

# Spectral Query-Key Product Weight Steering for Training-Free VLM Hallucination Mitigation

Karn Tiwari<sup>1</sup>, Varnith Chordia<sup>2</sup>, Prathosh AP<sup>1, 3</sup>

<sup>1</sup>Indian Institute of Science, Bengaluru, India, <sup>2</sup>Snap Inc., <sup>3</sup>LatentForce.ai

Correspondence: [karntiwari@iisc.ac.in](mailto:karntiwari@iisc.ac.in), [vhordia@snapchat.com](mailto:vhordia@snapchat.com), [prathosh@iisc.ac.in](mailto:prathosh@iisc.ac.in)

## Abstract

Vision-language models (VLMs) often generate fluent but visually unsupported descriptions, especially by mentioning objects absent from the image. We propose *QK Product Steering*, a data-free, training-free, and zero-inference-cost weight edit for reducing object hallucination. The method directly edits the per-head query-key product, the operator that produces pre-softmax attention logits, by suppressing a small number of dominant singular modes in selected middle layers. The edited product is then mapped back to the query weights through a closed-form query-only update while keeping shared key weights fixed, making the edit compatible with grouped-query attention. We further decompose the QK product into symmetric and antisymmetric components to distinguish mutual content-similarity patterns from directional attention patterns. Across three GQA-based VLMs, QK Product Steering achieves an average relative CHAIR<sub>s</sub> reduction of 5.92%, while matched random-mode controls show negligible change. Interpretability ablations show that the hallucination signal is specific to dominant QK modes and is primarily localized to the symmetric mutual-attention channel. Overall, QK Product Steering offers a simple alternative to decoding-time mitigation, requiring no additional data, fine-tuning, or inference-time overhead while largely preserving general multimodal capability.

## 1 Introduction

Vision-language models (VLMs) have become increasingly capable of generating fluent descriptions and answers from visual inputs. However, they still suffer from *object hallucination*: generating objects that are not present in the image. This failure is especially problematic because hallucinated descriptions are often coherent and confident, making them difficult to detect. Reducing object hallucination is therefore essential for building reliable and trustworthy VLMs.

A key source of this problem is the competition between visual evidence and learned vision-language priors. During generation, a VLM must decide whether to rely on the image or on object co-occurrence patterns learned during large-scale pretraining. When these priors dominate, the model may mention plausible objects that fit the textual context but are unsupported by the visual input. Effective hallucination mitigation should therefore weaken such prior-driven behavior while preserving the model’s general multimodal capability.

Existing methods address this problem mainly through additional training or inference-time intervention. Training-based approaches require extra data, supervision, and compute, making them expensive to apply across models. Training-free decoding methods avoid fine-tuning, but they alter generation through contrastive decoding, logit adjustment, attention reweighting, or penalty terms, introducing additional cost at every decoding step. A more practical alternative is a one-shot weight edit: modify the model once and then keep inference unchanged.

In this work, we propose QK Product Steering, a data-free, training-free, and zero-inference-cost weight-editing method for mitigating object hallucination in VLMs. Our key idea is to edit the query-key product, the operator that directly determines pre-softmax attention logits. Instead of modifying decoding, QK Product Steering identifies dominant singular modes of the per-head query-key product in selected middle layers and suppresses a small number of them. The edited product is then projected back into the query weights through a closed-form query-only update while keeping the shared key weights unchanged, making the method compatible with grouped-query attention.

The motivation behind QK Product Steering is that dominant query-key modes may capture high-energy attention directions that behave like generic vision-language priors. These directions support

fluent generation, but they can also overpower image evidence and produce visually unsupported objects. By damping only a small number of dominant modes in middle layers, QK Product Steering weakens these overactive priors while preserving most of the model’s original computation.

Beyond the main spectral edit, we further decompose the query-key product into symmetric and antisymmetric components. This allows us to separate *mutual* content-similarity patterns from *directional* attention patterns. Our analysis shows that the hallucination signal is primarily concentrated in the symmetric mutual-attention channel, while the antisymmetric channel carries little detectable hallucination effect. This provides a mechanistic explanation for why suppressing dominant query-key modes reduces hallucination: the edit weakens strong mutual vision-language priors rather than arbitrarily perturbing the model. Our contributions are summarized as follows:

- We introduce *QK Product Steering*, a data-free, training-free, and zero-inference-cost method for mitigating object hallucination in VLMs via a one-shot query-key product weight edit.
- We propose a GQA-safe query-only projection that realizes the edited query-key product while keeping shared key weights unchanged, preserving the standard inference procedure.
- We provide controlled ablations and symmetric/antisymmetric decomposition analysis showing that hallucination is specific to dominant QK modes and is primarily localized to the symmetric mutual-attention channel.

## 2 Related Work

**Large Vision-Language Models.** Large vision-language models (LVLMs) combine visual encoders with language models through cross-attention, query transformers, or multimodal instruction tuning, enabling open-ended captioning and visual question answering (Alayrac et al., 2022; Li et al., 2023a; Dai et al., 2023; Liu et al., 2023b; Bai et al., 2023). Recent models such as Qwen2.5-VL, InternVL3, and Pixtral-12B improve high-resolution perception and multimodal reasoning (Bai et al., 2025; Zhu et al., 2025a; Agrawal et al., 2024), but their generative flexibility also makes them prone to visually unsupported outputs.

**Object Hallucination Evaluation.** Object hallucination refers to generating objects not grounded in

the image. CHAIR evaluates hallucinated COCO object mentions in captions (Rohrbach et al., 2018; Lin et al., 2014), while POPE probes object existence through yes/no questions and highlights the role of object-frequency and co-occurrence priors (Li et al., 2023b). HallusionBench further tests visual-dependent and counterfactual reasoning (Guan et al., 2024), and benchmarks such as AMBER, MMHal-Bench, and M-HalDetect extend evaluation to broader factuality settings (Wang et al., 2023; Sun et al., 2024; Gunjal et al., 2023). We use CHAIR for long-form caption hallucination and report POPE, HallusionBench, MME, and MMMU to assess grounding and general capability.

**Hallucination Mitigation.** Training-based methods reduce hallucination through instruction tuning, preference optimization, or correction data (Liu et al., 2023a; Xiao et al., 2024; Yue et al., 2024b; Sun et al., 2024). Training-free methods avoid fine-tuning but usually modify inference: VCD contrasts logits from clean and perturbed images (Leng et al., 2024), OPERA penalizes over-trust during beam search (Huang et al., 2024), PAI increases image-token influence (Liu et al., 2024), Woodpecker verifies and corrects generated claims (Yin et al., 2023), and Summary-Guided Decoding reduces language-prior accumulation (Min et al., 2025). In contrast, QK Product Steering edits the model once and leaves the standard decoding path unchanged.

**Steering and Mechanistic Interventions.** Recent work localizes hallucination to activations, attention heads, and modality-specific subspaces. VTI, Activation Steering Decoding, VISTA, and Dynamic Multimodal Activation Steering intervene in hidden states or truthfulness directions (Liu et al., 2025; Su et al., 2025; Li et al., 2025; Yin et al., 2026). Attention-centered studies identify hallucination-relevant heads and layers, with middle layers playing an important role in visual information propagation (Yang et al., 2025b; Jiang et al., 2025; Zhu et al., 2025b); ASCD steers attention scores during contrastive decoding (Wang et al., 2025). Our method instead edits the query-key product itself, the bilinear operator that produces attention logits.

**Weight Editing and Grouped-Query Attention.** Modern decoder backbones often use grouped-query attention (GQA), where multiple query heads share key/value heads to reduce KV-cache cost (Ainslie et al., 2023). This sharing makes direct key-weight editing risky, since changing a shared key can unintentionally affect several query heads. Prior weight-editing methods such as Nullu reduce

hallucination without inference overhead, but rely on prompt-derived hallucination subspaces (Yang et al., 2025a). In contrast, QK Product Steering is data-free: it edits dominant spectral modes of each head’s query-key product and realizes the edit through a query-only update, preserving shared keys and standard inference.

### 3 Background and Preliminaries

**Query-Key Product in Attention.** Consider an attention layer with query head  $h$  and its assigned key head  $g$ . Let  $W_{q,h}, W_{k,g} \in \mathbb{R}^{r \times d}$  denote the query and key projection matrices, where  $r$  is the head dimension and  $d$  is the hidden dimension. Given hidden states  $X \in \mathbb{R}^{n \times d}$ , we have

$$Q_h = XW_{q,h}^\top, \quad K_g = XW_{k,g}^\top. \quad (1)$$

The pre-softmax attention logits are

$$Q_h K_g^\top = XW_{q,h}^\top W_{k,g} X^\top = X M_h X^\top, \quad (2)$$

where

$$M_h = W_{q,h}^\top W_{k,g} \in \mathbb{R}^{d \times d} \quad (3)$$

is the query-key product. Thus,  $M_h$  is the effective operator controlling the attention interaction. Since  $M_h$  is formed from two rank- $r$  factors,  $\text{rank}(M_h) \leq r$ , so each head admits a low-rank spectral description.

**Symmetric and Antisymmetric (skew-symmetric) Decomposition.** Any square QK product  $M_h$  can be uniquely decomposed as:

$$M_h = S_h + \Omega_h, \quad (4)$$

$$S_h = \frac{1}{2}(M_h + M_h^\top), \quad \Omega_h = \frac{1}{2}(M_h - M_h^\top), \quad (5)$$

where

$$S_h^\top = S_h, \quad \Omega_h^\top = -\Omega_h. \quad (6)$$

The two components lie in orthogonal subspaces under the Frobenius inner product:

$$\langle S_h, \Omega_h \rangle_F = \text{tr}(S_h^\top \Omega_h) = 0. \quad (7)$$

Therefore, the energy of the QK product decomposes additively:

$$\|M_h\|_F^2 = \|S_h\|_F^2 + \|\Omega_h\|_F^2. \quad (8)$$

This decomposition separates two attention channels. For token representations  $x_i$  and  $x_j$ , the logit contribution satisfies

$$x_i^\top M_h x_j = x_i^\top S_h x_j + x_i^\top \Omega_h x_j, \quad (9)$$

where the symmetric term is reciprocal,

$$x_i^\top S_h x_j = x_j^\top S_h x_i, \quad (10)$$

while the antisymmetric term changes sign under direction reversal,

$$x_i^\top \Omega_h x_j = -x_j^\top \Omega_h x_i. \quad (11)$$

Hence,  $S_h$  captures mutual content-similarity patterns, whereas  $\Omega_h$  captures directional attention preferences.

### 4 Proposed Method

We propose *QK Product Steering*, a data-free and training-free weight edit for mitigating object hallucination in VLMs. The method is applied once to a pretrained model and then leaves inference unchanged, requiring no auxiliary model, fine-tuning, or decoding-time intervention.

**Motivation.** The core idea is to steer the query-key product, which directly controls the pre-softmax attention logits. We hypothesize that its dominant spectral directions encode strong vision-language priors: useful for fluent generation, but harmful when they override visual evidence and produce plausible yet absent objects. QK Product Steering therefore suppresses a small number of dominant modes in selected middle-layer attention heads.

Concretely, for each selected head, we compute the thin SVD of the query-key product, damp its top singular modes, and project the edited product back into the query weights through a closed-form query-only update. The key weights are kept fixed, making the edit compatible with grouped-query attention. This yields a localized spectral intervention that weakens prior-driven attention patterns while preserving most of the model’s original computation; an overview is shown in Figure 1.

