SCAS: SYMMETRICAL CONCEPT-SPACE ACTIVATION STEERING FOR JOINT HALLUCINATION MITIGATION AND CAPABILITY ENHANCEMENT IN VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

A fundamental challenge in Vision-Language Models (VLMs) lies in the trade-off between mitigating hallucinations and preserving general reasoning capabilities, as existing safety interventions often degrade model utility or introduce severe biases. In this paper, we propose **SCAS** (**S**ymmetrical **C**oncept-space **A**ctivation **S**teering), a lightweight, train-free framework that achieves dual enhancement in both safety and utility. Utilizing only 100 calibration samples, SCAS extracts distinct semantic bases for factual groundedness (U_{pos}) and hallucination (U_{neg}). By orthogonalizing these bases, SCAS ensures decoupled, bidirectional steering of intermediate hidden states ($h' = h - \alpha P_{\text{neg}} + \beta P_{\text{pos}}$), simultaneously suppressing hallucination pathways while reinforcing core reasoning representations.

Evaluations on Qwen3-VL-2B, Qwen3-VL-8B, and InternVL3.5-4B demonstrate that SCAS effectively addresses the safety-utility conflict: it significantly reduces visual hallucinations on RePoPe and AMBER, while consistently improving general capabilities on MME and MMBench. Furthermore, on generative captioning, SCAS avoids the decoding collapse suffered by prior methods, maintaining rich descriptive coverage while maintaining safety. This confirms SCAS as a promising solution for co-promoting VLM reliability and general intelligence.

1 INTRODUCTION

Vision-Language Models (VLMs) have achieved remarkable progress in multimodal understanding and generation, enabling strong performance in tasks such as visual question answering and image captioning. Representative models such as CLIP Radford et al. (2021), LLaVA Liu et al. (2023a), InstructBLIP Dai et al. (2023), InternVL Chen et al. (2024), and Qwen-VL Bai et al. (2023) demonstrate strong cross-modal alignment and generalization capability.

Despite these advances, a fundamental limitation remains: the internal representations of VLMs are not fully controllable. This leads not only to local generation errors such as object hallucination Li et al. (2023b); Wang et al. (2024), but more importantly to a broader issue of entangled capability control, where improving one aspect of model behavior often degrades others. In particular, improving factual grounding may inadvertently harm general multimodal reasoning ability, indicating a lack of structured controllability in the representation space.

Recent studies have explored this issue from different perspectives. Training-based approaches, such as supervised fine-tuning Liu et al. (2023a) or reinforcement learning from human feedback Yu et al. (2024), improve behavioral alignment but often require expensive retraining and may reduce general capability. Inference-time decoding methods, such as Visual Contrastive Decoding (VCD) Leng et al. (2024), OPERA Huang et al. (2024), HALC Li et al. (2024), and Woodpecker Zhao et al. (2023), modify generation without retraining but rely on additional computation paths and provide limited improvements in overall capability.

More recently, latent-space intervention or steering methods have emerged as a promising direction, drawing insights from the emerging area of representation engineering Zou et al. (2023); Li et al. (2023a); Rimsky et al. (2024). These methods directly manipulate internal representations to control model behavior. For example, Revis Wu et al. (2026) demonstrates that steering directions in hidden space can effectively reduce hallucination. However, existing steering methods still suffer from two

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

key limitations: (1) steering directions are often model-dependent and lack cross-model stability; and (2) they primarily focus on suppressing undesired behavior, without explicitly improving general multimodal capability, often leading to performance trade-offs.

In this paper, we argue that these limitations stem from an underexplored aspect: the structure of semantic subspaces in VLM representations is not explicitly disentangled. We hypothesize that different behavioral aspects of VLMs are driven by distinct but entangled semantic directions in the latent space. This raises a fundamental question: can we explicitly construct and disentangle semantic subspaces to enable controllable and synergistic behavior optimization?

To address this question, we propose **SCAS (Symmetrical Concept-space Activation Steering)**, a training-free inference-time framework for controllable representation intervention. SCAS first constructs a small calibration dataset following prior latent steering protocols such as Revis Wu et al. (2026), and extracts semantic direction bases from paired factual and hallucinated samples. These directions are then used to construct two complementary subspaces: a factual semantic subspace and a deviation (hallucination-related) subspace.

To ensure stable and decoupled control, we introduce a structured disentanglement mechanism that orthogonalizes these subspaces and removes components correlated with model response bias. This results in a set of independent semantic directions that allow controllable and interpretable intervention in hidden space.

At inference time, SCAS performs symmetrical activation steering on intermediate hidden states, simultaneously suppressing deviation-related activations while enhancing factual semantic activations. This symmetric design enables bidirectional control in representation space, leading to improved behavioral consistency without sacrificing general capability.

We evaluate SCAS on multiple VLM architectures, including Qwen3-VL-2B, Qwen3-VL-8B, and InternVL3.5-4B. Experimental results show that SCAS significantly reduces object hallucination on REPOPE Neuhaus & Hein (2025) and AMBER Wang et al. (2023), while achieving consistent improvements on general multimodal benchmarks such as MME Fu et al. (2023) and MMBench Liu et al. (2023b). Notably, SCAS improves both robustness and general capability simultaneously, in contrast to existing steering methods that typically exhibit trade-offs. Further analysis demonstrates strong cross-model generalizability of the framework and stable behavior across diverse architectures.

Our contributions are summarized as follows:

- We propose SCAS, a symmetrical concept-space activation steering framework for controllable representation intervention in VLMs.
- We introduce a subspace disentanglement mechanism that enables stable and decoupled control of semantic directions in latent space.
- We demonstrate that SCAS achieves simultaneous hallucination reduction and general capability improvement across multiple VLM architectures and benchmarks.

2 Related Work

2.1 Multimodal Evaluation: From Hallucination to General Capability Assessment

Vision-language models (VLMs) have achieved strong performance in multimodal understanding, but still suffer from reliability issues such as hallucinations and inconsistent reasoning. To systematically evaluate these models, existing benchmarks can be broadly categorized into hallucination-centric evaluation and general multimodal capability evaluation.

Hallucination-Centric Evaluation. Object hallucination has been widely studied as a key failure mode of VLMs. Early discriminative benchmarks such as POPE Li et al. (2023b) evaluate whether models correctly identify object existence in a binary question-answering setting. However, POPE has been shown to suffer from annotation bias, leading to the refined RePOPE Neuhaus & Hein (2025). To further enable fine-grained evaluation, AMBER Wang et al. (2023) introduces a multi-dimensional framework covering object existence, attribute correctness, and relational reasoning, providing a more systematic analysis of hallucination behaviors.

