbones, training or inference settings, and auxiliary mechanisms. For example, in hallucination mitigation, visual textual tokens in VCD (Leng et al., 2024), specific position tokens in OPERA (Huang et al., 2024), and attention blocks in PAI (Liu et al., 2024) have all shown strong performance under their respective settings. More recent work, such as VISTA (Li et al., 2025c), further shows that interventions on visual tokens and layer outputs can outperform these methods on certain hallucination-related metrics. These results indicate that different targets may remain effective as methods continue to evolve within specific domains. Therefore, our taxonomy should not be interpreted as a practical ranking of intervention targets. More controlled studies are needed to establish reliable target selection criteria across models, datasets, metrics, and application goals.

References

Iñigo Alonso, Gorka Azkune, Ander Salaberria, Jeremy Barnes, and Oier Lopez de Lacalle. 2025. Vision-language models struggle to align entities across modalities. *arXiv preprint arXiv:2503.03854*.

Ravikumar Balakrishnan and Mansi Phute. 2025. Visor++: Universal visual inputs based steering for large vision language models. *arXiv preprint arXiv:2509.25533*.

Federico Barbero, Álvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. 2025. Why do llms attend to the first token? *arXiv preprint arXiv:2504.02732*.

Lukasz Bartoszcze, Sarthak Munshi, Bryan Sukidi, Jennifer Yen, Zejia Yang, David Williams-King, Linh Le, Kosi Asuzu, and Carsten Maple. 2025. Representation engineering for large-language models: Survey and research challenges. *arXiv preprint arXiv:2502.17601*.

Karuna Bhaila, Aneesh Komanduri, Minh-Hao Van, and Xintao Wu. 2025. Cross-modal attention guided unlearning in vision-language models. *arXiv preprint arXiv:2510.07567*.

Kaibing Chen, Dong Shen, Hanwen Zhong, Huasong Zhong, Kui Xia, Di Xu, Wei Yuan, Yifei Hu, Bin Wen, Tianke Zhang, Changyi Liu, Dewen Fan, Huihui Xiao, Jiahong Wu, Fan Yang, Size Li, and Di Zhang. 2024. Evlm: An efficient vision-language model for visual understanding. *arXiv preprint arXiv:2407.14177*.

Qizhou Chen, Taolin Zhang, Chengyu Wang, Xiaofeng He, Dakan Wang, and Tingting Liu. 2025. Attribution analysis meets model editing: Advancing knowledge correction in vision language models with visedit. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 2168–2176.

Dawei Dai, Long Xu, Yutang Li, Yuanhui Zhang, and Shuyin Xia. 2025. Humanvlm: Foundation for human-scene vision-language model. *Inf. Fusion*, 123:103271.

Yunkai Dang, Kaichen Huang, Jiahao Huo, Yibo Yan, Sirui Huang, Dongrui Liu, Mengxi Gao, Jie Zhang, Chen Qian, Kun Wang, Yong Liu, Jing Shao, Hui Xiong, and Xuming Hu. 2024. Explainable and interpretable multimodal large language models: A comprehensive survey. *arXiv preprint arXiv:2412.02104*.

Moreno D'Incà, Elia Peruzzo, Xingqian Xu, Humphrey Shi, Nicu Sebe, and Massimiliano Mancini. 2025. Safe vision-language models via unsafe weights manipulation. *arXiv preprint arXiv:2503.11742*.

Jiafei Duan, Wilbert Pumacay, Nishanth Kumar, Yi Ru Wang, Shulin Tian, Wentao Yuan, Ranjay Krishna, Dieter Fox, Ajay Mandlekar, and Yijie Guo. 2025a. AHA: A vision-language-model for detecting and reasoning over failures in robotic manipulation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*.

Jinhao Duan, Fei Kong, Hao Cheng, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. 2025b. Truthprint: Mitigating large vision-language models object hallucination via latent truthful-guided pre-intervention. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pages 7372–7382.

Michal Golovanevsky, William Rudman, Michael A. Lepori, Amir Bar, Ritambhara Singh, and Carsten Eickhoff. 2025. Pixels versus priors: Controlling knowledge priors in vision-language models through visual counterfacts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language*

Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025, pages 24837–24852.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. OPERA: alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13418–13427.

Bear Häon, Kaylene Stocking, Ian Chuang, and Claire Tomlin. 2025. Mechanistic interpretability for steering vision-language-action models. *arXiv preprint arXiv:2509.00328*.

Eun Woo Im, Muhammad Kashif Ali, and Vivek Gupta. 2026. Self-aug: Query and entropy adaptive decoding for large vision-language models. In *The Fourteenth International Conference on Learning Representations*.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. 2025. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 25004–25014.