**Thin SVD of the QK Product.** Since  $M_h \in \mathbb{R}^{d \times d}$  has rank at most  $r$ , we compute its thin SVD from the smaller factors. Using economy QR decompositions,

$$W_{q,h}^\top = Q_q R_q, \quad W_{k,g}^\top = Q_k R_k, \quad (12)$$

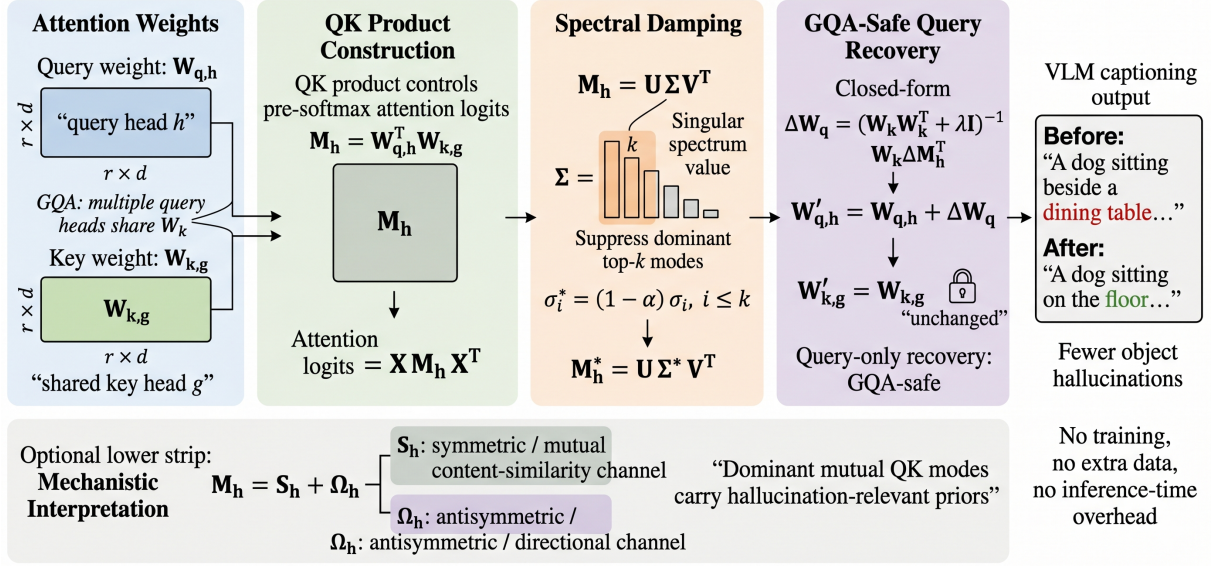


Figure 1: **Overview of QK Product Steering.** For each selected query head and its assigned shared key head, we form the QK product  $M_h = W_{q,h}^T W_{k,g}$ , suppress its dominant singular modes, and recover an edited query weight through a closed-form query-only projection. The shared key weight is kept unchanged, making the edit compatible with grouped-query attention. The same framework also enables symmetric and antisymmetric QK decompositions to identify whether hallucination-relevant priors arise from mutual or directional attention channels.

where  $Q_q, Q_k \in \mathbb{R}^{d \times r}$  have orthonormal columns and  $R_q, R_k \in \mathbb{R}^{r \times r}$ . Then

$$M_h = W_{q,h}^T W_{k,g} = Q_q C_h Q_k^T, \quad C_h = R_q R_k^T. \quad (13)$$

Taking the SVD  $C_h = \hat{U} \Sigma \hat{V}^T$  gives

$$M_h = U \Sigma V^T, \quad U = Q_q \hat{U}, \quad V = Q_k \hat{V}. \quad (14)$$

Thus, we obtain the singular modes of  $M_h$  from an  $r \times r$  SVD instead of a  $d \times d$  decomposition.

**Dominant Mode Suppression.** Let

$$M_h = U \Sigma V^T, \quad (15)$$

be the thin SVD of the QK product, where  $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_r)$  with  $\sigma_1 \geq \dots \geq \sigma_r$ . We edit a mode set  $\mathcal{S} \subseteq \{1, \dots, r\}$ , with the default choice  $\mathcal{S} = \{1, \dots, k\}$  corresponding to the top- $k$  modes. Given damping strength  $\alpha$ , we define

$$\Sigma_{ii}^* = \begin{cases} (1 - \alpha) \Sigma_{ii}, & i \in \mathcal{S}, \\ \Sigma_{ii}, & i \notin \mathcal{S}, \end{cases} \quad (16)$$

and reconstruct the edited product as

$$M_h^* = U \Sigma^* V^T. \quad (17)$$

Here  $\alpha = 0$  gives no edit, while  $\alpha = 1$  fully removes the selected modes. In our main setting,

we suppress the top singular modes in the middle layers, targeting the strongest attention directions while preserving most of the QK product structure. This produces an edited product  $M_h^*$  in the spectral space. To make the edit usable in the original VLM, we must realize  $M_h^*$  as an update to the model weights without disrupting shared key projections. **GQA-Safe Query Recovery.** After editing  $M_h$ , we recover model weights while keeping the shared key projection fixed. This avoids conflicts under grouped-query attention. Let

$$\Delta M_h = M_h^* - M_h. \quad (18)$$

We compute the smallest query-weight update that realizes the edited product:

$$\begin{aligned} \min_{\Delta W_{q,h}} \quad & \|\Delta W_{q,h}\|_F^2 \\ \text{s.t.} \quad & (W_{q,h} + \Delta W_{q,h})^T W_{k,g} = M_h^*. \end{aligned} \quad (19)$$

The ridge-stabilized closed-form update is

$$\Delta W_{q,h} = (W_{k,g} W_{k,g}^T + \lambda I)^{-1} W_{k,g} \Delta M_h^T, \quad (20)$$

$$\lambda = \epsilon \cdot \frac{1}{r} \text{tr}(W_{k,g} W_{k,g}^T). \quad (21)$$

The final recovered weights are:

$$W_{q,h}^* = W_{q,h} + \Delta W_{q,h}, \quad W_{k,g}^* = W_{k,g}. \quad (22)$$

| Variant      | Edit                                         | Product                      |
|--------------|----------------------------------------------|------------------------------|
| Sym-only     | Damp top- $k_s$ eigenmodes of $S_h$          | $M_h^* = S_h^* + \Omega_h$   |
| Antisym-only | Damp top- $k_a$ singular modes of $\Omega_h$ | $M_h^* = S_h + \Omega_h^*$   |
| Sym+Antisym  | Edit both $S_h$ and $\Omega_h$               | $M_h^* = S_h^* + \Omega_h^*$ |

Table 1: Symmetric and antisymmetric edit variants. Sym-only edits the mutual content-similarity channel  $S_h$ , Antisym-only edits the directional channel  $\Omega_h$ , and Sym+Antisym edits both.

Thus, the edited QK product is implemented through a query-only update, while the shared key weights remain unchanged.

**Symmetric and Antisymmetric Edit Variants.** Building on the decomposition in Section 3, we use the symmetric and antisymmetric components of  $M_h$  as diagnostic edit targets. These variants provide a mechanistic view of which attention channel contributes to hallucination. Given  $M_h = S_h + \Omega_h$ , we define three edit variants as shown in Table 1.

Each edited product  $M_h^*$  is mapped back to the query weights using the same GQA-safe query recovery step. These variants let us test whether hallucination is mainly associated with mutual content-similarity patterns in  $S_h$  or directional attention patterns in  $\Omega_h$  for mechanistic interpretability.

**Layer and Head Selection.** We apply QK Product Steering to selected layers and heads. Empirically, middle layers provide the best intervention point: early-layer edits can disrupt basic visual-textual representations, while late-layer edits have limited effect on object hallucination. Middle layers are therefore well positioned to weaken hallucination-related attention patterns while preserving caption quality. For each selected layer, we edit every query head using its assigned shared key head under grouped-query attention. The shared key weights are kept fixed throughout, and only the corresponding query weights are updated. A detailed discussion is provided in Section 5.

**Interpretation and Theoretical Insights.** *QK Product Steering* is motivated by the structure of attention itself. The QK product is the exact bilinear operator that produces pre-softmax attention logits, so editing it directly steers attention behavior. Its top singular modes correspond to the highest-energy low-rank directions of this operator, making their suppression a principled way to weaken dominant attention patterns rather than applying an arbitrary weight perturbation. In VLMs, these dominant directions can behave like generic vision-language

priors: useful for fluent generation, but harmful when they override visual evidence and produce unsupported objects. By suppressing only a few dominant modes in middle layers, QK Product Steering weakens such prior-driven attention while preserving most of the model’s computation.

The symmetric–antisymmetric decomposition further separates mutual content-similarity patterns from directional attention patterns, providing more interpretability. This provides a mechanistic way to test where hallucination-relevant priors reside. Finally, because the edit is low-rank, localized to selected middle layers, and recovered through a minimum-change query-only update, its off-target effect is bounded (more details in Appendix B).

**One-Time Offline Editing Complexity.** The proposed QK Product Steering requires only a one-time offline weight-editing cost. By exploiting the low-rank structure of the QK product, we avoid computing a full  $d \times d$  SVD and instead perform a thin SVD in the  $r$ -dimensional head space, where  $r$  is the attention head dimension. The resulting per-head editing cost is  $O(dr^2 + r^3)$ . Thus, for  $L_e$  edited layers and  $H_q$  query heads, the total one-time offline cost is  $O(L_e H_q (dr^2 + r^3))$ . This cost is incurred only once before deployment; after the edited weights are written back, the model uses the original inference procedure with no additional inference-time FLOPs, memory, KV-cache overhead, or decoding latency unlike previous methods.

## 5 Experiments

We evaluate QK Product Steering on three grouped-query-attention vision-language models: Qwen2.5-VL-7B (Bai et al., 2025), InternVL3-8B (Zhu et al., 2025a), and Pixtral-12B (Agrawal et al., 2024). Qwen2.5-VL-7B is the primary model where the full hyperparameter sweep is run; the other two are used to test cross-architecture generalization at each model’s preferred  $(k, \alpha)$ . Per-model architecture details (layer counts, head dimensions, GQA group sizes) and the hardware configuration used to run all experiments are deferred to Appendix C.4.

Our evaluation is generative object hallucination, measured with CHAIR (Rohrbach et al., 2018) on COCO 2014 val; all sweeps and ablations in the paper are scored on CHAIR. POPE (Li et al., 2023b) and HallusionBench (Guan et al., 2024) are reported alongside as probe-based hallucination benchmarks (templated yes/no and adversarial paired questions, respectively), and MME (Fu et al.,

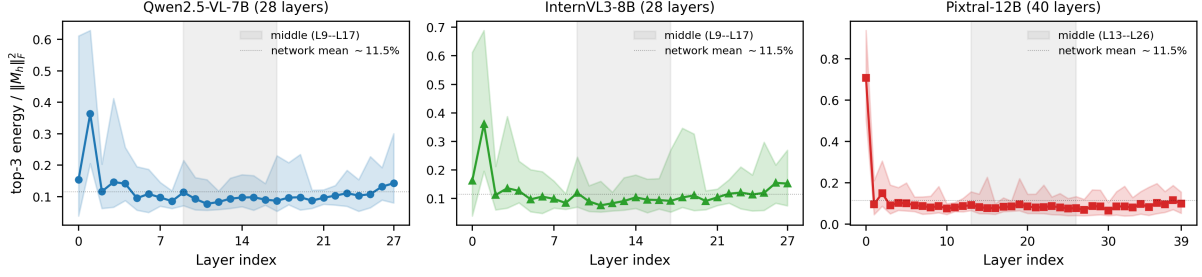


Figure 2: Per layer and head, the fraction of  $M_h$ 's Frobenius energy in its top-3 singular modes,  $E_3 = (\sigma_1^2 + \sigma_2^2 + \sigma_3^2) / \|M_h\|_F^2$ . Solid line: mean over query heads in each layer; shaded band: head min/max; grey vertical band: middle-third edit locus. A small  $E_3$  means QK Product Steering top-3 touches a small share of the operator (a surgical edit); a large  $E_3$  means it removes a large share (a destructive edit). Across all three models,  $E_3$  is highest in the first few layers and lowest in the middle third, where the same low-rank intervention is most surgical.

2023) and MMMU val (Yue et al., 2024a) are reported only to verify that general capability is preserved. Our probe and capability benchmarks are run through lmms-eval v0.7 (Zhang et al., 2025) with each model's reference inference config. Per-benchmark sample sizes, decoding settings, and scoring protocols are deferred to Appendix D.

Decoding is greedy throughout; hallucination deltas are reported with paired-bootstrap CIs (iteration count per experiment in Appendix C.3) and capability deltas as raw within-pipeline differences. Controls and protocol details are in Appendix D.

### 5.1 Middle Layer Analysis

**Why edit the middle layers? Top- $k$  surgicality across depth.** QK Product Steering damps the top- $k$  singular modes of  $M_h = W_q^\top W_k$ , and the natural a priori measure of how *surgical* that edit is at a given layer is the fraction of the operator's Frobenius energy carried by those modes,  $E_k(M_h) = \sum_{i \leq k} \sigma_i^2 / \|M_h\|_F^2$ . A head with  $E_3 = 0.10$  has 90% of its operator left untouched by QK Product Steering top-3; a head with  $E_3 = 0.36$  loses a third of the operator. We compute  $E_3$  per head for every query head across all three models and use it to localise the edit before running any CHAIR experiment.

Figure 2 reveals a single, model-consistent depth pattern:  $E_3$  rises sharply over the first few layers, falls to a model-internal minimum in the middle third, and creeps back up toward the output. Early layers are rank-concentrated, with  $E_3$  reaching as much as 0.36 on Qwen and 0.71 on Pixtral, so damping their top three modes removes a substantial fraction of the operator and collapses generation. The middle band is the unique cell where the same low-rank intervention is small relative to  $\|M_h\|_F^2$

and therefore structurally consistent with a surgical edit. We adopt the middle-third locus as the default edit band on this structural basis alone, before any CHAIR experiment.