In addition to discriminative evaluation, generative hallucination is commonly measured in free-form text generation tasks. Representative metrics include CHAIR Rohrbach et al. (2018) and object coverage, which quantify inconsistencies between generated descriptions and visual content. These studies suggest that hallucinations are not isolated perception errors, but are closely related to entangled semantic representations across objects, attributes, and relations in latent space.

General Multimodal Capability Evaluation. Beyond hallucination-specific benchmarks, recent work has proposed comprehensive evaluation suites for assessing overall multimodal understanding ability. Benchmarks such as MME and MMBench evaluate VLMs across perception, reasoning, and knowledge-intensive tasks. Unlike hallucination metrics that focus on local inconsistencies, these benchmarks reflect global multimodal competence, including visual recognition, commonsense reasoning, and compositional understanding.

Importantly, improvements on hallucination metrics do not necessarily translate to general capability gains. Therefore, we additionally evaluate our method on MME and MMBench to verify that it improves not only factual consistency but also overall multimodal representation quality.

Training-Time Alignment. To mitigate hallucination and improve alignment, prior work explores supervised fine-tuning with negative instructions, reinforcement learning from human feedback (RLHF Yu et al. (2024)), and direct preference optimization (DPO). However, these approaches require updating large-scale model parameters and often introduce an alignment tax, potentially degrading generalization ability.

2.2 Inference-Time Decoding and Correction

To avoid expensive retraining, a large body of work focuses on inference-time methods, including decoding-time intervention and post-hoc correction.

Contrastive Decoding. These methods modify token-level logits during autoregressive generation. Visual Contrastive Decoding (VCD) Leng et al. (2024) suppresses hallucinated tokens by contrasting clean and perturbed visual inputs. OPERA Huang et al. (2024) reduces over-trust in attention via retrospective allocation, while HALC Li et al. (2024) leverages local-to-global contrastive signals for hallucination mitigation.

Post-hoc Correction. Another line of work performs post-generation refinement using external detectors or auxiliary models, such as the Woodpecker Zhao et al. (2023) framework.

Limitations. Although training-free, these methods often rely on additional computation or heuristic pipelines. More importantly, they operate at the output probability space and cannot directly modify intermediate representations, limiting their ability to fundamentally improve multimodal reasoning.

2.3 Latent Representation Steering

Our work is closely related to representation engineering Zou et al. (2023) and activation steering methods Li et al. (2023a); Rimsky et al. (2024), which control model behavior by manipulating internal hidden states.

Activation Steering. Inference-Time Intervention (ITI) Li et al. (2023a) demonstrates that steering activations along linear probe directions can improve model truthfulness. Representation engineering Zou et al. (2023) further generalizes this idea to controllable manipulation of abstract attributes such as emotion, safety, and morality. In multimodal settings, Revis Wu et al. (2026) extends latent steering to VLMs by injecting sparse visual factual vectors into intermediate representations during decoding.

Limitations and Motivation. However, existing additive steering methods often introduce side effects due to dynamic hidden-state scaling during generation. Fixed-magnitude interventions may cause representational drift Zou et al. (2023), leading to degeneration such as repetitive outputs or biased responses. These limitations motivate more stable and decoupled steering mechanisms.

To address these issues, we propose SCAS, a symmetric and decoupled subspace steering method with explicit debiasing, enabling stable and generalizable multimodal control across tasks.

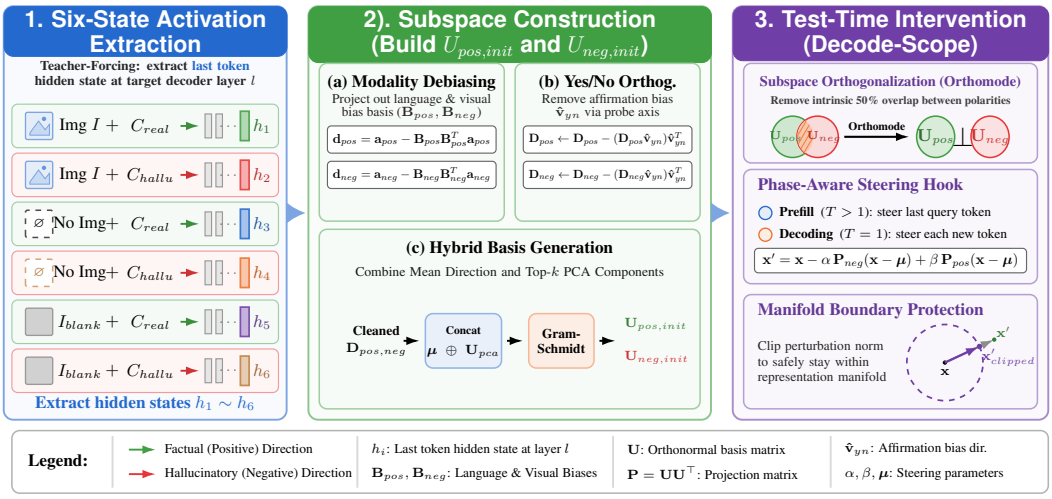


Figure 1: Overview of SCAS. Six activation states are extracted from the last token. Raw directional vectors undergo modality debiasing and yes/no probe orthogonalization. Factual subspaces are derived via hybrid basis generation and mutually orthogonalized. Finally, phase-aware test-time steering adjusts the latent representations with boundary protection.(Norm clipping is optionally applied to constrain latent perturbation norms; see Appendix for details).

3 METHODOLOGY

The proposed **Symmetric Contrastive Activation Steering (SCAS)** framework dynamically modifies intermediate representation manifolds during test-time inference. SCAS mathematically isolates factual steering directions from confounding multimodal biases using a projection-based cleaning protocol. The pipeline consists of two primary phases: (1) an offline, debiased subspace calibration phase to construct mutually orthogonal factual subspaces, and (2) a low-overhead, test-time selective intervention phase that steers the decoder's activation manifold without deteriorating base representation capabilities.

3.1 SIX-STATE ACTIVATION REPRESENTATION

To capture the specific activation patterns that govern factual hallucination, we extract intermediate representation states from target decoder layers. During the offline calibration phase, we operate under a **teacher-forcing paradigm** (Rimsky et al., 2024), feeding the entire concatenated sequence composed of the image (or blank control), the textual query, and the positive or negative target caption.