Omri Kaduri, Shai Bagon, and Tali Dekel. 2025. What's in the image? A deep-dive into the vision of vision language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 14549–14558.

Pegah Khayatan, Mustafa Shukor, Jayneel Parekh, Arnaud Dapogny, and Matthieu Cord. 2025. Analyzing finetuning representation shift for multimodal llms steering. *arXiv preprint arXiv:2501.03012*.

Jieun Kim, Jinmyeong Kim, Yoonji Kim, and Sung-Bae Cho. 2025a. Fuzzy contrastive decoding to alleviate object hallucination in large vision-language models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pages 20572–20581.

Sohee Kim, Soohyun Ryu, Joonhyung Park, and Eunho Yang. 2025b. Unveiling the response of large vision-language models to visually absent tokens. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 21535–21557.

Byung-Kwan Lee, Ryo Hachiuma, Yu-Chiang Frank Wang, Yong Man Ro, and Yueh-Hua Wu. 2025. Vlsi: Verbalized layers-to-interactions from large to small vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 29545–29557.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13872–13882.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024a. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Kunhao Li, Wenhao Li, Di Wu, Lei Yang, Jun Bai, Ju Jia, and Jason Xue. 2025a. Cross-modal unlearning via influential neuron path editing in multimodal large language models. *arXiv preprint arXiv:2511.06793*.

Qiming Li, Zekai Ye, Xiaocheng Feng, Weihong Zhong, Libo Qin, Ruihan Chen, Baohang Li, Kui Jiang, Yaowei Wang, Ting Liu, and Bing Qin. 2025b. Cai: Caption-sensitive attention intervention for mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2506.23590*.

Shen Li, Liuyi Yao, Lan Zhang, and Yaliang Li. 2024b. Safety layers in aligned large language models: The key to llm security. *arXiv preprint arXiv:2408.17003*.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. 2025c. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. *arXiv preprint arXiv:2502.03628*.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. 2025d. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. 2025e. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. *arXiv preprint arXiv:2501.02189*.

Zongxia Li, Wenhao Yu, Chengsong Huang, Zhenwen Liang, Rui Liu, Fuxiao Liu, Jingxi Chen, Dian Yu, Jordan Boyd-Graber, Haitao Mi, and Dong Yu. 2025f. Self-rewarding vision-language model via reasoning decomposition. *arXiv preprint arXiv:2508.19652*.

Chih-Ting Liao, Zhangquan Chen, Chunlei Meng, Tzu-Yu Huang, Xin Cao, and Xu Zheng. 2025. Adversarial robustness for unified multi-modal encoders via efficient calibration. *arXiv preprint arXiv:2505.11895*.

Shi Liu, Kecheng Zheng, and Wei Chen. 2024. Paying more attention to image: A training-free method

for alleviating hallucination in lvlms. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXIII*, pages 125–140.

Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, Julian J. McAuley, Wei Ai, and Furong Huang. 2025. Large language models and causal inference in collaboration: A comprehensive survey. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 7668–7684.

Jiayun Luo, Wan-Cyuan Fan, Lyuyang Wang, Xiangteng He, Tanzila Rahman, Purang Abolmaesumi, and Leonid Sigal. 2025. To sink or not to sink: Visual information pathways in large vision-language models. *arXiv preprint arXiv:2510.08510*.

Ahmed Masry, Juan A. Rodriguez, Tianyu Zhang, Suyuchen Wang, Chao Wang, Aarash Feizi, Akshay Kalkunte Suresh, Abhay Puri, Xiangru Jian, Pierre-Andre Noel, Sathwik Tejaswi Madhusudhan, Marco Pedersoli, Bang Liu, Nicolas Chapados, Yoshua Bengio, Enamul Hoque, Christopher Pal, Issam H. Laradji, David Vazquez, and 3 others. 2025. AlignVLM: Bridging vision and language latent spaces for multimodal understanding. In *Second Workshop on Representational Alignment at ICLR 2025*.

Mansi Phute and Ravikumar Balakrishnan. 2025. Visor: Visual input-based steering for output redirection in vision-language models. *arXiv preprint arXiv:2508.08521*.

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. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763.

Sreetama Sarkar, Yue Che, Alex Gavin, Peter Anthony Beerel, and Souvik Kundu. 2025. Mitigating hallucinations in vision-language models through image-guided head suppression. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 12481–12500.

Alessandro Pietro Serra, Francesco Ortu, Emanuele Panizon, Lucrezia Valeriani, Lorenzo Basile, Alessio Ansuini, Diego Doimo, and Alberto Cazzaniga. 2024. The narrow gate: Localized image-text communication in native multimodal models. *arXiv preprint arXiv:2412.06646*.