### 5.2 Main Result

Table 2 reports QK Product Steering and its channel decompositions across three GQA vision-language models and six benchmarks, each at the model's preferred  $(k, \alpha)$  in the middle-third edit band. The ablations in Section 6 drive the  $(k=3, \alpha=1)$  default on Qwen; for cross-architecture transfer, we keep  $\alpha=1$  and the middle-third band fixed and select  $k$  per model by a small CHAIR sweep, with details in Appendix D.7. QK Product Steering top- $k$  is the lowest-CHAIR<sub>s</sub> non-control row on every model, while the matched random- $k$  control at the same  $k$  moves CHAIR<sub>s</sub> negligibly on all three – ruling out arbitrary low-rank perturbations of  $W_q$  and identifying the dominant query-key modes as the carrier of the hallucination signal. Paired-bootstrap CIs supporting this picture, including the symmetric-channel edit on Qwen and the QK Product Steering top- $k$  edit on Pixtral, are reported in Appendix C.3. The channel decompositions separate cleanly: Sym-only TOP captures most of the CHAIR<sub>s</sub> gain on Qwen and InternVL3, whereas Antisym-only TOP is the worst non-control row on CHAIR<sub>s</sub> across all three models – antisymmetric content carries no hallucination signal under this edit. Capability is broadly preserved within the random-control noise budget; the only systematic exception is the Pixtral middle band at  $k=3$ , where MMMU is sensitive to which spectral channel is touched (Sym best, Antisym worst). Per-method  $(k, \alpha)$  sweeps and bottom- $k$  / amplification controls are reported in Appendix D.

| Model         | Method                                          | Hallucination ( $\downarrow$ ) |                    | Probe ( $\uparrow$ ) |              | Capability ( $\uparrow$ ) |              |
|---------------|-------------------------------------------------|--------------------------------|--------------------|----------------------|--------------|---------------------------|--------------|
|               |                                                 | CHAIR <sub>s</sub>             | CHAIR <sub>i</sub> | POPE F1              | HB AAcc      | MME                       | MMMU         |
| Qwen2.5-VL-7B | Baseline                                        | 31.44                          | 65.26              | 86.18                | 64.30        | 2339.2                    | <b>51.56</b> |
|               | QK Product Steering random $k=3$                | 31.33                          | 64.79              | 85.33                | 64.30        | 2332.8                    | 51.33        |
|               | <b>QK Product Steering top <math>k=3</math></b> | <b>28.56</b>                   | <b>63.75</b>       | <b>86.68</b>         | <b>64.98</b> | 2293.3                    | 50.78        |
|               | Sym-only TOP $k_s=3$                            | 29.53                          | 65.75              | 86.50                | 63.68        | <b>2343.2</b>             | 50.89        |
|               | Antisym-only TOP $k_a=6$                        | 31.60                          | 66.73              | 82.87                | 61.12        | 2290.9                    | 49.44        |
| InternVL3-8B  | Baseline                                        | 33.81                          | 62.98              | 90.85                | 60.58        | <b>2388.0</b>             | 54.33        |
|               | QK Product Steering top $k=1$                   | 33.23                          | 62.98              | 91.09                | 60.14        | 2377.2                    | 54.67        |
|               | QK Product Steering random $k=3$                | 34.12                          | 63.24              | 89.01                | 59.89        | 2377.0                    | 54.81        |
|               | <b>QK Product Steering top <math>k=3</math></b> | <b>31.23</b>                   | <b>62.07</b>       | <b>91.26</b>         | <b>60.84</b> | 2382.2                    | <b>55.84</b> |
|               | Sym-only TOP $k_s=1$                            | 33.43                          | 63.20              | 91.14                | 60.67        | 2381.4                    | 55.78        |
|               | Antisym-only TOP $k_a=2$                        | 33.98                          | 63.79              | 90.88                | 59.79        | 2371.7                    | 54.89        |
| Pixtral-12B   | Baseline                                        | 34.14                          | 64.86              | 84.99                | 72.26        | 2009.1                    | 45.44        |
|               | QK Product Steering random $k=3$                | 34.00                          | 64.51              | 84.44                | 71.59        | 1963.2                    | <b>45.67</b> |
|               | <b>QK Product Steering top <math>k=3</math></b> | <b>33.81</b>                   | <b>63.04</b>       | <b>85.23</b>         | <b>72.93</b> | <b>2018.6</b>             | 43.11        |
|               | Sym-only TOP $k_s=3$                            | 34.10                          | 64.70              | 85.21                | 72.71        | 1975.2                    | 44.11        |
|               | Antisym-only TOP $k_a=6$                        | 35.09                          | 64.50              | 84.24                | 71.38        | 1995.4                    | 42.46        |

Table 2: **Main Results.** Results across three VLMs and six benchmarks. Paired-bootstrap CIs and statistical significance for CHAIR are reported in Appendix C.3; full benchmark configurations and evaluation details and comparison against representative training-free hallucination-mitigation methods, including VCD (Leng et al., 2024), VTI (Liu et al., 2025), and VAR (Jiang et al., 2025); detailed results are provided in the Appendix D.

## 6 Ablations

The spectral diagnostic of Section 5.1 identifies *which layers* carry the profile compatible with a small- $k$  surgical edit; it does not fix the mode count  $k$ , the damping strength  $\alpha$ , or empirically validate the layer choice. We close those gaps with three ablations on Qwen2.5-VL-7B: a layer-locus check, a  $k$ -sweep, and an  $\alpha$ -sweep. Together they fix the default configuration ( $k=3$ ,  $\alpha=1$ , middle) used in our main results (see Table 2).

**Layer Locus.** Table 3 matches the  $E_3$  prediction of Figure 2: early-band and full-stack edits collapse generation entirely (zero detectable COCO objects per caption); late-band edits preserve fluency but do not move CHAIR<sub>s</sub>; only the middle band converts the surgical operator footprint into a measurable hallucination reduction.

**Mode Count.** At fixed  $\alpha=1$  on Qwen, the symmetric-channel top- $k$  edit peaks at  $k_s=3$  on CHAIR<sub>s</sub> and saturates beyond it, while the antisymmetric channel is non-monotone and weak across all  $k_a$  tested (Appendix Table 6). Matched-rank *bottom- $k$*  controls are null on both channels (CIs include zero), confirming that the signal lives in the dominant modes of  $S_h$  rather than in arbitrary low-rank perturbations or in the antisymmetric channel. On the full joint  $(k, \alpha)$  grid (Appendix Section D.5), CHAIR<sub>s</sub> decreases monotonically with  $k$  at  $\alpha \geq 1$ , but at increasing capability cost, which we examine next.

| Layers        | $\Delta\text{CHAIR}_s$ | $\Delta\text{CHAIR}_i$ | avg_obj         |
|---------------|------------------------|------------------------|-----------------|
| early         | -28.81                 | -66.60                 | 0.00 (collapse) |
| <b>middle</b> | <b>-3.04</b>           | -2.06                  | 2.67            |
| late          | +0.09                  | -2.41                  | 2.62            |
| all           | -28.81                 | -66.60                 | 0.00 (collapse) |

Table 3: Layer sensitivity on Qwen2.5-VL-7B at  $k=3$ ,  $\alpha=1$ ,  $N=2000$ . avg\_obj: mean COCO objects per caption (baseline 2.54).

**Damping Strength and Capability Tradeoff.** At fixed  $k=3$  on Qwen, the Sym-only edit’s CHAIR<sub>s</sub> response is monotone in  $\alpha$ , with CIs excluding zero from  $\alpha=0.5$  onward (Appendix Table 7); Antisym-only on  $\Omega$  is flat across the sweep. Pushing past zero-out into  $\alpha > 1$  continues to lower CHAIR<sub>s</sub> on the joint grid, but at a measurable capability cost on MMMU and MME; the same cost appears, at smaller magnitude, when  $k$  is increased at safe  $\alpha=1$  (Appendix Table 9).  $\alpha$  is the larger lever on capability. We therefore report  $(k=3, \alpha=1)$  as the default: the smallest principled cell where the CHAIR drop is CI-significant and capability is preserved within the random-control noise budget.

## 7 Discussion and Analysis

We analyze how QK Product Steering affects the model beyond its behavioral improvements. We focus on three questions: (i) which spectral components are responsible for hallucination reduction, (ii) how the intervention propagates through the network, and (iii) whether the targeted modes are

associated with naturally occurring hallucinations before intervention.

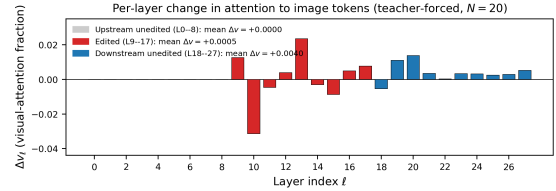
## 7.1 Symmetric Modes Drive Hallucination Reduction

QK Product Steering acts primarily on the symmetric part of the per-head QK product. On Qwen and InternVL3, damping the dominant eigenmodes of  $S_h$  alone reproduces roughly two-thirds of the CHAIR<sub>s</sub> reduction obtained by damping the full product; on Pixtral, the symmetric contribution is smaller. Editing only the antisymmetric part  $\Omega_h$  does *not* reduce hallucination and, across all three models, degrades probe and capability metrics relative to baseline (Table 2). Antisymmetric content therefore appears to carry no hallucination signal but still couples to general task performance, so editing it costs capability without any hallucination return. The behavioural consequence of that damping is non-local: at the edited layers themselves the layer-mean attention shift is essentially zero – per-head reorganisations cancel within the layer – and the change surfaces only in the unedited layers downstream of the edit site, as a small, uniform tilt of attention probability toward image tokens (Section 7.2). That per-step tilt is amplified across the roughly two-hundred decoding steps of a free-form caption into a CHAIR-level effect, but is too small to flip a single short answer on MME or MMMU, which is why hallucination drops while capability is preserved (Section 7.2).

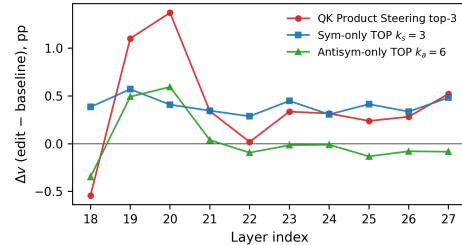
We further find that the intervention does not simply suppress the probability of hallucinated nouns at a fixed prefix or directly re-ground attention at the corresponding noun position under teacher forcing (Appendix D.8). These null results suggest that the observed behavioral effect is neither purely token-local nor position-local.

## 7.2 Propagation of Attention Shifts

QK Product Steering’s effect on attention has two structural features that make the discussion follow: it is *small* per layer and *non-local* in depth. The layer-mean image-attention shift at the edited layers is essentially zero; the signal appears as a uniform tilt toward image tokens in every unedited downstream layer, around 0.4pp per layer. Both features are essential: the locus tells us *where* the mechanism lives (downstream of the edit, not at the edit), and the magnitude tells us *which benchmarks* can resolve it (long-horizon captioning yes, single-decision MCQ no). We measure the shift under teacher forcing



(a) QK Product Steering top-3,  $N=200$ .



(b) Three spectral methods,  $N=200$ .

Figure 3: Per-layer attention-to-image shift  $\Delta v$  (edit – baseline, percentage points) under teacher forcing. **(Top)** (a) Full 28-layer profile for QK Product Steering: upstream is bit-exactly zero, edited band cancels in the layer mean, downstream is uniformly positive. **(Bottom)** (b) Downstream zoom across the three spectral methods: QK Product Steering and Sym-only both lift attention to image tokens (mean +0.4pp); Antisym-only is flat.

– identical [prompt + baseline caption] sequence through both edited and unedited models, so any attention difference is attributable to the weight edit alone.

Upstream of the edit ( $L_0$ – $L_8$ ) the shift is bit-exactly zero – a protocol check, since unedited layers with identical input must produce identical attention. Inside the edited band ( $L_9$ – $L_{17}$ ) QK Product Steering’s layer-mean is effectively zero (+0.05pp, sign-mixed). The signal lives in the downstream unedited layers ( $L_{18}$ – $L_{27}$ ): every one of them shifts in the same direction, with mean +0.4pp. Figure 3b resolves this layer-by-layer: the largest shifts are at  $L_{19}$ – $L_{20}$  immediately after the edit, taper through  $L_{22}$ , and hold a smaller positive offset to  $L_{27}$  – the decay shape of a perturbation that enters at  $L_{17}$  and propagates through subsequent attention and MLP blocks. Sym-only shows the same downstream pattern at comparable magnitude; Antisym-only, which is inert on CHAIR, has no such downstream signal (Appendix D.9, Figure 4).

The zero layer-mean inside the edited band reflects per-head cancellation rather than absence of effect: per-head shifts reach  $\pm 0.48$ pp with 138 positive and 114 negative cells across the 252 (layer, head) cells in the edited band, summing to a near-zero layer mean but a non-zero residual stream that the downstream layers respond to (see Appendix

Section D.9, Figure 5).

The per-layer magnitude of  $\Delta v$  is small by design of the mechanism, not by limitation of the edit. A shift of this size cannot flip a single short answer on MME or MMMU; applied at every one of the  $\sim 200$  greedy decoding steps of a free-form caption, it accumulates into the CHAIR-level effect we observe. The next subsection makes that asymmetry concrete.