Following representation probing methodologies (Zou et al., 2023), for each calibration sample, we extract the **last hidden state** $\mathbf{h} \in \mathbb{R}^D$ at target decoder layer l, corresponding to the activation of the final token of the target caption. Under a teacher-forcing paradigm, this final token representation fully encodes the integrated semantic dependencies of both the multimodal context and the factual or hallucinatory alignment of the target caption. Specifically, given a calibration triplet $(I, C_{\text{real}}, C_{\text{hallu}})$, we construct a six-state activation profile consisting of distinct hidden states $\mathbf{h}_1, \ldots, \mathbf{h}_6$ to symmetrically decompose the hidden space. Detailed input configurations, modality controls (including the text-only baseline $\emptyset$ and solid background control I_{blank}), and symmetric pairing details are illustrated in Figure 1.

3.2 MULTI-MODAL DEBIASING VIA SUBSPACE PROJECTION

Directly subtracting raw activation states (e.g., $\mathbf{h}_1 - \mathbf{h}_2$) introduces severe confounding factors, such as text-only semantic priors and non-factual visual object representations. To isolate a clean representation of factual consistency, we implement a multimodal subspace cleaning protocol inspired by causal projection and editing of latent representations (Meng et al., 2022).

For the positive and negative channels, we define the raw discrepancy vectors as $\mathbf{a}_{\text{pos}} = \mathbf{h}_1 - \mathbf{h}_2$ and $\mathbf{a}_{\text{neg}} = \mathbf{h}_2 - \mathbf{h}_1$. The corresponding language bias and visual bias vectors are formulated as:

$$\mathbf{t}_{\text{pos}} = \mathbf{h}_3 - \mathbf{h}_4, \quad \mathbf{v}_{\text{pos}} = \mathbf{h}_1 - \mathbf{h}_5; \qquad \mathbf{t}_{\text{neg}} = \mathbf{h}_4 - \mathbf{h}_3, \quad \mathbf{v}_{\text{neg}} = \mathbf{h}_2 - \mathbf{h}_6 \tag{1}$$

To strip away textual priors and unaligned visual features, we project each $\mathbf{a}_*$ onto the orthogonal complement of the subspace spanned by its corresponding bias basis matrix $\mathbf{B}_* = [\hat{\mathbf{t}}_*, \hat{\mathbf{v}}_*] \in \mathbb{R}^{D\times 2}$. The cleaned directional vectors $\mathbf{d}_*$ are obtained symmetrically:

$$\mathbf{d}_* = \text{res}(\mathbf{a}_*, \mathbf{B}_*) = \mathbf{a}_* - \mathbf{B}_*\mathbf{B}_*^T\mathbf{a}_*, \quad \forall * \in \{\text{pos}, \text{neg}\} \tag{2}$$

where $\mathbf{B}_*$ is the orthonormalized basis matrix of $[\mathbf{t}_*, \mathbf{v}_*]$.

3.3 Affirmation-Negation (Yes/No) Probe Orthogonalization

Visual question answering benchmarks often suffer from systemic affirmation bias (i.e., a baseline tendency to output "Yes"), which distorts the extracted activation representations (Li et al., 2023b). To prevent our factual steering vectors from being corrupted by generic decision boundaries associated with answering style, we introduce a task-specific **Yes/No Probe Orthogonalization** step (Alain & Bengio, 2016).

We compile a balanced set of $M = 20$ common-sense textual binary questions (half designed to elicit true "Yes" and half true "No"). By recording activations at the last prompt token, we calculate the mean affirmation-negation decision axis:

$$\mathbf{v}_{\text{yn}} = \frac{1}{M/2}\sum_{i\in\text{Yes}} \mathbf{h}_i^{\text{probe}} - \frac{1}{M/2}\sum_{j\in\text{No}} \mathbf{h}_j^{\text{probe}}, \quad \hat{\mathbf{v}}_{\text{yn}} = \frac{\mathbf{v}_{\text{yn}}}{\|\mathbf{v}_{\text{yn}}\|_2} \tag{3}$$

Before performing low-rank decomposition, we explicitly remove the projection of this decision axis from our calibration direction matrices $\mathbf{D}_{\text{pos}}, \mathbf{D}_{\text{neg}} \in \mathbb{R}^{N\times D}$:

$$\mathbf{D}_* \leftarrow \mathbf{D}_* - (\mathbf{D}_*\hat{\mathbf{v}}_{\text{yn}})\hat{\mathbf{v}}_{\text{yn}}^T, \quad \forall * \in \{\text{pos}, \text{neg}\} \tag{4}$$

This step ensures that downstream steering operations selectively adjust factual content rather than artificially manipulating the model's answering style.

3.4 Hybrid Subspace Construction and Overlap Resolution

Hybrid Basis Generation. Rather than relying solely on the mean shift or principal components, our hybrid approach combines both representations. Specifically, for each cleaned matrix $\mathbf{D} \in \{\mathbf{D}_{\text{pos}}, \mathbf{D}_{\text{neg}}\}$, we extract: (i) the normalized global mean vector $\mathbf{u}_{\text{mean}} = \bar{\mathbf{d}}/\|\bar{\mathbf{d}}\|_2$, capturing the general direction of activation translation; and (ii) the top-k singular vectors $\mathbf{U}_{\text{pca}} \in \mathbb{R}^{D\times k}$ via Singular Value Decomposition (SVD) on the centered direction matrix, capturing primary variance axes (Burns et al., 2022; Zou et al., 2023). The initial basis matrix is obtained via Gram-Schmidt orthonormalization:

$$\mathbf{U}_{\text{init}} = \text{orthonormalize}([\mathbf{u}_{\text{mean}}, \mathbf{U}_{\text{pca}}]) \tag{5}$$

Intrinsic Latent Overlap & Resolution. Empirical evaluations on the REVIS (Wu et al., 2026) calibration dataset reveal that $\mathbf{U}_{\text{pos,init}}$ and $\mathbf{U}_{\text{neg,init}}$ consistently exhibit a high mutual subspace overlap (~ 0.5), indicating an intrinsic, global shared "conceptual backbone" in the pre-trained VLM's latent manifold (see Supp. for details). Because projection operators ($\mathbf{P} = \mathbf{U}\mathbf{U}^T$) are sign-agnostic, simultaneous steering without adjustment causes destructive interference. We resolve this via two directional orthogonalization modes:

$$\begin{aligned} &\textbf{Mode A (}\texttt{pos}\textbf{)}: && \begin{cases} \mathbf{U}_{\text{pos}} = \text{orth}(\text{res}(\mathbf{U}_{\text{pos,init}}, \mathbf{U}_{\text{neg,init}})) \\ \mathbf{U}_{\text{neg}} = \mathbf{U}_{\text{neg,init}} \end{cases} \\ &\textbf{Mode B (}\texttt{both}\textbf{)}: && \begin{cases} \mathbf{U}_{\text{pos}} = \text{orth}(\text{res}(\mathbf{U}_{\text{pos,init}}, \mathbf{U}_{\text{neg,init}})) \\ \mathbf{U}_{\text{neg}} = \text{orth}(\text{res}(\mathbf{U}_{\text{neg,init}}, \mathbf{U}_{\text{pos,init}})) \end{cases} \end{aligned} \tag{6}$$

This formulation yields mutually orthogonal spaces ($\mathbf{P}_{\text{neg}}\mathbf{P}_{\text{pos}} = \mathbf{0}$), effectively decoupling opposing steering dynamics.