Zhiyi Shi, Binjie Wang, Chongjie Si, Yichen Wu, Junsik Kim, and Hanspeter Pfister. 2025. Dualedit: Dual editing for knowledge updating in vision-language models. *arXiv preprint arXiv:2506.13638*.

Anushka Sivakumar, Andrew Zhang, Zaber Ibn Abdul Hakim, and Christopher Thomas. 2025. Steervlm: Robust model control through lightweight activation steering for vision language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 23640–23665.

Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Li Qing, and Zhaoxiang Zhang. 2025. 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 (Volume 1: Long Papers)*, pages 12964–12974.

Fengyuan Sun, Hui Chen, Xinhao Xu, Dandan Zheng, Jingdong Chen, Jun Zhou, Jungong Han, and Guiguang Ding. 2025. Prunehal: Reducing hallucinations in multi-modal large language models through adaptive kv cache pruning. *arXiv preprint arXiv:2510.19183*.

Xiaosen Wang, Shaokang Wang, Zhijin Ge, Yuyang Luo, and Shudong Zhang. 2025a. Attention! your vision language model could be maliciously manipulated. *arXiv preprint arXiv:2505.19911*.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 15840–15853.

Yining Wang, Mi Zhang, Junjie Sun, Chenyue Wang, Min Yang, Hui Xue, Jialing Tao, Ranjie Duan, and Jiexi Liu. 2025b. Mirage in the eyes: Hallucination attack on multi-modal large language models with only attention sink. *arXiv preprint arXiv:2501.15269*.

Ziyue Wang, Yurui Dong, Fuwen Luo, Minyuan Ruan, Zhili Cheng, Chi Chen, Peng Li, and Yang Liu. 2025c. How do multimodal large language models handle complex multimodal reasoning? placing them in an extensible escape game. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pages 4807–4817.

Jan Wehner, Sahar Abdelnabi, Daniel Tan, David Krueger, and Mario Fritz. 2025. Taxonomy, opportunities, and challenges of representation engineering for large language models. *arXiv preprint arXiv:2502.19649*.

Laura Wenderoth, Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. 2025. Measuring cross-modal interactions in multimodal models. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium*

on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025, pages 21501–21509.

Junfei Wu, Yue Ding, Guofan Liu, Tianze Xia, Ziyue Huang, Dianbo Sui, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2025. SHARP: steering hallucination in lvlms via representation engineering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 14346–14361.

Junhao Xia, Haotian Zhu, Shuchao Pang, Zhigang Lu, Bing Li, Yongbin Zhou, and Jason Xue. 2026. One head to rule them all: Amplifying LVLM safety through a single critical attention head. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Ziwen Xu, Chenyan WU, Hengyu Sun, Haiwen Hong, Mengru Wang, Yunzhi Yao, Longtao Huang, Hui Xue, Shumin Deng, Zhixuan Chu, Huajun Chen, and Ningyu Zhang. 2026. Why steering works: Toward a unified view of language model parameter dynamics. In *ICLR 2026 Workshop on Principled Design for Trustworthy AI - Interpretability, Robustness, and Safety across Modalities*.

Zihe Yan, Jiaping Gui, Zhuosheng Zhang, and Gongshen Liu. 2025. Lasm: Layer-wise scaling mechanism for defending pop-up attack on gui agents. *arXiv preprint arXiv:2507.10610*.

Kichang Yang, Seonjun Kim, Minjae Kim, Nairan Zhang, Chi Zhang, and Youngki Lee. 2025a. Vlm in a flash: I/o-efficient sparsification of vision-language model via neuron chunking. *arXiv preprint arXiv:2511.18692*.

Senqiao Yang, Junyi Li, Xin Lai, Bei Yu, Hengshuang Zhao, and Jiaya Jia. 2025b. Visionthink: Smart and efficient vision language model via reinforcement learning. *arXiv preprint arXiv:2507.13348*.

Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang, and Yansong Tang. 2025a. Atp-llava: Adaptive token pruning for large vision language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 24972–24982.

Zekai Ye, Qiming Li, Xiaocheng Feng, Libo Qin, Yichong Huang, Baohang Li, Kui Jiang, Yang Xiang, Zhirui Zhang, Yunfei Lu, Duyu Tang, Dandan Tu, and Bing Qin. 2025b. Claim: Mitigating multilingual object hallucination in large vision-language models with cross-lingual attention intervention. *arXiv preprint arXiv:2506.11073*.

Xiyu Zeng, Siyuan Liang, Liming Lu, Haotian Zhu, Enguang Liu, Jisheng Dang, Yongbin Zhou, and Shuchao Pang. 2025. Safesteer: Adaptive subspace steering for efficient jailbreak defense in vision-language models. *arXiv preprint arXiv:2509.21400*.

Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, XinQiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, He Wang, Zhizheng Zhang, Li Yi, Wenjun Zeng, and Xin Jin. 2026. DreamVLA: A vision-language-action model dreamed with comprehensive world knowledge. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yang Zhou, Pengfei Cao, Yubo Chen, Qingbin Liu, Dianbo Sui, Xi Chen, Kang Liu, and Jun Zhao. 2025. M2edit: Locate and edit multi-granularity knowledge in multimodal large language model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 29029–29042.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, and 1 others. 2022. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114.

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, and 2 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

Control Function	Mathematical form
Directional shift control	$\tau' = \tau + \lambda v$
Multiplicative control	$\tau' = \alpha\tau$ or $\tau' = D\tau$
Normalized control	$\tau' = \frac{\tau+\lambda v}{\lVert\tau+\lambda v\rVert}\lVert\tau\rVert$
Projection control	$\tau' = P\tau$
Affine control	$\tau' = A\tau + b$
Gated control	$\tau' = \tau + g(\Omega)\lambda v$

Table 2: Abstract control functions for representation intervention.

A Control Functions for Activation Intervention.

A.1 Definition for Control Functions

Under Definition from Section 2.2, the intervention operator $\mathcal{I}_\mathcal{C}$ instantiates a control function that maps a selected target τ to a controlled target τ'. The target τ may correspond to an activation component, such as a visual token, an embedding, a layer output, an attention block output, or an MLP block output. It may also correspond to a weight component, such as a projection matrix or another parameter subset. Although existing methods implement $\mathcal{I}_\mathcal{C}$ in different ways, many control functions can be summarized by a small set of elementary transformations as illustrated in Table 2.

Directional shift control. This method moves the selected target along a control direction, where the sign and magnitude of the coefficient determine whether the direction is strengthened or weakened.

Multiplicative control. Such control method rescales the target globally or along selected dimensions, which amplifies or attenuates existing information without necessarily changing its direction.

Normalized control. This method adjusts the magnitude after intervention, which helps prevent abnormal target norms.

Projection-based control. This method maps the target onto or away from a subspace, which enables structured preservation or removal of concept-related information.

Affine control. More expressive variants, such as affine control, combine scaling, rotation, and translation into a unified transformation.

Gated control further makes the intervention dependent on the input, model state, or intervention configuration.

We hope that presenting representation intervention through these atomic control functions can provide outside researchers with an accessible entry point for understanding and developing new steering methods. The next subsection presents a concrete case study to illustrate how these control functions can be combined within a complete intervention framework.

A.2 Case Study: Hybrid Control Functions in VISTA.

As shown in Table 1, a single study does not necessarily instantiate only one control function. Many representation intervention methods combine several elementary operations into a hybrid intervention pipeline. Therefore, instead of assigning each method to a single abstract formula, we use VISTA (Li et al., 2025c) as a representative case study to demonstrate how our taxonomy can analyze a complete method. VISTA first constructs a visual steering vector by contrasting a positive context that preserves visual tokens with a negative context in which visual tokens are removed:

$$V_{\text{steer}} = V_p - V_n = \{v^l_{\text{steer}}\}_{l=1}^{L}, \tag{5}$$

where $V_p = \mathcal{F}(X_p)$ and $V_n = \mathcal{F}(X_n)$. During inference, the method injects the steering vector into the residual stream at each generation step:

$$\tilde{h}^l_t = h^l_t + \lambda v^l_s, \quad l \in [1, L], \tag{6}$$

which corresponds to **directional shift control**. To maintain generation stability, VISTA further normalizes the modified hidden state:

$$\tilde{h}^l_t = \tilde{h}^l_t \cdot \frac{\lVert h^l_t\rVert_2}{\lVert \tilde{h}^l_t\rVert_2}, \quad l \in [1, L], \tag{7}$$

which corresponds to **normalized control**. In addition to activation intervention, VISTA introduces self logits augmentation by averaging logits from the w layers before the final layer:

$$o^{\text{aug}}_t = \frac{1}{w}\sum_{l=L-w}^{L-1} \mathcal{H}(h^l_t), \tag{8}$$

and combining them with the final layer logits:

$$\tilde{o}_t = (1-\gamma)\cdot o^L_t + \gamma \cdot o^{\text{aug}}_t, \tag{9}$$

where $o^L_t = \mathcal{H}(h^L_t)$. This component performs logit-level aggregation rather than a direct transformation of the selected internal target τ. Therefore, VISTA combines **directional shift control**,

normalized control, and output level aggregation within a single framework. This case illustrates why assigning one closed-form control function to each method may obscure practical implementations.