**Why the attention shift reduces CHAIR but not MME or MMMU.** The per-layer image-attention shift of Section 7.2 is small and applies to every forward pass equally, yet it produces a clean CHAIR<sub>s</sub> reduction on free-form captioning while leaving MME and MMMU val essentially flat. Two factors account for the asymmetry. The first is decoding length: CHAIR scores captions of  $\sim 200$  greedy steps, so a small per-step bias toward image-grounded continuations accumulates over the trajectory and concentrates at the positions where the baseline model was borderline between a hallucinated and a grounded noun. MME and MMMU score a single short answer dominated by the recognition head over a handful of candidate tokens; there is no long generation to integrate over. The second is which spectral subspace QK Product Steering damps: the top modes of  $M_h$  are the high-energy, generic content-similarity defaults the model leans on when image evidence is weak – the same defaults that drive hallucination on free-form captioning. Damping them forces decoding toward image-specific signal without touching the lower-energy modes that the recognition-style benchmarks reward, so the cost surfaces as a free-form hallucination drop with capability preserved.

### 7.3 Pre-Intervention Spectral Evidence

To examine whether the spectral modes targeted by QK Product Steering are associated with naturally occurring hallucinations, we analyze 500 baseline captions from Qwen2.5-VL-7B on COCO using teacher forcing, with no intervention. At each object-mention position, we measure the logit-energy contribution of the top-3 QK singular modes toward image versus text tokens and compare hallucinated and grounded mentions within the same caption. Across 252 paired captions, hallucinated mentions show significantly lower top-3 image-directed energy ( $\Delta = -0.0091$ ,  $p \approx 10^{-16}$ ) and lower image-vs-text balance ( $\Delta = -0.0050$ ,  $p \approx 3 \times 10^{-10}$ ), whereas the bottom-3 control modes do not exhibit the same behavior. The

top-mode statistic provides a modest mention-level hallucination signal (AUC = 0.58), while the signal largely disappears when averaged over entire captions (AUC  $\approx 0.48$ ), indicating that the effect is localized near object-generation decisions. These results provide pre-intervention evidence linking the dominant modes targeted by our edit to naturally occurring hallucination behavior.

### 7.4 Qualitative Analysis

Appendix Section E provides qualitative COCO examples comparing baseline and QK Product Steering captions side by side. These examples illustrate how suppressing the dominant top-3 spectral modes can remove hallucinated objects while preserving grounded content. Across the full COCO validation set, QK Product Steering with top-3 mode editing satisfies a strict improvement criterion on 756 images: it removes at least one hallucinated object, introduces no new hallucinations, and preserves all grounded object mentions. This provides additional evidence that the observed CHAIR reduction is not solely due to changes in caption length or indiscriminate object removal but reflects targeted hallucination mitigation.

## 8 Conclusion

We introduced *QK Product Steering*, a training-free and zero-inference-cost method for reducing object hallucination in VLMs through query-key product editing. By suppressing a few dominant middle-layer modes, it weakens prior-driven attention while keeping architecture and inference unchanged. Our results show that these modes are closely linked to hallucination, making QK Product Steering a practical alternative to decoding-time interventions.

## 9 Limitations

*QK Product Steering* is a lightweight post hoc edit focused on addressing object hallucination in free-form VLM generation. It weakens dominant QK attention modes but does not add visual knowledge or improve underlying perception, so failures due to poor recognition, ambiguous images, or insufficient visual evidence may persist. The method requires a small set of hyperparameters, including the edited layers, number of modes, and damping strength, whose optimal values may vary across architectures. Finally, extending the method to adaptive and position-aware steering is an important direction for future work.

## References

- Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, and 23 others. 2024. [Pixtral 12b](#). *Preprint*, arXiv:2410.07073.
- Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. [GQA: Training generalized multi-query transformer models from multi-head checkpoints](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4895–4901, Singapore. Association for Computational Linguistics.
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. [Flamingo: A visual language model for few-shot learning](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. [Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond](#). *Preprint*, arXiv:2308.12966.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. [Qwen2.5-VL technical report](#). *Preprint*, arXiv:2502.13923.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. [InstructBLIP: Towards general-purpose vision-language models with instruction tuning](#). In *Advances in Neural Information Processing Systems*, volume 36.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. 2023. [MME: A comprehensive evaluation benchmark for multimodal large language models](#). *Preprint*, arXiv:2306.13394.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. [HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385.
- Anisha Gunjal, Jihan Yin, and Erhan Bas. 2023. [Detecting and preventing hallucinations in large vision language models](#). *Preprint*, arXiv:2308.06394.
- 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13418–13427.
- 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25004–25014.
- 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13872–13882.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. [BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. [Evaluating object hallucination in large vision-language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.
- Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenxing Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. 2025. [The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering](#). *Preprint*, arXiv:2502.03628.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. [Microsoft COCO: Common objects in context](#). In *Computer Vision – ECCV 2014*, pages 740–755. Springer.
- Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. [Mitigating hallucination in large multi-modal models via robust instruction tuning](#). *Preprint*, arXiv:2306.14565.

- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. [Visual instruction tuning](#). In *Advances in Neural Information Processing Systems*, volume 36.
- Sheng Liu, Haotian Ye, and James Y. Zou. 2025. [Reducing hallucinations in large vision-language models via latent space steering](#). In *International Conference on Learning Representations*.
- Shi Liu, Kecheng Zheng, and Wei Chen. 2024. [Paying more attention to images: A training-free method for alleviating hallucination in LVLMs](#). In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- Kyungmin Min, Minbeom Kim, Kang-il Lee, Dongryeol Lee, and Kyomin Jung. 2025. [Mitigating hallucinations in large vision-language models via summary-guided decoding](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4183–4198, Albuquerque, New Mexico. Association for Computational Linguistics.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. [Object hallucination in image captioning](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.
- 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, Vienna, Austria. Association for Computational Linguistics.
- Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024. [Aligning large multimodal models with factually augmented RLHF](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110, Bangkok, Thailand. Association for Computational Linguistics.
- Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. 2023. [AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation](#). *Preprint*, arXiv:2311.07397.
- Yujun Wang, Aniri, Jinhe Bi, Soeren Pirk, and Yunpu Ma. 2025. [ASCD: Attention-steerable contrastive decoding for reducing hallucination in MLLM](#). *Preprint*, arXiv:2506.14766.
- Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Fangxun Shu, Hao Jiang, and Linchao Zhu. 2024. [Detecting and mitigating hallucination in large vision language models via fine-grained AI feedback](#). *Preprint*, arXiv:2404.14233.
- Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. 2025a. [Nullu: Mitigating object hallucinations in large vision-language models via HalluSpace projection](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14635–14645.
- Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. 2025b. [Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention](#). In *International Conference on Learning Representations*.
- Jianghao Yin, Qin Chen, Kedi Chen, Jie Zhou, Xingjiao Wu, and Liang He. 2026. [Dynamic multimodal activation steering for hallucination mitigation in large vision-language models](#). *Preprint*, arXiv:2602.21704. Accepted at ICLR 2026.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. [Woodpecker: Hallucination correction for multimodal large language models](#). *Preprint*, arXiv:2310.16045.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024a. [MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI](#). *Preprint*, arXiv:2311.16502.
- Zihao Yue, Liang Zhang, and Qin Jin. 2024b. [Less is more: Mitigating multimodal hallucination from an EOS decision perspective](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11766–11781, Bangkok, Thailand. Association for Computational Linguistics.
- Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2025. [LMMs-eval: Reality check on the evaluation of large multimodal models](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916, Albuquerque, New Mexico. Association for Computational Linguistics.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Yue Cao, Yangzhou Liu, Weiye Xu, Hao Li, Jiahao Wang, Han Lv, Dengnian Chen, and 28 others. 2025a. [InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models](#). *Preprint*, arXiv:2504.10479.
- Younan Zhu, Linwei Tao, Minjing Dong, and Chang Xu. 2025b. [Mitigating object hallucinations in large vision-language models via attention calibration](#). *Preprint*, arXiv:2502.01969.

## Appendix Contents

|          |                                                                     |           |
|----------|---------------------------------------------------------------------|-----------|
| <b>A</b> | <b>Notation and Conventions</b>                                     | <b>13</b> |
| <b>B</b> | <b>Theoretical Insights</b>                                         | <b>13</b> |
| B.1      | QK Product as the Attention-Logit Operator . . . . .                | 13        |
| B.2      | Why Top Singular Modes Are Natural Edit Targets . . . . .           | 14        |
| B.3      | Symmetric and Antisymmetric Attention Channels . . . . .            | 14        |
| B.4      | GQA-Safe Query Recovery as a Minimum-Change Edit . . . . .          | 15        |
| B.5      | Global SVD vs. Component-Specific Edits . . . . .                   | 16        |
| B.6      | Bounded Off-Target Effect . . . . .                                 | 16        |
| B.7      | Prediction Stability Under a Margin . . . . .                       | 17        |
| <b>C</b> | <b>Implementation Details</b>                                       | <b>18</b> |
| C.1      | Proposed Algorithm . . . . .                                        | 18        |
| C.2      | Training-Free and Zero-Cost Edit . . . . .                          | 18        |
| C.3      | Statistical Significance . . . . .                                  | 18        |
| C.4      | Model Architectures and Details . . . . .                           | 19        |
| C.5      | Hardware and Compute Cost . . . . .                                 | 19        |
| C.6      | Artifact Use . . . . .                                              | 19        |
| <b>D</b> | <b>Additional Experimental Results</b>                              | <b>19</b> |
| D.1      | Benchmarks and Evaluation Protocol . . . . .                        | 19        |
| D.2      | Controls and Ablations . . . . .                                    | 20        |
| D.3      | Sym/Antisym k-Sweep . . . . .                                       | 20        |
| D.4      | $\alpha$ -Sweep with Confidence Intervals . . . . .                 | 20        |
| D.5      | Joint $(k, \alpha)$ Grid . . . . .                                  | 21        |
| D.6      | Capability Tradeoff at Higher $(k, \alpha)$ . . . . .               | 21        |
| D.7      | Cross-Architecture k Selection . . . . .                            | 21        |
| D.8      | Per-Token and Pre-Noun Null Checks . . . . .                        | 21        |
| D.9      | Per-Layer Attention Trajectory . . . . .                            | 21        |
| D.10     | Comparison with Existing Hallucination Mitigation Methods . . . . . | 22        |
| D.11     | Pre-Intervention Spectral Analysis . . . . .                        | 23        |
| <b>E</b> | <b>Qualitative Examples with COCO Images</b>                        | <b>24</b> |
| <b>F</b> | <b>Limitations and Future Work</b>                                  | <b>24</b> |
| <b>G</b> | <b>Ethics Statement</b>                                             | <b>24</b> |
| <b>H</b> | <b>Use of Large Language Models (LLMs)</b>                          | <b>24</b> |

## A Notation and Conventions

Table 4 summarizes the main notation used throughout the paper. We use  $h$  to index query heads and  $g$  to index the corresponding shared key head under grouped-query attention. The central object in our method is the per-head QK product  $M_h = W_{q,h}^\top W_{k,g}$ , which directly determines the pre-softmax attention logits. QK Product Steering edits this product spectrally, maps the edited product back to the query weights, and keeps the shared key weights fixed.

| Notation                              | Description                                                      |
|---------------------------------------|------------------------------------------------------------------|
| $X \in \mathbb{R}^{n \times d}$       | Hidden states with sequence length $n$ and dimension $d$         |
| $r$                                   | Attention head dimension                                         |
| $d$                                   | Model hidden dimension                                           |
| $h$                                   | Query head index                                                 |
| $g$                                   | Shared key head index assigned to query head $h$                 |
| $W_{q,h} \in \mathbb{R}^{r \times d}$ | Query projection for head $h$                                    |
| $W_{k,g} \in \mathbb{R}^{r \times d}$ | Key projection for shared key head $g$                           |
| $Q_h = XW_{q,h}^\top$                 | Query representation for head $h$                                |
| $K_g = XW_{k,g}^\top$                 | Key representation for head $g$                                  |
| $M_h = W_{q,h}^\top W_{k,g}$          | Per-head QK product controlling attention logits                 |
| $S_h$                                 | Symmetric component of $M_h$ ; mutual attention channel          |
| $\Omega_h$                            | Antisymmetric component of $M_h$ ; directional attention channel |
| $U\Sigma V^\top$                      | Thin SVD of the QK product $M_h$                                 |
| $\sigma_i$                            | $i$ -th singular value of $M_h$                                  |
| $\mathcal{S}$                         | Set of selected modes to edit                                    |
| $k$                                   | Number of singular modes selected for suppression                |
| $\alpha$                              | Damping strength for selected modes                              |
| $M_h^*$                               | Edited QK product after spectral damping                         |
| $\Delta M_h = M_h^* - M_h$            | Desired change in the QK product                                 |
| $\Delta W_{q,h}$                      | Query-weight update used to realize $M_h^*$                      |
| $\lambda$                             | Ridge coefficient for stable query recovery                      |
| $W_{q,h}^*$                           | Edited query projection                                          |
| $W_{k,g}^*$                           | Key projection after editing; kept unchanged                     |

Table 4: Summary of notation used in QK Product Steering.