3.5 Test-Time Activation Steering via Decode Hook

During test-time generation, we execute a localized **decode-scope intervention** to preserve multimodal contextual reasoning, extending inference-time intervention paradigms (Li et al., 2023a; Rimsky et al., 2024; Turner et al., 2023). The intervention hook adaptively identifies the generation phase based on the sequence length T of the hidden state tensor $\mathbf{H} \in \mathbb{R}^{B\times T\times D}$.

Layer Selection Strategy. Steering is applied to a specific intermediate decoder layer l: Layer 15 for Qwen3-VL-2B and Layer 24 for InternVL3.5-4B. Grounded in VLM mechanistic analyses (Kaduri et al., 2025), cross-modal fusion predominantly peaks in intermediate layers. Mid-layer steering provides a computational buffer for subsequent layers to naturally contextualize and refine perturbed activations, mitigating decoding collapse. As detailed in the supplementary material, optimal layers are pinpointed via auxiliary probing AUC and subspace overlap diagnostics on calibration data.

Phase-Aware Steering. The intervention is executed adaptively across generation phases:

- **Prefill Phase ($T > 1$):** To avoid corrupting raw visual tokens and instruction prompts, the hook only steers the final query token that aggregates global sequence context: $\mathbf{H}[:, -1, :] \leftarrow \mathrm{Steer}(\mathbf{H}[:, -1, :])$.
- **Decoding Phase ($T = 1$):** During autoregressive generation steps (Vaswani et al., 2017), the hook is applied directly to each newly generated token state: $\mathbf{H} \leftarrow \mathrm{Steer}(\mathbf{H})$.

The steering function $\mathrm{Steer}(\mathbf{x})$ is defined as:

$$\mathrm{Steer}(\mathbf{x}) = \mathbf{x} - \alpha\mathbf{P}_{\text{neg}}(\mathbf{x} - \boldsymbol{\mu}) + \beta\mathbf{P}_{\text{pos}}(\mathbf{x} - \boldsymbol{\mu}) \tag{7}$$

where $\mathbf{P} = \mathbf{U}\mathbf{U}^T$ are the orthogonal projection matrices, α and β govern steering intensities, and $\boldsymbol{\mu} = \frac{1}{2}(\bar{\mathbf{h}}_1 + \bar{\mathbf{h}}_2)$ is the reference centering bias from calibration.

4 Experiments

4.1 Experimental Setup

Models. We evaluate SCAS on two recent and highly optimized Vision-Language Models (VLMs): **Qwen3-VL-2B** Qwen Team (2026), a compact 28-layer VLM, and **InternVL3.5-4B** InternVL Team (2026), a stronger 36-layer VLM with competitive visual grounding capability. To further examine scalability, we additionally report supplementary results on **Qwen3-VL-8B** in Section 4.5. We evaluate modern, well-grounded VLMs rather than older architectures to rigorously test intervention efficacy. All runs employ deterministic greedy decoding (T=0).

Benchmarks. We evaluate both hallucination robustness and general multimodal capability. For hallucination evaluation, we use **RePoPe** Li et al. (2023b), which measures object existence hallucination under random, popular, and adversarial settings, and **AMBER** Wang et al. (2023), which contains both discriminative subtasks and free-form generation evaluation. For general multimodal utility, we use **MME** Fu et al. (2023) and **MMBench** Liu et al. (2023b). This joint evaluation is important because hallucination metrics alone can be misleading: a method may reduce hallucination scores by over-producing negative answers, under-describing the image, or causing repetitive generation collapse.

Baselines and Implementation Details. We compare SCAS against standard **Baseline** decoding, **VCD** Leng et al. (2024) (α=1.0, β=0.1, noise level 500), and **Revis** Wu et al. (2026) (coeff=1.6, τ=None, where omitting thresholding prevents intervention bypassing on modern VLMs and provides a stronger, more competitive baseline). For SCAS, calibration directions are extracted from OH-100 (Wu et al., 2026) with Yes/No debiasing and PCA rank k=8 (`both` strategy). By default, we use α=1.0, β=1.0, `ortho_mode=pos` during decoding, intervening at Layer 15 for Qwen3-VL-2B and Layer 24 for InternVL3.5-4B.

4.2 Main Results: Safety-Utility Pareto Evaluation

To prevent artificial metric inflation via degenerate predictions (e.g., blind negation), we assess hallucination suppression jointly with multimodal utility.

Table 1: Joint evaluation of safety and utility on Qwen3-VL-2B and InternVL3.5-4B. Safety is measured by RePoPe Adversarial and AMBER-d Overall. Utility is measured by MMBench and MME. For Qwen3-VL-2B and InternVL3.5-4B, SCAS uses the primary Layer 15 and Layer 24 configurations, respectively. Best results are in **bold**, and second-best results are underlined.

Model	Method	Safety: RePoPe Adversarial				Safety: AMBER-d Overall		Utility: General	
		Acc↑	F1↑	FP↓	FN↓	Acc↑	F1↑	MMBench Acc↑	MME Total↑
Qwen3-VL-2B	Baseline	92.29	91.58	145	62	86.27	89.14	83.16	2033.27
	VCD	92.36	91.68	147	**58**	85.64	88.67	83.33	2058.68
	Revis	**93.11**	**92.26**	**101**	84	**86.75**	**89.90**	83.93	2086.79
	SCAS (Ours)	92.92	92.08	107	83	86.68	89.86	**84.62**	**2101.92**
InternVL3.5-4B	Baseline	83.64	84.11	414	25	84.84	87.84	86.77	2192.15
	Revis	83.27	83.87	429	**20**	85.33	88.29	85.65	**2237.00**
	SCAS (Ours)	**86.14**	**86.16**	**343**	29	**85.64**	**88.69**	**87.03**	2165.25

Table 1 reports the joint evaluation on RePoPe, AMBER-d, MMBench, and MME. Across the two primary models, SCAS provides the most stable safety-utility trade-off. On **InternVL3.5-4B**, SCAS improves RePoPe Adversarial F1 from 84.11% to **86.16%**, while reducing false positives from 414 to **343**. Importantly, this hallucination reduction does not come at the cost of general capability: MMBench accuracy increases from 86.77% to **87.03%**. In contrast, Revis becomes unstable on this stronger architecture. It decreases RePoPe Adversarial F1 from 84.11% to 83.87%, increases false positives to 429, and reduces MMBench accuracy by 1.12 percentage points.