## B Theoretical Insights

This section provides theoretical support for QK Product Steering. Our goal is not to prove universal hallucination removal, but to justify why editing the query-key product is a direct, targeted, and localized intervention. We show that the QK product is the exact attention-logit operator, that top singular modes are principled edit targets, that symmetric and antisymmetric components form orthogonal attention channels, and that the query-only recovery is a minimum-change GQA-safe realization of the edited product.

### B.1 QK Product as the Attention-Logit Operator

For query head  $h$  and its assigned key head  $g$ , let

$$Q_h = XW_{q,h}^\top, \quad K_g = XW_{k,g}^\top, \quad (23)$$

where  $X \in \mathbb{R}^{n \times d}$  is the hidden-state matrix. The pre-softmax attention logits are

$$Q_h K_g^\top = XW_{q,h}^\top W_{k,g} X^\top = X M_h X^\top, \quad M_h = W_{q,h}^\top W_{k,g}. \quad (24)$$

Thus,  $M_h$  is the content-based bilinear operator controlling the query-key interaction for head  $h$ ; standard scaling and position-dependent transformations such as RoPE act around this product. Editing  $M_h$  therefore directly modifies the attention interaction, rather than intervening only at the decoding logits or output probabilities.

## B.2 Why Top Singular Modes Are Natural Edit Targets

Let the thin SVD of the QK product be

$$M_h = \sum_{i=1}^r \sigma_i u_i v_i^\top, \quad \sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r. \quad (25)$$

The top- $k$  component is

$$M_{h,k} = \sum_{i=1}^k \sigma_i u_i v_i^\top. \quad (26)$$

**Proposition B.1** (Dominant QK modes are maximum-energy directions). *The matrix  $M_{h,k}$  is the best rank- $k$  approximation of  $M_h$  in Frobenius norm:*

$$M_{h,k} = \arg \min_{\text{rank}(A) \leq k} \|M_h - A\|_F. \quad (27)$$

Moreover,

$$\|M_{h,k}\|_F^2 = \sum_{i=1}^k \sigma_i^2. \quad (28)$$

*Proof.* This follows directly from the Eckart–Young–Mirsky theorem. The truncated SVD gives the optimal rank- $k$  approximation in Frobenius norm, and the Frobenius energy of an SVD component is the sum of squared singular values. Since the singular values are sorted in decreasing order, the top- $k$  modes carry the largest rank- $k$  spectral energy.  $\square$

QK Product Steering suppresses this dominant component:

$$M_h^* = M_h - \alpha M_{h,k} = M_h - \alpha \sum_{i=1}^k \sigma_i u_i v_i^\top. \quad (29)$$

Hence, the edit is not an arbitrary perturbation: it targets the strongest low-rank directions of the actual attention-logit operator.

## B.3 Symmetric and Antisymmetric Attention Channels

Every QK product admits the unique decomposition

$$M_h = S_h + \Omega_h, \quad S_h = \frac{1}{2}(M_h + M_h^\top), \quad \Omega_h = \frac{1}{2}(M_h - M_h^\top), \quad (30)$$

where  $S_h^\top = S_h$  and  $\Omega_h^\top = -\Omega_h$ .

**Proposition B.2** (Orthogonal mutual and directional channels). *The symmetric and antisymmetric components are orthogonal under the Frobenius inner product:*

$$\langle S_h, \Omega_h \rangle_F = \text{tr}(S_h^\top \Omega_h) = 0. \quad (31)$$

Consequently,

$$\|M_h\|_F^2 = \|S_h\|_F^2 + \|\Omega_h\|_F^2. \quad (32)$$

*Proof.* Since  $S_h^\top = S_h$  and  $\Omega_h^\top = -\Omega_h$ ,

$$\text{tr}(S_h^\top \Omega_h) = \text{tr}(S_h \Omega_h). \quad (33)$$

Using trace invariance under transpose,

$$\text{tr}(S_h \Omega_h) = \text{tr}((S_h \Omega_h)^\top) = \text{tr}(\Omega_h^\top S_h^\top) = \text{tr}(-\Omega_h S_h) = -\text{tr}(S_h \Omega_h). \quad (34)$$

Therefore,  $\text{tr}(S_h \Omega_h) = 0$ . The additive energy decomposition follows from orthogonality:

$$\|M_h\|_F^2 = \|S_h + \Omega_h\|_F^2 = \|S_h\|_F^2 + \|\Omega_h\|_F^2 + 2\langle S_h, \Omega_h \rangle_F. \quad (35)$$

□

For token representations  $x_i$  and  $x_j$ , the attention-logit contribution decomposes as

$$x_i^\top M_h x_j = x_i^\top S_h x_j + x_i^\top \Omega_h x_j. \quad (36)$$

The symmetric term is reciprocal:

$$x_i^\top S_h x_j = x_j^\top S_h x_i, \quad (37)$$

whereas the antisymmetric term changes sign under reversal:

$$x_i^\top \Omega_h x_j = -x_j^\top \Omega_h x_i. \quad (38)$$

Thus,  $S_h$  captures mutual content-similarity patterns, while  $\Omega_h$  captures directional attention patterns. This justifies Sym-only and Antisym-only edits as mechanistic probes of whether hallucination is associated with reciprocal content priors or directional attention effects.

#### B.4 GQA-Safe Query Recovery as a Minimum-Change Edit

After obtaining an edited product  $M_h^*$ , we must realize it in the model weights. Under grouped-query attention, the key projection  $W_{k,g}$  is shared across multiple query heads, so modifying it can unintentionally affect other heads. We therefore keep  $W_{k,g}$  fixed and update only  $W_{q,h}$ .

Let

$$\Delta M_h = M_h^* - M_h. \quad (39)$$

We seek the smallest query-weight update that realizes the edited product:

$$\begin{aligned} \min_{\Delta W_{q,h}} \quad & \|\Delta W_{q,h}\|_F^2 \\ \text{s.t.} \quad & (W_{q,h} + \Delta W_{q,h})^\top W_{k,g} = M_h^*. \end{aligned} \quad (40)$$

Equivalently,

$$\Delta W_{q,h}^\top W_{k,g} = \Delta M_h. \quad (41)$$

Taking transposes gives

$$W_{k,g}^\top \Delta W_{q,h} = \Delta M_h^\top. \quad (42)$$

**Proposition B.3** (Minimum-change query recovery). *If  $W_{k,g}$  has full row rank, the minimum-Frobenius-norm solution of Eq. (40) is*

$$\Delta W_{q,h} = (W_{k,g} W_{k,g}^\top)^{-1} W_{k,g} \Delta M_h^\top. \quad (43)$$

*Proof.* The transposed constraint has the form

$$A \Delta W_{q,h} = B, \quad A = W_{k,g}^\top, \quad B = \Delta M_h^\top. \quad (44)$$

The minimum-norm solution of a linear system is given by the Moore–Penrose pseudoinverse:

$$\Delta W_{q,h} = A^\dagger B. \quad (45)$$

Since  $A = W_{k,g}^\top$  and  $W_{k,g}$  has full row rank,

$$A^\dagger = (A^\top A)^{-1} A^\top = (W_{k,g} W_{k,g}^\top)^{-1} W_{k,g}. \quad (46)$$

Substituting gives the stated update. The ridge version replaces the inverse with a numerically stable regularized inverse.  $\square$

The final recovered weights are

$$W_{q,h}^* = W_{q,h} + \Delta W_{q,h}, \quad W_{k,g}^* = W_{k,g}. \quad (47)$$

Thus, the edited QK product is realized through a query-only update while leaving shared key weights unchanged, making the edit GQA-safe.

*Remark B.4.* In practice, we use the ridge-stabilized least-squares update

$$\Delta W_{q,h} = (W_{k,g} W_{k,g}^\top + \lambda I)^{-1} W_{k,g} \Delta M_h^\top. \quad (48)$$

where

$$\lambda = \epsilon \cdot \frac{1}{r} \text{tr}(W_{k,g} W_{k,g}^\top). \quad (49)$$

For small  $\lambda$ , this approximates the minimum-norm recovery while improving numerical stability.

*Remark B.5.* For edits constructed from the QK product SVD,  $\Delta M_h$  lies in the appropriate row/column subspace, so the unregularized solution can realize the product edit exactly under the full-row-rank condition. For component-specific symmetric or antisymmetric edits, the ridge update should be interpreted as the closest query-only realization.

## B.5 Global SVD vs. Component-Specific Edits

The global SVD of  $M_h$  does not, in general, preserve the symmetric and antisymmetric decomposition. Since

$$M_h = S_h + \Omega_h, \quad (50)$$

the singular vectors of  $M_h$  can mix energy from both channels. Therefore, damping the top singular modes of  $M_h$  acts as a broad spectral intervention on the full attention-logit operator.

In contrast, Sym-only and Antisym-only edits operate on the two orthogonal channels separately. Sym-only edits the eigenmodes of  $S_h$ , isolating reciprocal content-similarity patterns. Antisym-only edits the singular modes of  $\Omega_h$ , isolating directional attention patterns. These component-specific edits are therefore diagnostic: they test which attention channel carries hallucination-relevant signal.

## B.6 Bounded Off-Target Effect

We next show why QK Product Steering is expected to have limited off-target effects. For each selected head,

$$\Delta M_h = M_h^* - M_h = -\alpha \sum_{i=1}^k \sigma_i u_i v_i^\top. \quad (51)$$

Thus,

$$\text{rank}(\Delta M_h) \leq k, \quad \|\Delta M_h\|_F = |\alpha| \left( \sum_{i=1}^k \sigma_i^2 \right)^{1/2}. \quad (52)$$

For hidden states  $X$ , the induced attention-logit perturbation is

$$\Delta L_h = X \Delta M_h X^\top. \quad (53)$$

If  $\|X\|_2 \leq B_X$ , then

$$\|\Delta L_h\|_F \leq \|X\|_2^2 \|\Delta M_h\|_F \leq B_X^2 |\alpha| \left( \sum_{i=1}^k \sigma_i^2 \right)^{1/2}. \quad (54)$$

Therefore, the change in attention logits is controlled by the energy of the suppressed low-rank component.

If the downstream computation from the edited attention logits to the final output logits is  $L_f$ -Lipschitz, then

$$\|\Delta z\|_2 \leq L_f B_X^2 |\alpha| \left( \sum_{i=1}^k \sigma_i^2 \right)^{1/2}. \quad (55)$$

This does not guarantee identical predictions on every input. Rather, it shows that the edit is localized and bounded: only a small set of attention directions in selected middle layers is modified. This provides a mathematical explanation for why broad capability benchmarks can remain close to the baseline while hallucination-sensitive generation changes more noticeably.

### B.7 Prediction Stability Under a Margin

For classification or multiple-choice VLM tasks, the bounded-output result gives a simple stability condition. Let  $z$  be the original output logits and let  $y^* = \arg \max_i z_i$ . Define the prediction margin

$$\gamma = z_{y^*} - \max_{j \neq y^*} z_j. \quad (56)$$

Suppose the edit changes every output logit by at most  $\eta$ :

$$\|\Delta z\|_\infty \leq \eta. \quad (57)$$

If

$$2\eta < \gamma, \quad (58)$$

then the predicted answer is unchanged.

Indeed, the original top logit can decrease by at most  $\eta$ , while any competing logit can increase by at most  $\eta$ . Hence the margin can shrink by at most  $2\eta$ . If  $2\eta < \gamma$ , the original top class remains the top class. This explains why single-decision benchmarks can remain stable when examples have sufficiently large margins, while long-form captioning can still be affected through many small generation-step changes.

---

**Algorithm 1** QK Product Steering

---

**Require:** Pretrained VLM, selected layers  $\mathcal{L}$ , modes  $k$ , damping  $\alpha$ , ridge scale  $\epsilon$

**Ensure:** Edited VLM

- 1: **for** each layer  $\ell \in \mathcal{L}$  **do**
- 2:   **for** each query head  $h$  **do**
- 3:     Find the shared key head  $g$  assigned to  $h$ .
- 4:     Extract  $W_{q,h}$  and  $W_{k,g}$ .
- 5:     Compute thin SVD of the QK product:

$$M_h = W_{q,h}^\top W_{k,g} = U \Sigma V^\top.$$

- 6:     Dampen top- $k$  singular values:

$$\Sigma_{ii}^* = \begin{cases} (1 - \alpha)\Sigma_{ii}, & i \leq k, \\ \Sigma_{ii}, & i > k. \end{cases}$$

- 7:     Form edited product:

$$M_h^* = U \Sigma^* V^\top, \quad \Delta M_h = M_h^* - M_h.$$

- 8:     Set

$$\lambda = \epsilon \cdot \frac{1}{r} \text{tr}(W_{k,g} W_{k,g}^\top).$$

- 9:     Compute query-only update:

$$\Delta W_{q,h} = (W_{k,g} W_{k,g}^\top + \lambda I)^{-1} W_{k,g} \Delta M_h^\top.$$

- 10:     Update query weight:

$$W_{q,h} \leftarrow W_{q,h} + \Delta W_{q,h}.$$

- 11:     Keep key weight fixed:

$$W_{k,g} \leftarrow W_{k,g}.$$

- 12:   **end for**

- 13: **end for**

- 14: **return** edited VLM
- 

## C Implementation Details

We provide implementation details to make the proposed edit reproducible. QK Product Steering is applied as a one-time offline modification to the pretrained model weights. For each selected layer and query head, we identify the corresponding shared key head under grouped-query attention, compute the per-head QK product through the efficient thin-SVD procedure, damp the selected spectral modes, and recover the edited query weight using the closed-form query-only update. All computations are performed directly on the attention projection weights; no training data, gradient updates, or loss optimization are used. After the edited weights are written back to the model, inference uses the original architecture, tokenizer, image processor, and decoding configuration unchanged.

### C.1 Proposed Algorithm

Algorithm 1 summarizes the full QK Product Steering procedure. The method is applied once to a pretrained VLM over a selected set of layers and query heads. For each query head, we first identify its corresponding shared key head under grouped-query attention and form the per-head QK product. We then compute its spectral decomposition, suppress the top- $k$  singular modes, and obtain an edited product. Finally, the edited product is mapped back to the model through a closed-form query-only update, while the shared key weight is kept fixed. This makes the edit GQA-safe and preserves the original inference procedure. Since all steps are computed directly from pretrained weights, the algorithm requires no training data, gradient updates, or decoding-time modification.

### C.2 Training-Free and Zero-Cost Edit

QK Product Steering is computed directly from pretrained attention weights. It does not use training data, optimize a loss, or backpropagate through the model. Once the edited query weights are written back, the architecture and decoding procedure remain unchanged. Therefore, the edited model has the same inference cost as the original model.

### C.3 Statistical Significance

We assess statistical significance using paired bootstrap resampling over image IDs. For CHAIR evaluation, each edited model is paired with the corresponding unedited baseline on the same COCO 2014 val images and we compute the distribution of  $\Delta\text{CHAIR}_s$  and  $\Delta\text{CHAIR}_i$  over paired-bootstrap samples. We use  $n_{\text{boot}}=200$  resamples for the full-split  $N=4,946$  runs (where each bootstrap pass is expensive) and  $n_{\text{boot}}=1,000$  for the  $N=1,000$  Qwen  $k$ - and  $\alpha$ -sweeps used in the ablations. We report 95% percentile confidence intervals and the bootstrap-estimated  $P(\Delta > 0)$ , treating an edit as statistically reliable only when its CI excludes zero. The same paired setup is used for all CHAIR controls, including bottom- $k$ , random- $k$ , symmetric-only, antisymmetric-only, and amplification.

Table 5 reports paired-bootstrap CIs at the preferred middle-band setting ( $k=3, \alpha=1$ ) on Qwen2.5-VL-7B (Sym top  $k_s=3$ ,  $N=1,000$ ) and Pixtral-12B (QK Product Steering top- $k$ ,  $N=4,946$ ). On Qwen, the symmetric-channel edit excludes zero on both  $\Delta\text{CHAIR}_s$  and  $\Delta\text{CHAIR}_i$ , while the matched bottom- $k$  control on the same

channel does not – the signal lives in the top eigenmodes of  $S_h$ , not in arbitrary low-rank perturbations of  $\mathbf{W}_q$ . On Pixtral, QK Product Steering top- $k$  excludes zero on  $\Delta\text{CHAIR}_i$  with  $P(\Delta > 0) = 0.000$ , whereas its matched random- $k$  control does not. Together these CIs back the body’s main hallucination claims at the cleanest available sample size per model.

For probe and capability benchmarks (POPE, HallusionBench, MME, MMMU) we report raw within-pipeline differences and interpret them relative to the matched random-mode control, since these evaluations are used to verify capability preservation rather than to establish the main hallucination claim.

#### C.4 Model Architectures and Details

**Per-model architectures.** The three GQA VLMs used in the main paper have the following backbones. Qwen/Qwen2.5-VL-7B-Instruct: 28 transformer blocks, 28 query heads, 4 key/value heads (group size 7), head dimension 128. OpenGVLab/InternVL3-8B: 28 transformer blocks, 28 query heads, 7 key/value heads (group size 4), head dimension 128. mistral-community/pixtral-12b (the HF-native repackage of Mistral’s Pixtral-12B): 40 transformer blocks, 32 query heads, 8 key/value heads (group size 4), head dimension 128. All three are loaded in bfloat16 with `attn_implementation=eager`; the eager attention path is what permits the per-(layer, head) query weight rewrites used by QK Product Steering.

#### C.5 Hardware and Compute Cost

All experiments – the apply-edit step, the CHAIR / POPE / HallusionBench / MME / MMMU evaluations, the teacher-forced attention measurements, and the per-head spectral decomposition diagnostics – run on a single NVIDIA A100 (80GB) GPU. Each job loads one model into memory once, applies the edit in place, and then runs the benchmark serially with greedy decoding; no multi-GPU sharding or quantization is used. The apply-edit step itself completes in  $\sim 6$ – $10$  seconds for all three models and contributes negligibly to wall-clock time relative to generation. Approximate per-job wall-clock times: CHAIR on the full COCO 2014 val split ( $N=4,946$ ) takes  $\sim 1.5$ – $2.5$  hours on Qwen2.5-VL-7B and InternVL3-8B and  $\sim 11$  hours on Pixtral-12B (longer due to model size); POPE ( $N=9,000$ ) and MME ( $N\approx 2,374$ ) each take  $\sim 20$ – $35$  minutes per variant; HallusionBench ( $N=447$ ) and MMMU

val ( $N=900$ ) each take  $\sim 5$ – $15$  minutes per variant; the per-head spectral decomposition is GPU-only weight math and finishes in under a minute per model.

#### C.6 Artifact Use

We use existing models, datasets, and benchmarks consistently with their intended research use. The evaluated VLM checkpoints and benchmarks, including COCO, CHAIR, POPE, HallusionBench, MME, and MMMU, are used only for research evaluation of hallucination mitigation and multimodal capability. We do not introduce a new dataset or pretrained model artifact. Any released code is intended for research use, and users should comply with the licenses and access conditions of the underlying models and datasets.

### D Additional Experimental Results

This section provides implementation and evaluation details that support the main experimental claims. We first describe the per-model architectures and the hardware used to run all experiments. We then describe the benchmark protocols, sample sizes, decoding settings, and scoring procedures used for hallucination, probe, and capability evaluation. We then define the control edits used to test whether the observed gains come from dominant QK modes rather than arbitrary spectral perturbations. Finally, we report additional null checks and attention-trajectory analyses, which show that the effect is not explained by local token suppression or immediate pre-noun re-grounding, but instead emerges through downstream attention changes after the edited middle layers.

#### D.1 Benchmarks and Evaluation Protocol

**Per-benchmark protocols and sample sizes.** Generative hallucination is evaluated with CHAIR on the full COCO 2014 val split ( $N = 4,946$ ) under the standard descriptive prompt with greedy decoding and no image-resolution cap; we report both per-sentence  $\text{CHAIR}_s$  and per-mention  $\text{CHAIR}_i$  and pair every edit against the unedited baseline on the same image IDs for paired-bootstrap CIs (resample counts in Appendix C.3). Probe hallucination uses POPE on  $N = 9,000$  random / popular / adversarial yes/no items and HallusionBench on its  $N = 447$  figure-paired items (covering 1,129 underlying question instances aggregated to the paired figure level under the official scoring protocol), both scored with a regex extractor matching the official

| Variant                            | $\Delta_s$ | 95% CI <sub>s</sub>   | $\Delta_i$ | 95% CI <sub>i</sub> |
|------------------------------------|------------|-----------------------|------------|---------------------|
| <i>Qwen2.5-VL-7B</i> ( $N=1,000$ ) |            |                       |            |                     |
| Sym top $k_s=3$                    | -2.17      | <b>[-2.95, -1.33]</b> | -0.67      | [-1.44, +0.03]      |
| Antisym top $k_a=6$                | -0.12      | [-0.92, +0.62]        | +0.31      | [-0.44, +1.10]      |
| Sym bot $k_s=3$ (ctrl)             | +0.14      | [-0.56, +0.80]        | -0.52      | [-1.18, +0.13]      |
| Antisym bot $k_a=6$ (ctrl)         | +0.15      | [-0.57, +0.86]        | -0.58      | [-1.28, +0.09]      |
| <i>Pixtral-12B</i> ( $N=4,946$ )   |            |                       |            |                     |
| QK Product Steering top $k=3$      | -0.36      | [-0.85, +0.17]        | -0.83      | [-1.26, -0.33]      |
| Random $k=3$ (ctrl)                | -0.17      | [-0.56, +0.21]        | -0.35      | [-0.74, +0.13]      |
| Sym top $k_s=3$                    | -0.06      | [-0.41, +0.32]        | -0.17      | [-0.56, +0.21]      |
| Antisym top $k_a=6$                | -0.10      | [-0.50, +0.32]        | -0.37      | [-0.73, +0.04]      |

Table 5: Paired-bootstrap 95% CIs on  $\Delta\text{CHAIR}_s$  ( $\Delta_s$ ) and  $\Delta\text{CHAIR}_i$  ( $\Delta_i$ ) at the preferred middle-band ( $k=3$ ,  $\alpha=1$ ) edit,  $n_{\text{boot}}=200$ . Bolded CIs (above) exclude zero. The Qwen Sym top- $k_s=3$  row defends the main-paper claim that the symmetric-channel  $\text{CHAIR}_s$  drop on Qwen is CI-significant; the Pixtral QK Product Steering top- $k$  row defends the corresponding claim on Pixtral  $\Delta\text{CHAIR}_i$ . Matched bottom- $k$  and random- $k$  controls in both blocks have CIs that include zero, ruling out arbitrary low-rank perturbations of  $\mathbf{W}_q$ .

protocol. General capability uses MME (perception + cognition total) and MMMU val under the standard lmms-eval v0.7 multiple-choice protocol. POPE, MME, and MMMU all run through the mmmu\_val\_qwen, pope, and mme tasks of the same harness with each model’s reference inference config; we exclude any task that requires a paid LLM judge to keep evaluation reproducible.

## D.2 Controls and Ablations

At the default hyperparameters we additionally report a bottom- $k$  control selecting the smallest-magnitude modes, a random- $k$  control selecting modes uniformly at random, and (for QK Product Steering only) an amplify control at  $\alpha < 0$  that doubles rather than removes the selected modes. Definitions of the three edit families (QK Product Steering, Sym-only, Antisym-only, Sym+Antisym) are in Section 4, Table 1.

## D.3 Sym/Antisym k-Sweep

Table 6 reports the sym/antisym top- $k$  and bottom- $k$  sweep on Qwen2.5-VL-7B at fixed  $\alpha=1$  on  $N=1,000$  COCO images. The symmetric top- $k$  edit peaks at  $k_s=3$  on  $\Delta\text{CHAIR}_s$  (CI excludes zero) and does not improve at  $k_s=5$ . The antisymmetric channel is non-monotone and weak across  $k_a \in \{2, 6, 12\}$  – no setting recovers the symmetric top-3 effect. Bottom- $k$  at matched rank is null on both channels (CIs include zero), supporting the conclusion that the hallucination signal lives in the top eigenmodes of  $S_h$  rather than in arbitrary low-rank perturbations of  $\mathbf{W}_q$  or in the antisymmetric channel.

| Variant        | $k$      | $\Delta\text{CHAIR}_s$ | 95% CI                |
|----------------|----------|------------------------|-----------------------|
| Sym top        | 1        | -2.38                  | –                     |
| <b>Sym top</b> | <b>3</b> | <b>-2.17</b>           | <b>[-2.95, -1.33]</b> |
| Sym top        | 5        | -2.93                  | –                     |
| Sym bot        | 3        | +0.14                  | [-0.56, +0.80]        |
| Antisym top    | 2        | -1.41                  | –                     |
| Antisym top    | 6        | -0.12                  | [-0.92, +0.62]        |
| Antisym top    | 12       | -1.67                  | –                     |
| Antisym bot    | 6        | +0.15                  | [-0.57, +0.86]        |

Table 6: Sym/Antisym top- $k$  and bottom- $k$  sweep on Qwen2.5-VL-7B at  $\alpha=1$ , middle-band edit,  $N=1,000$  (baseline  $\text{CHAIR}_s = 32.75$ ). Bootstrap CIs ( $n_{\text{boot}}=1,000$ ) are computed for the four CI-reported rows; rows without CIs are point estimates from chair\_metrics.json.