On **Qwen3-VL-2B**, Revis obtains a slightly higher RePoPe Adversarial F1 than SCAS, with a margin of only 0.18 percentage points. However, this marginal gain is accompanied by lower general utility. SCAS outperforms Revis on both MMBench, 84.62% versus 83.93%, and MME, 2101.92 versus 2086.79. SCAS also substantially reduces RePoPe false positives compared with the baseline, from 145 to 107, while improving AMBER-d Overall F1 from 89.14% to 89.86%. These results suggest that SCAS achieves hallucination suppression through a more controlled representation correction rather than by simply making the model overly conservative.

Overall, SCAS is the only method that consistently achieves substantial gains on hallucination-related benchmarks while largely preserving general multimodal reasoning capabilities across model families (with only negligible variance on InternVL3.5 MME).

4.3 Free-form Generation: Avoiding Degenerate Suppression

We further evaluate free-form captioning on AMBER-g, where the model must remain descriptive while avoiding unsupported objects. This setting is particularly important because hallucination metrics can be distorted by degenerate generation. A method may obtain a low CHAIR score not because it is more faithful to the image, but because it avoids mentioning objects, produces incomplete captions, or falls into repetitive decoding loops.

Table 2 reports the AMBER-g results. On Qwen3-VL-2B, Revis obtains the lowest CHAIR score of 4.4. However, its Coverage drops catastrophically from 75.7% to 44.1%. Qualitative inspection shows that Revis frequently enters autoregressive repetition loops, producing uninformative outputs with very few distinct object mentions. Since CHAIR penalizes hallucinated object mentions but does not reward completeness or informativeness, such degenerate captions can receive deceptively favorable hallucination scores.

SCAS avoids this failure mode. Although its CHAIR score is slightly higher than Revis by 0.3 points, it preserves much higher Coverage, 69.2% versus 44.1%. This indicates that SCAS suppresses hallucination while maintaining descriptive generation. On InternVL3.5-4B, SCAS reduces image-level hallucination Hal_i from 2.0 to **1.6** and sentence-level hallucination Hal_s from 32.7 to **31.9**, while maintaining comparable coverage. In contrast, Revis worsens both CHAIR and hallucination rates on InternVL3.5-4B. These results show that SCAS does not rely on under-generation or repetition collapse to improve hallucination behavior.

Table 2: Generative captioning results on AMBER-g. Best results are in **bold**, and second-best results are underlined.

Model	Method	CHAIR↓	Cover↑	$\text{Hal}_i \downarrow$	$\text{Hal}_s \downarrow$
Qwen3-VL-2B	Baseline	6.2	**75.7**	4.7	56.0
	VCD	7.3	77.5	4.8	61.8
	Revis	**4.4**	44.1	**2.0**	**27.1**
	SCAS (Ours)	4.7	69.2	2.6	45.0
InternVL3.5-4B	Baseline	**5.6**	**68.0**	2.0	32.7
	Revis	6.6	67.2	2.6	35.7
	SCAS (Ours)	5.7	66.5	**1.6**	**31.9**

Table 3: Ablation of steering coefficients α and β on AMBER-d and RePoPe subtasks on Qwen3-VL-2B (Acc / F1, %). Best results are in **bold**, and second-best results are underlined.

Variant Configuration	AMBER-d Subtasks				RePoPe Subtasks		
	Overall↑	Existence↑	Attribute↑	Relation↑	Random↑	Popular↑	Adversarial↑
(1) Baseline	86.27 / 89.14	89.03 / 94.19	85.27 / 84.34	**82.69 / 80.77**	96.14 / 95.39	93.55 / 92.76	92.29 / 91.58
(2) Supp. Only (α=1.0, β=0.0)	86.41 / 89.54	90.98 / 95.27	85.30 / 84.87	77.95 / 77.79	**96.21 / 95.45**	93.84 / 93.01	92.70 / 91.91
(3) Enhanc. Only (α=0.0, β=1.0)	**86.88** / 89.78	90.41 / 94.96	**85.76** / 85.16	81.55 / 80.25	96.03 / 95.24	93.73 / 92.90	92.59 / 91.80
(4) Full SCAS (α=1.0, β=1.0)	86.68 / **89.86**	**91.71 / 95.67**	85.72 / **85.54**	76.20 / 76.73	96.00 / 95.17	**93.99 / 93.12**	**92.92 / 92.08**

4.4 Ablation Study: What Does Each Component Protect?

We conduct ablations on Qwen3-VL-2B to examine how each component governs specific failure modes—evaluating steering intensities (α, β), Yes/No debiasing, and orthogonalization constraints. The results confirm that our designs not only suppress hallucinations, but crucially prevent shortcut negation and preserve multimodal reasoning.

Symmetrical steering dimensions (α and β) are complementary. Table 3 report the effects of α and β on AMBER-d and RePoPe. Solely activating deviation-subspace suppression, i.e., α=1.0, β=0.0, effectively clears out absolute hallucinatory features. On AMBER Existence, it increases F1 from 94.19% to 95.27%. On RePoPe Adversarial, it improves F1 from 91.58% to 91.91%. This supports our hypothesis that coarse-grained entity existence boundaries are strongly encoded in absolute hidden-state manifolds.

However, isolated suppression can overly prune valid semantic spaces, introducing distortions to fine-grained compositional reasoning. As shown in Table 3, deviation suppression alone decays AMBER Relation F1 from 80.77% to 77.79%. Conversely, activating factual semantic enhancement alone (α=0.0, β=1.0) plays a stabilizing role, maintaining Relation F1 at 80.25% while yielding moderate gains on existence metrics. While Full SCAS (α=1.0, β=1.0) exhibits a minor trade-off on fine-grained Relation F1 (76.73%), this spatial synergy is critical for driving out pervasive polar hallucinations, ultimately yielding peak overall performance across RePoPe and AMBER existence tasks.

Yes/No debiasing prevents shortcut negation. Table 4 studies the effect of our debiasing and orthogonalization constraints. Removing Yes/No debiasing slightly increases aggregate AMBER and RePoPe scores; however, this improvement is misleading. Removing debiasing severely biases the model toward negation (63.1% "No" predictions), plunging AMBER Relation Accuracy from 82.69% to 75.18%. This verifies that debiasing eliminates task-prior shortcuts rather than just shifting answer distributions.

This confirms that the raw calibration direction contains not only hallucination-related semantics, but also a dataset-level answer prior. Yes/No debiasing removes this shortcut component and prevents SCAS from becoming a blind negation mechanism. Therefore, debiasing is essential not because it maximizes aggregate accuracy, but because it preserves the semantic validity of the intervention.

Asymmetric orthogonalization preserves utility. We also compare asymmetric orthogonalization (Mode A, which orthogonalizes the factual subspace $\mathbf{U}_{\text{pos}}$ with respect to $\mathbf{U}_{\text{neg}}$) against symmetric orthogonalization of both subspaces (Mode B). The two variants achieve nearly identical AMBER

Table 4: Impact of Yes/No debiasing and orthogonalization space on the safety-utility trade-off on Qwen3-VL-2B (evaluated at Layer 15 with $\alpha = 1.0, \beta = 1.0$). Metrics are reported as Accuracy / F1-score for safety benchmarks.

Variant Configuration	Safety Performance Acc / F1		Utility Performance	
	AMBER Overall	RePoPe Adv.	MMBench Acc	MME Total
(1) Debias SCAS (`ortho=pos`)	86.68 / 89.86	92.92 / 92.08	84.62%	**2101.92**
(2) Debias SCAS (`ortho=both`)	86.66 / 89.84	92.92 / 92.08	84.62%	2093.68
(3) Non-Debiased SCAS (`ortho=pos`)	**86.70 / 89.93**	**93.00 / 92.12**	**84.71%**	2066.43

Table 5: Cross-scale supplementary validation on Qwen3-VL-8B.

Method	MME Total↑	MMBench Acc↑	RePoPe Adv. Acc↑	RePoPe Adv. F1↑
Baseline	2402.68	90.38%	92.18	91.45
SCAS (α=1.0, β=1.0)	**2407.66**	**90.64%**	**92.40**	**91.68**

and RePoPe safety scores. However, symmetric orthogonalization lowers MME from 2101.92 to 2093.68. This suggests that the negative subspace may still contain useful grounded visual evidence. Aggressively removing components aligned with both sides can therefore damage general perception and reasoning. SCAS adopts asymmetric orthogonalization as a softer constraint, suppressing hallucination-related directions while preserving useful grounded semantics.

4.5 Supplementary Scaling Experiment on Qwen3-VL-8B

Finally, we evaluate whether SCAS remains stable on a larger VLM. Table 5 reports supplementary results on Qwen3-VL-8B under Layer 20 intervention. SCAS improves MMBench accuracy from 90.38% to **90.64%** and MME from 2402.68 to **2407.66**. At the same time, it improves RePoPe Adversarial Accuracy from 92.18% to **92.40%** and F1 from 91.45% to **91.68%**. Although the absolute gains are moderate, this experiment shows that SCAS does not rely on weaknesses of small models and remains non-destructive on a stronger 8B-scale VLM.

5 Conclusion and Limitations

In this paper, we presented Symmetric Contrastive Activation Steering (SCAS), a training-free, test-time framework designed to improve representation controllability and address entangled capabilities in Vision-Language Models. By extracting and orthogonalizing a six-state activation profile, SCAS partially decouples factual steering directions from confounding multimodal biases and decision-style artifacts. Evaluations on Qwen3-VL-2B, Qwen3-VL-8B, and InternVL3.5-4B demonstrate that SCAS consistently mitigates multimodal hallucinations on REPOPE and AMBER, while simultaneously improving general reasoning capabilities on MME and MMBench.

Limitations and Future Work. Despite its effectiveness, SCAS has several limitations that represent promising directions for future study. First, while multiple subspace extraction methods exist (e.g., PCA, mean shift), determining the mathematically optimal strategy remains highly empirical and relies on trial-and-error validation. Second, our layer selection strategy is largely heuristic and lacks deep theoretical interpretability. Third, calibration remains dependent on paired task-specific samples, and the steering intensities require manual empirical tuning. Future work will focus on automatically discovering semantic subspaces, developing adaptive layer-wise modulation, and exploring the framework's scalability to larger multi-modal foundation models.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Jinze Bai, Shuang Shu, Jiabo Zhou, et al. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

Zhe Chen, Jiannan Wu, Wenhai Wang, et al. How far are we to GPT-4V? closing the gap with open-source vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Wenliang Dai, Chunyuan Li, et al. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Chaoyou Fu, Peixian Zhang, et al. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

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 (CVPR)*, 2024.

InternVL Team. InternVL-3.5: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2601.00000*, 2026.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14558, 2025.

Sicong Leng, Kun Zhang, et al. Mitigating object hallucination in large vision-language models via visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Kenneth Li, Ojas Patel, Fernanda Viégas, Martin Wattenberg, and Neel Nanda. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 292–305, 2023b.

Zujie Li et al. HALC: Object hallucination reduction via local-to-global contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

Yuan Liu, Haodong Duan, et al. MMBench: Is your multi-modal model all-in-one? *arXiv preprint arXiv:2307.06281*, 2023b.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 17359–17372, 2022.

Yannic Neuhaus and Matthias Hein. RePOPE: Impact of annotation errors on the POPE benchmark. *arXiv preprint arXiv:2504.15707*, 2025.

Qwen Team. Qwen3-VL: A comprehensive vision-language model. *arXiv preprint arXiv:2601.00000*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 15504–15522, 2024.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, 2017.

Junyang Wang, Mengting Wang, et al. AMBER: An LLM-free multi-dimensional benchmark for MLLM hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 9440–9450, 2024.

Jialin Wu, Wei Shi, Han Shen, Peigui Qi, Kunsheng Tang, Zhicong Huang, Binghao Wang, and Zhou Yang. Revis: Sparse latent steering to mitigate object hallucination in large vision-language models. *arXiv preprint arXiv:2602.11824*, 2026.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Shuo Zhao, Jiaming Guo, et al. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Sahil Shashidhar, Katherine Zhu, Brittany Yee, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Hardware and Implementation Details

Hardware and Computing Environment. All experiments and evaluations are conducted locally on a workstation equipped with $4\times$ NVIDIA GeForce RTX 4090 GPUs (24GB VRAM per GPU). Each evaluation task is executed independently on a single GPU under `bfloat16` precision, confirming that our framework is deployable in standard consumer-grade environments without requiring enterprise server clusters.

Visual Token Budget and Decoding Settings. Our codebase is implemented in PyTorch and HuggingFace Transformers. All evaluations employ deterministic greedy decoding (T=0). To avoid inference latency bottlenecks caused by excessively long visual token sequences on high-resolution inputs, we explicitly regulate the visual token budget across both model families. Specifically, for **Qwen3-VL**, we constrain the dynamic image resolution using `min_pixels` ($64 \times 28 \times 28$) and `max_pixels` ($512 \times 28 \times 28$) such that the visual token count fed into the LLM backbone is strictly bounded between 64 and 512. Similarly, for **InternVL3.5**, we configure dynamic patch partitioning to align with an equivalent visual token budget. This guarantees a fair trade-off between fine-grained visual perception and computational throughput across all benchmarks.