## D.4 $\alpha$ -Sweep with Confidence Intervals

Table 7 reports the Sym top- $k_s=3$   $\alpha$ -sweep on Qwen2.5-VL-7B,  $N=1,000$ , with paired-bootstrap CIs at each  $\alpha$ . The  $\Delta\text{CHAIR}_s$  response is monotonically more negative in  $\alpha$  on the zero-out range  $[0, 1]$ , with CIs excluding zero from  $\alpha=0.5$  onward. The  $\alpha>1$  rows continue to lower  $\text{CHAIR}_s$  – the selected modes are sign-flipped rather than removed – but as Section D.6 shows, this CHAIR gain comes with a measurable capability cost.

| $\alpha$ | $\Delta\text{CHAIR}_s$ | 95% CI         |
|----------|------------------------|----------------|
| 0.3      | -0.56                  | [-1.34, +0.15] |
| 0.5      | -1.02                  | [-1.74, -0.21] |
| 0.7      | -1.39                  | [-2.20, -0.61] |
| 1.0      | -2.17                  | [-2.95, -1.33] |

Table 7:  $\alpha$ -sweep for Sym top- $k_s=3$  on Qwen2.5-VL-7B, middle-band edit,  $N=1,000$ ,  $n_{\text{boot}}=1,000$ .  $\Delta\text{CHAIR}_s$  becomes more negative monotonically in  $\alpha$ ; CIs exclude zero from  $\alpha=0.5$  onward.

## D.5 Joint $(k, \alpha)$ Grid

We additionally run a 122-cell joint  $(k, \alpha)$  grid on Qwen2.5-VL-7B at  $N=500$  across all three edit families: QK Product Steering top- $k$ , Sym top- $k_s$ , and Antisym top- $k_a$ . CHAIR<sub>s</sub> at fixed  $\alpha \geq 1$  decreases monotonically with  $k$  for PSVD and Sym; Antisym remains weak at every  $(k, \alpha)$  cell. The per-method grid minima are reported in Table 8. The default cell ( $k=3, \alpha=1$ ) is not the global CHAIR<sub>s</sub> minimum: the PSVD grid minimum sits at ( $k=6, \alpha=2$ ) at CHAIR<sub>s</sub> = 24.78 versus the baseline 31.44. However, the larger- $k$  and larger- $\alpha$  cells carry a capability cost that the default cell does not, as quantified in Section D.6.

| Method             | best $k^*$ | best $\alpha^*$ | CHAIR <sub>s</sub> |
|--------------------|------------|-----------------|--------------------|
| Baseline           | –          | –               | 31.44              |
| PSVD top- $k$      | 6          | 2.0             | <b>24.78</b>       |
| Sym top- $k_s$     | 4          | 2.0             | 27.28              |
| Antisym top- $k_a$ | 10         | 2.0             | 28.16              |

Table 8: Per-method grid minima from the 122-cell joint  $(k, \alpha)$  sweep on Qwen2.5-VL-7B,  $N=500$ . Antisym’s minimum sits well above PSVD’s and Sym’s at every  $(k, \alpha)$  cell.

## D.6 Capability Tradeoff at Higher $(k, \alpha)$

To isolate the capability cost of each axis, we evaluate POPE / MME / MMMU at three Qwen cells: the default ( $k=3, \alpha=1$ ); the matched- $k$  stronger- $\alpha$  cell ( $k=3, \alpha=2$ ); and the matched- $\alpha$  higher- $k$  cell ( $k=6, \alpha=1$ ). Table 9 shows the result. Both axes independently cost capability, with  $\alpha$  the larger lever: MMMU loses 7.11pp at  $\alpha=2$  versus 3.78pp at  $k=6$ , and MME loses  $\sim 4\times$  more at  $\alpha=2$  than at  $k=6$ . POPE F1 is essentially unchanged in all three cells. The default cell ( $k=3, \alpha=1$ ) is therefore the smallest principled cell where the CHAIR drop is CI-significant and capability is preserved within the random-control noise budget.

| Setting                           | MMMU  | MME    | POPE F1 |
|-----------------------------------|-------|--------|---------|
| Baseline                          | 51.56 | 2339.2 | 86.18   |
| $(k=3, \alpha=1)$ default         | 50.89 | 2344.4 | 86.68   |
| $(k=6, \alpha=1) - k$ effect      | 47.11 | 2321.0 | 87.07   |
| $(k=3, \alpha=2) - \alpha$ effect | 43.78 | 2253.5 | 85.74   |

Table 9: Capability tradeoff at higher  $(k, \alpha)$  on Qwen2.5-VL-7B. MMMU and MME degrade monotonically with both  $k$  and  $\alpha$ ;  $\alpha$  is the larger lever. POPE F1 is essentially unchanged across the three cells.

## D.7 Cross-Architecture $k$ Selection

The full  $(k, \alpha)$  ablation in Sections D.3–D.6 is conducted on Qwen2.5-VL-7B and drives the ( $k=3, \alpha=1$ ) default on that model. For the cross-architecture rows in Table 2, we keep  $\alpha=1$  and the middle-third edit band fixed and select  $k$  per model as the lower-CHAIR<sub>s</sub> cell at the same protocol:  $k=1$  for InternVL3-8B and  $k=3$  for Pixtral-12B. The InternVL3 choice is consistent with its higher mid-band  $E_3$  profile (Figure 2), where a smaller  $k$  keeps the edit surgical.

## D.8 Per-Token and Pre-Noun Null Checks

Two natural null hypotheses for the CHAIR<sub>s</sub> reduction would be that QK Product Steering selectively suppresses the per-token log-probability of hallucinated noun tokens at fixed prefix, or that it visibly re-grounds attention at the position immediately before a hallucinated noun. Under teacher-forced prefixes neither holds. QK Product Steering top-3 leaves the hallucinated-token log-probability essentially unchanged ( $-0.004$ ) while reducing the grounded-token log-probability *more* ( $-0.049$ ); the attention-to-image probability at the pre-noun position is within 0.002 of every control (Table 10). The mechanism is therefore neither token-local nor position-local; the rest of the discussion in the main text locates it.

| Variant                   | $\Delta \log P$ |        | attn→img |       |
|---------------------------|-----------------|--------|----------|-------|
|                           | hallu           | gr.    | hallu    | gr.   |
| Matched-norm rand.        | -0.015          | +0.006 | 0.126    | 0.131 |
| Random- $k$               | +0.003          | +0.013 | 0.126    | 0.131 |
| QK-Prod top-3, $\alpha=1$ | -0.004          | -0.049 | 0.125    | 0.129 |

Table 10: Per-token log-probability deltas and attention-to-image probability at the pre-noun position, split by hallucinated vs. grounded noun tokens.  $N=291$  hallucinated noun tokens and  $N=214$  grounded noun tokens across 100 COCO images, middle layers.

## D.9 Per-Layer Attention Trajectory

Figure 5 further supports the trajectory-cascade explanation in Section 7.2. The edited layers show nearly zero layer-mean shift due to per-head cancellation, while the effect emerges in the downstream unedited layers, with the strongest increase immediately after the edit site.

Figure 5 disaggregates the layer-mean  $\Delta v$  inside the edited band over its 252 (layer, head) cells. The per-head shifts are large in magnitude ( $\pm 0.48$ pp)

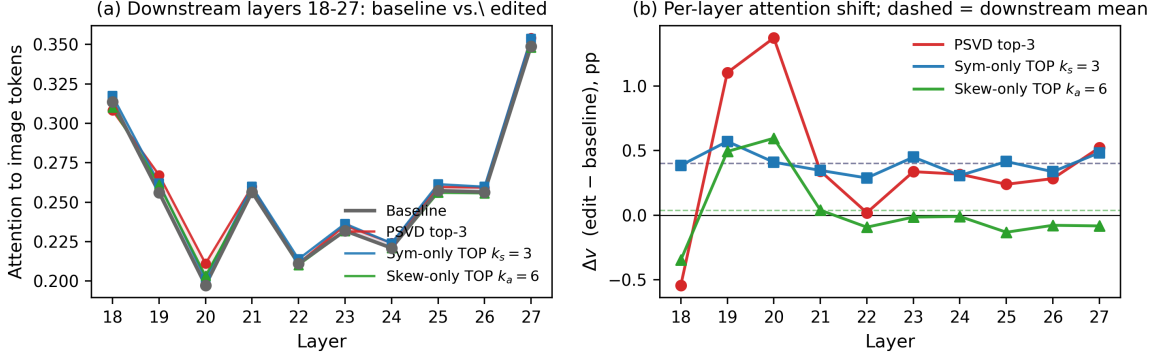


Figure 4: Per-layer shift  $\Delta v$  (edit – baseline, percentage points) across all 28 layers of Qwen2.5-VL-7B for the three spectral methods, teacher-forced. Shaded band: edited middle layers  $L_9-L_{17}$ . The two CHAIR-active methods (QK Product Steering, Sym-only) produce a uniform positive shift across every downstream unedited layer; Antisym-only, which is inert on CHAIR, does not.

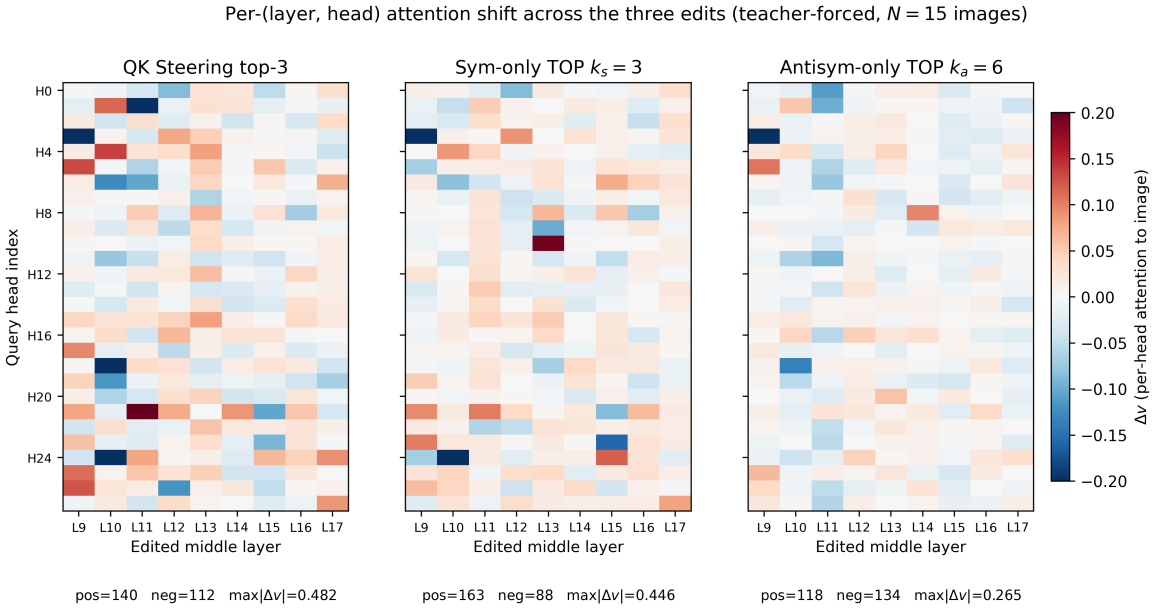


Figure 5: Per-(layer, head) attention-to-image shift  $\Delta v$  inside the edited middle band ( $9 \times 28 = 252$  cells), QK Product Steering top-3 on Qwen2.5-VL-7B. Per-head shifts reach  $\pm 0.48$ pp with 138 positive and 114 negative cells; the layer mean is  $+0.019$ pp.

and roughly balanced in sign (138 positive, 114 negative), so they cancel in the layer mean ( $+0.019$ pp). They do not cancel in the residual stream that those heads jointly write, which is what the downstream layers in Figure 4 respond to with the uniform positive shift.

#### D.10 Comparison with Existing Hallucination Mitigation Methods

We compare QK Product Steering with representative training-free hallucination-mitigation methods, including VCD (Leng et al., 2024), VTI (Liu et al., 2025), and VAR (Jiang et al., 2025). The results are presented in Table 11.

QK Product Steering is competitive with existing

hallucination-mitigation approaches while offering a distinct advantage. It performs a one-time, data-free, training-free weight-space edit that leaves the inference procedure unchanged and introduces no decoding-time overhead. Furthermore, operating directly on the QK product provides a mechanistic perspective on the spectral modes associated with hallucination-related behavior.

For VAR, we report results on Qwen2.5-VL-7B, where the method can be directly evaluated under its standard multi-head attention formulation. Its performance is lower in this setting, while direct application to the GQA-based InternVL3-8B and Pixtral-12B is not reliable. Overall, these results demonstrate that QK Product Steering provides

| Model         | Method                     | CHAIR <sub>s</sub> ↓ | CHAIR <sub>i</sub> ↓ |
|---------------|----------------------------|----------------------|----------------------|
| Qwen2.5-VL-7B | Baseline                   | 31.44                | 65.26                |
|               | <b>QK Product Steering</b> | <b>28.56</b>         | <b>63.75</b>         |
|               | VCD                        | 31.19                | 66.55                |
|               | VTI                        | 31.00                | 64.14                |
|               | VAR                        | 34.26                | 65.86                |
| InternVL3-8B  | Baseline                   | 33.81                | 62.98                |
|               | <b>QK Product Steering</b> | <b>31.23</b>         | <b>62.07</b>         |
|               | VCD                        | 35.94                | 65.46                |
|               | VTI                        | 35.57                | 65.70                |
| Pixtral-12B   | Baseline                   | 34.14                | 64.86                |
|               | <b>QK Product Steering</b> | <b>33.81</b>         | <b>63.04</b>         |
|               | VCD                        | 34.68                | 66.03                |
|               | VTI                        | 37.14                | 66.05                |

Table 11: Comparison with representative training-free hallucination-mitigation methods. Lower CHAIR scores indicate fewer object hallucinations.

a complementary approach to decoding-time and activation-level interventions for hallucination mitigation.