Intervention Configuration. For SCAS, calibration directions are extracted from the OH-100 dataset. We utilize the `both` projection strategy combining the mean direction with PCA components of rank k=8, with Yes/No debiasing applied before PCA decomposition. Unless otherwise specified, default steering hyperparameters are set to α=1.0, β=1.0, and `ortho_mode=pos`, intervening strictly during the autoregressive decoding phase (`scope=decode`).

B Clip Detail

To prevent the steering operation from shifting the activations outside the model's nominal representation manifold, we apply a relative norm constraint to the perturbation vector $\mathbf{\Delta} = \mathbf{H}_{\text{steered}} - \mathbf{H}$. If the norm of the perturbation exceeds a specified fraction C of the original hidden state norm, we scale it back:

$$\mathbf{\Delta}_{\text{clipped}} = \mathbf{\Delta} \cdot \min\left(1.0, \frac{C \cdot \|\mathbf{H}\|}{\|\mathbf{\Delta}\| + \epsilon}\right) \tag{8}$$

The final representation passed to subsequent layers is:

$$\mathbf{H}_{\text{final}} = \mathbf{H} + \mathbf{\Delta}_{\text{clipped}} \tag{9}$$

where $\epsilon = 10^{-8}$ is a small constant for numerical stability. This constraint preserves the semantic consistency and generation quality of the model while applying the factual correction.

C Qualitative Examples on AMBER-g

In Section 4.3, we observed that aggressive activation-editing methods, such as Revis, may report artificially low CHAIR scores while suffering from severe generation degeneration. These methods can avoid explicit hallucinated object nouns by producing repetitive loops, overly generic descriptions, or prematurely terminated captions. Such behavior reduces the number of evaluated object mentions, but it also sacrifices coverage, fluency, and descriptive completeness.

Figure C.1 provides a representative AMBER-g example comparing the baseline model, Revis, and SCAS. The comparison shows three distinct behaviors: the baseline model is fluent but susceptible to priors; Revis suppresses hallucinations at the expense of severe repetition collapse; while SCAS preserves descriptive completeness and visual fidelity.

D Analysis of Intrinsic Latent Manifold Overlap

We observed that the initial positive and negative factual basis matrices, denoted as $\mathbf{U}_{\text{pos_init}}$ and $\mathbf{U}_{\text{neg_init}}$, exhibit substantial mutual subspace overlap. This suggests that positive and negative factual directions are not naturally disentangled in the raw latent space. Instead, they partially share a sign-agnostic semantic backbone, which may encode general visual salience, objectness, or topical information.



(a) Input image from AMBER-g, ID: 204.

Baseline (Qwen3-VL-2B): Object Hallucination & Premature Termination

This is a black and white photograph capturing a quiet moment in a bedroom. The image focuses on a classic twin-bell alarm clock resting on a wooden nightstand. In the background, the bed is visible, with white sheets and **a person's dark clothing visible in the background [Object Hallucination]**. The clock face is black, and shows the **time as approximately 10:10 [Time Hallucination / Prior Bias]**. In the background, **a person's dark clothing, possibly pants... The overall [Premature termination]**

Revis (Qwen3-VL-2B): Repetition Collapse

This is a detailed description... The alarm clock is the central subject... indicating the **time is 10:10 [Time Hallucination / Prior Bias]**. The background is the bed covered with a light sheet. **The bed is not a clean, tidy bed, but rather a bed that is not being used. [Endless Repetition Loop] . . . repeated 6 times . . . [Repetition collapse]**

SCAS (Ours): Factual, Complete & High-Recall Description

This is a black and white photograph of a classic twin-bell alarm clock placed on a bedside table. The clock face is dark, with white numerals and hands, and the **time displayed is approximately 10:55 [Accurate hands alignment, avoiding 10:10 stereotype]**. The background is softly blurred, showing rumpled white sheets and **a dark fabric, suggesting a comforter [Factual description, avoids hallucinating a non-existent person]**.

Figure C.1: **Qualitative comparison on AMBER-g (ID: 204).** The Baseline model fabricates non-existent persons; Revis collapses into repetitive loops; SCAS preserves fluency, factual grounding, and descriptive coverage.

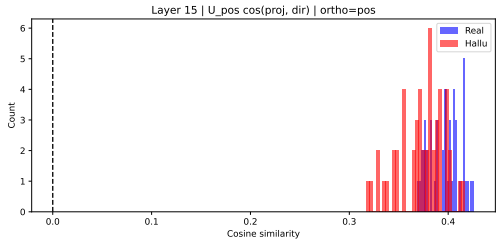
To quantify this property, we define the normalized subspace overlap between two arbitrary subspaces $\mathcal{S}_{\text{pos}}$ and $\mathcal{S}_{\text{neg}}$ as:

$$O(\mathcal{S}_{\text{pos}}, \mathcal{S}_{\text{neg}}) = \frac{1}{\min(k_{\text{pos}}, k_{\text{neg}})} \operatorname{Tr}\left(\mathbf{P}_{\text{pos}} \mathbf{P}_{\text{neg}}\right), \tag{10}$$

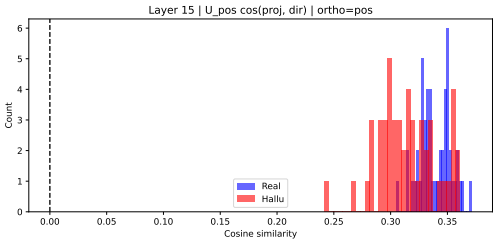
where $\mathbf{P}_{\text{pos}} = \mathbf{U}_{\text{pos}} \mathbf{U}_{\text{pos}}^{\top}$ and $\mathbf{P}_{\text{neg}} = \mathbf{U}_{\text{neg}} \mathbf{U}_{\text{neg}}^{\top}$ are the orthogonal projection matrices onto $\mathcal{S}_{\text{pos}}$ and $\mathcal{S}_{\text{neg}}$, respectively.

Figure D.1 compares the cosine-similarity distributions at Layer 15. Without orthogonalization, simultaneous multi-directional steering causes destructive interference. The `both` configuration symmetrically residualizes positive and negative bases against each other, isolating sign-specific factual components.

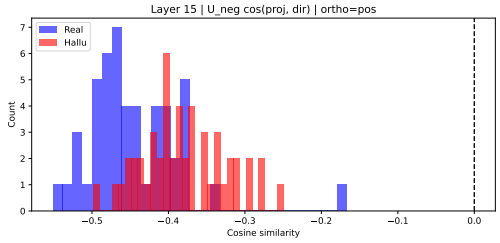
Figure D.2 further confirms that intermediate layers possess substantial semantic overlap, justifying our orthogonal steering formulation in SCAS.



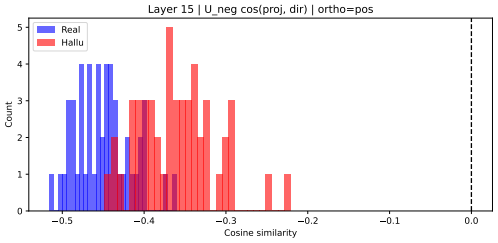
(a) $\mathbf{U}_{\text{pos}}$ under the `vec50_2, data50_1` configuration.



(b) $\mathbf{U}_{\text{pos}}$ under the `pos` (`vec50_1, data50_2`) configuration.

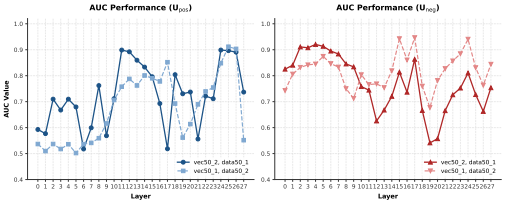


(c) $\mathbf{U}_{\text{neg}}$ under the `vec50_2, data50_1` configuration.

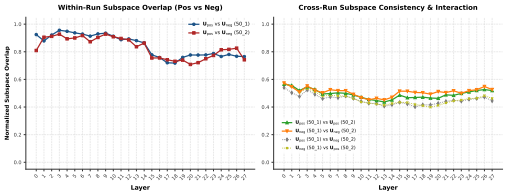


(d) $\mathbf{U}_{\text{neg}}$ under the `pos` (`vec50_1, data50_2`) configuration.

Figure D.1: **Cosine-similarity distributions of positive and negative projection bases under different configurations.**



(a) Layer-wise classification AUC (U_{pos} vs. U_{neg}).



(b) Subspace overlap and cross-run consistency.

Figure D.2: **Performance and geometric analysis of extracted subspaces across layers.** (a) Factual classification AUC of U_{pos} and U_{neg}. (b) Normalized subspace overlap (> 0.7, peak ≈ 0.95) versus random baseline ($k/d \approx 0.0044$).

E Implementation Details and Yes/No Debiasing Setup

This section provides implementation details omitted from the main paper, including the construction of the yes/no response-style axis and the intervention-layer selection protocol.

E.1 Yes/No Probe Question Template

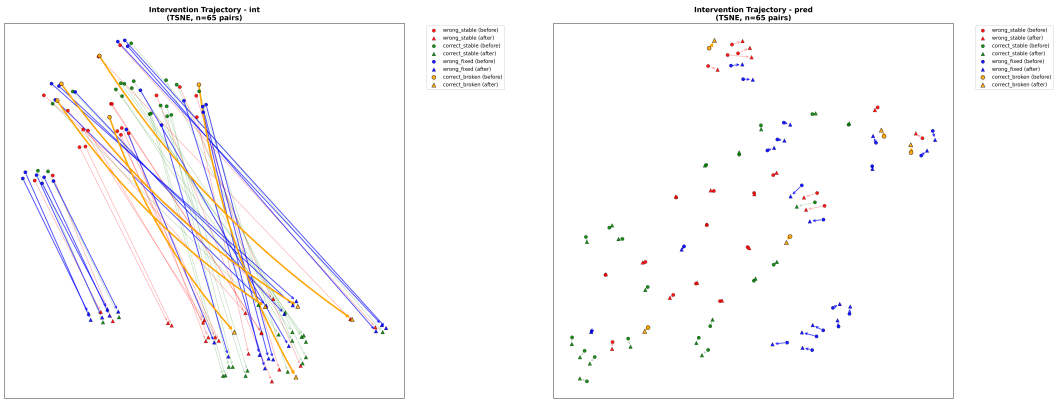
To construct the affirmation-negation decision axis $\mathbf{v}_{\text{yn}}$, we use $M = 20$ balanced common-sense questions (Table E.1) to isolate the stylistic tendency toward affirmative or negative answers from visual content.

E.2 Hyperparameter Sensitivity and Layer Selection

We evaluate SCAS under different intervention layers and steering strengths. In practice, intervening at middle-to-late decoder layers yields the best balance between factual steering and generation fluency.

Table E.1: **Balanced probe questions for extracting $\mathbf{v}_{\text{yn}}$.**

Yes	No
Is ice cold?	Is fire cold?
Does water flow downward?	Do stones float naturally on water?
Is the sun a star?	Is the moon a star?
Do humans breathe oxygen?	Do humans breathe only carbon dioxide?
Are triangles three-sided shapes?	Are squares three-sided shapes?
Does paper burn in fire?	Does iron melt at room temperature?
Is grass generally green?	Is snow generally pink?
Does a year contain twelve months?	Does a week contain ten days?
Is concrete a solid material?	Is water solid at room temperature?
Are oceans filled with saltwater?	Are clouds made of solid iron?



(a) Intervention-layer representation $\mathbf{z}_{\text{int}}$. (b) Prediction-layer representation $\mathbf{z}_{\text{pred}}$.

Figure E.1: **Visualization of representation trajectories under causal activation intervention.** Circles denote pre-intervention states, triangles denote post-intervention states, and arrows show sample-wise shifts induced by SCAS.

Figure E.1 visualizes how the intervention modifies representations at the intervention layer versus the final prediction layer. In our experiments, the most stable performance is obtained when applying interventions at Layers 15–20 for Qwen3-VL-2B and Layers 22–26 for InternVL3.5-4B.

F Limitations and Failure Modes

We summarize the main limitations of SCAS below.

Dependence on pre-existing visual knowledge. SCAS steers latent representations already learned by the underlying VLM. It cannot introduce visual knowledge that is entirely absent from the pre-trained vision encoder.

Offline calibration overhead. Constructing factual steering subspaces requires an offline calibration phase on a small set of curated examples.

Layer-selection heuristic. A fixed layer choice may not be universally optimal for every prompt or context. Dynamic layer-adaptive modulation is a promising future direction.

Handling compositional hallucinations. SCAS mainly targets factual directions associated with object presence. Extending the method to complex compositional relations and actions remains an important direction for future work.

G AI DISCLOSURE

Large Language Models (LLMs) were utilized to assist in drafting and polishing sections of this manuscript based on outlines, experimental results, and methodology designed by the authors. All generated content was thoroughly reviewed, verified, and edited by the authors to ensure factual and technical accuracy.