### D.11 Pre-Intervention Spectral Analysis

To examine whether the spectral modes targeted by QK Product Steering are associated with naturally occurring hallucination behavior, we perform a pre-intervention analysis on the unedited Qwen2.5-VL-7B model using 500 baseline captions generated on the COCO validation set. Each generated caption is fed back through the same unedited model using teacher forcing, with no weight editing or intervention applied at any stage.

Since each attention logit can be decomposed into contributions from the singular modes of the QK product, using the same thin SVD employed by our method, we quantify the contribution of a subset of modes to attention logits through its *logit energy*, defined as the mean squared contribution toward a specified token set. At the token position immediately preceding each mentioned object, we measure two statistics for each middle-layer head: (i) *top-3 image-energy share*, the fraction of image-directed logit energy contributed by the top-3 singular modes, and (ii) *top-3 image-text balance*, the fraction of the top-3 mode energy directed toward image tokens rather than text tokens. The statistics are averaged across the middle-layer heads.

We compare naturally occurring hallucinated object mentions with grounded object mentions within the same caption using a paired Wilcoxon signed-rank test. In total, 252 captions contain both hallucinated and grounded mentions. The bottom-3 singular modes from the same heads are used as a

control.

The results show that, immediately before hallucinated object mentions, the dominant QK modes contribute a smaller share of image-directed logit energy and exhibit a shift in energy from image tokens toward text context. In contrast, the bottom-3 control modes do not exhibit the same deficit, with their image-text balance changing in the opposite direction. This provides pre-intervention evidence that the spectral modes targeted by our method are associated with naturally occurring hallucination behavior.

We further evaluate whether these statistics can predict hallucination risk. At the mention level, the top-mode image-energy share provides a modest separation between hallucinated and grounded mentions, achieving an AUC of 0.58 (1,584 hallucinated vs. 1,093 grounded mentions; Mann-Whitney  $p \approx 5 \times 10^{-13}$ ). At the whole-caption level, averaging over all tokens gives an AUC of approximately 0.48, indicating that the signal is highly localized around object-mention positions. Restricting the analysis to mention positions yields an AUC of 0.57 ( $p = 0.06$ ). These results suggest that the spectral signal is primarily localized near the object-generation decision rather than uniformly distributed across the caption.

Overall, this analysis provides evidence that the same dominant spectral modes targeted by QK Product Steering are associated with naturally occurring hallucination-related behavior before intervention. We interpret this as supporting mechanistic evidence rather than proof that these modes universally encode generic vision-language priors.

| Statistic                   | Mean Diff. | % Hallu. Higher | Wilcoxon                      |
|-----------------------------|------------|-----------------|-------------------------------|
| Top-3 image-energy share    | -0.0091    | 27.8%           | $p \approx 10^{-16}$          |
| Bottom-3 image-energy share | +0.00003   | 51.2%           | $p > 0.1$                     |
| Top-3 image-text balance    | -0.0050    | 32.5%           | $p \approx 3 \times 10^{-10}$ |
| Bottom-3 image-text balance | +0.0083    | 69.0%           | $p \approx 5 \times 10^{-12}$ |

Table 12: Pre-intervention comparison between hallucinated and grounded object mentions. Mean differences are computed as hallucinated minus grounded.

## E Qualitative Examples with COCO Images

This section provides qualitative examples illustrating how QK Product Steering changes generated captions on COCO images. We show cases where the edited model removes hallucinated object mentions from the baseline caption while preserving grounded visual content. These examples are intended to complement the quantitative CHAIR results by showing that the reduction in hallucination corresponds to more image-grounded descriptions rather than shorter or less informative captions.

Across the full COCO validation set ( $N=4,946$ ), we identify 756 images where QK Product Steering top-3 satisfies a strict improvement criterion: it removes at least one hallucinated COCO object from the baseline caption, introduces no new hallucinated objects, and preserves all grounded object mentions. Figure H presents eight representative examples, ranked by the number of hallucinations removed. For each case, we show the COCO image together with truncated baseline and QK Product Steering top-3 captions.

## F Limitations and Future Work

QK Product Steering is a lightweight post-hoc edit with an intentionally focused scope. It targets the content-based query-key product and keeps the model architecture and inference procedure unchanged. This design makes the method efficient, but it does not explicitly model position-dependent attention effects, such as those introduced by rotary positional embeddings. Extending QK Product Steering to position-aware query-key interactions is therefore an important direction for future work.

The method also uses a small set of hyperparameters, including the edited layers, heads, number of modes, and damping strength. While our experiments show that middle-layer dominant modes are effective across multiple VLMs, the optimal configuration may vary across architectures. Future work can make these choices adaptive using

spectral statistics or a small calibration set.

Finally, QK Product Steering is designed to reduce object hallucination in free-form generation by weakening misleading attention priors. It does not add new visual knowledge or solve all forms of hallucination, such as failures caused by weak visual perception, ambiguous images, or adversarial reasoning. This makes it complementary to training-based alignment, decoding-time verification, uncertainty calibration, and external grounding methods. Combining these directions could lead to more robust hallucination-mitigation pipelines for VLMs.

## G Ethics Statement

QK Product Steering is a lightweight post-hoc intervention for reducing object hallucination in VLM-generated captions. It requires no additional data, annotations, gradient updates, or decoding-time changes, making it easy to apply to existing models. However, QK Product Steering is not a universal hallucination solution. It mainly targets free-form object hallucination by weakening dominant prior-driven attention modes, and may not address failures from weak visual perception, ambiguous images, or adversarial yes/no reasoning. Therefore, in safety-critical settings, QK Product Steering should be used as a complementary reliability tool alongside stronger grounding, verification, and human oversight.

## H Use of Large Language Models (LLMs)

Large language models (LLMs) were used only as writing aids during manuscript preparation. Their use was limited to improving grammar, clarity, wording, and organization. The authors reviewed and revised all LLM-assisted text to ensure that the final manuscript accurately represents the intended meaning, technical claims, and scientific contributions.

## Color Reference

**Red** = hallucinated by the baseline, **green** = grounded, **blue** = grounded object the baseline missed and the edit recovers.

**Example 1 (image 120853): plated sandwich with coleslaw.** REMOVED: *fork*; KEPT: *bowl, carrot, cell phone, dining table, sandwich.*



**Baseline.** This image shows a meal served on a green plate at an outdoor dining setting, likely a cafe or restaurant. The main dish is a **sandwich** made with toasted bread, which appears to be filled with grilled chicken and possibly some vegetables. The **sandwich** is cut diagonally into two halves, with one half placed on top of the other. Accompanying the **sandwich** is a small white **bowl** containing coleslaw, which includes shredded cabbage, **carrots**, and possibly other vegetables, dressed with a light sauce. A small white plastic **fork** is placed inside the **bowl** for serving. In the background, there is a glass of red liquid, possibly iced tea or juice, and a **smartphone** resting on the **table**. A salt shaker is also visible on the left side of the frame. The **table** has a metal mesh design, and the overall ambiance suggests a casual dining experience.

**QK Product Steering.** This image shows a meal served on a **table**, likely at an outdoor dining area given the metal mesh **table** and the presence of sunlight. The main focus is a plate with a **sandwich** cut in half, revealing its contents. The **sandwich** appears to be made with toasted bread, possibly sourdough or a similar type, and contains slices of grilled chicken as the filling. Next to the **sandwich**, there is a small **bowl** filled with coleslaw, which includes shredded cabbage, **carrots**, and other vegetables, dressed in a creamy dressing. The coleslaw adds a fresh and colorful element to the meal. In the background, there's a glass of what looks like iced tea, a **smartphone** resting on the **table**, and a pepper shaker. The overall setting suggests a casual dining experience, perhaps at a café or restaurant. The lighting indicates it might be daytime, and the scene has a relaxed atmosphere.

**Example 2 (image 150417): birthday cake in a home kitchen.** REMOVED: *toaster*; KEPT: *bowl, cake, cup, dining table, person, refrigerator.*



**Baseline.** The image shows an indoor setting, likely a kitchen or dining area, where two individuals are present. The focus is on the older **woman** seated at a **table**, wearing glasses and a red jacket over a blue top. She appears to be looking at a small **cake** placed in front of her on the **table**. The **cake** has multiple lit candles, suggesting it might be for a birthday celebration. The **table** is cluttered with various items, including a pink **cup**, a clear **bowl**, some papers, and what looks like a decorative rabbit figurine. A **man**'s arm is visible on the right side of the frame, indicating he is standing next to the **table**. He seems to be interacting with the **cake** or the **woman**, possibly lighting another candle or assisting with the celebration. In the background, there is a **refrigerator** adorned with magnets, notes, and pictures, as well as a countertop with various kitchen items such as **containers**, utensils, and a **toaster**. The overall atmosphere suggests a casual and warm family gathering.

**QK Product Steering.** The image shows an indoor setting, likely a kitchen or dining area, where two individuals are present. The focus is on an elderly **woman** seated at a **table**, wearing glasses and a red cardigan over a blue top. She appears to be looking at the **cake** in front of her with a thoughtful expression. The **cake** is round, white, and decorated with colorful candles, suggesting it might be for a birthday celebration. A hand, presumably belonging to another **person**, is seen lighting one of the candles. The individual lighting the candle is partially visible, with only their arm and part of their torso shown. The **table** has various items on it, including a pink **cup**, a clear **bowl**, and some papers. In the background, there's a **refrigerator** adorned with magnets and notes, along with other kitchen appliances and cabinets. The room is lit by warm artificial light, giving the scene a cozy, homey feel. The overall atmosphere suggests a casual and intimate gathering, possibly a family celebration.

**Example 3 (image 237517): patio dinner on a wooden table.** REMOVED: *bench*; RECOVERED: *spoon*; KEPT: *bowl, dining table, fork, knife, potted plant*.



**Baseline.** This image depicts an outdoor dining setup on a wooden **table**, likely in a backyard or garden area. The **table** is set with various dishes and utensils, suggesting a meal is either being enjoyed or prepared. 1. Foreground: - A white plate holds a piece of steak, roasted potatoes, and grilled zucchini slices. - A **knife** and **fork** rest on the plate, indicating that someone is eating or has been eating. - To the right of the plate, there's a small **bowl** containing what appears to be a salad with greens and possibly some crumbled cheese or dressing. 2. Midground: - Another plate further back contains a baked dish, possibly a quiche or pie, served with a side of bread or toast. - Next to it, there's a small **bowl** filled with fresh greens, possibly a salad or herbs. 3. Background: - The setting includes a white wooden **bench** and a **potted plant** with succulent-like leaves on the left side of the **table**. - Beyond the **table**, there is a brick patio leading to a grassy area with a hedge in the distance, suggesting a private garden space. 4. Atmosphere: - The scene conveys a relaxed, casual dining experience, likely during the daytime given the natural light. - The presence of plants and the outdoor setting adds a tranquil and cozy ambiance to the meal. The overall composition suggests a simple yet enjoyable meal in a peaceful outdoor environment.

**QK Product Steering.** This image depicts an outdoor dining setup on a wooden **table**, likely in a backyard or patio area. The **table** is set with various dishes and utensils, suggesting a meal is being enjoyed. 1. Foreground Plate: - A white plate holds a piece of steak, a baked potato, and two slices of grilled zucchini. - A **knife** and **fork** rest on the plate, indicating that the meal is ready to be eaten. 2. Background Plate: - Another plate contains what appears to be a baked dish, possibly a quiche or pie, served in a small bread **bowl**. - A **spoon** is placed inside the dish, and it seems to have a creamy filling. 3. **Bowl** of Salad: - Two **bowls** are visible on the right side of the **table**. One contains a green salad with leafy greens and possibly some croutons or cheese. - The other **bowl** has a similar salad but with a different mix of greens and toppings. 4. Decorative Elements: - To the left of the **table**, there is a **potted plant** with succulent-like leaves, adding a touch of greenery to the scene. - A small pot with pink flowers is also visible, enhancing the aesthetic appeal of the setting. 5. Outdoor Setting: - The **table** is surrounded by a white wooden railing, which suggests the dining area is part of a deck or patio. - Beyond the railing, there is a grassy area and a brick pathway leading into the garden. - The background includes a hedge and some trees, indicating a private, well-maintained outdoor space. The overall ambiance of the image is cozy and inviting, with a focus on fresh food and a pleasant outdoor environment.

**Example 4 (image 384661): stove with a utensil cart.** REMOVED: *scissors*; RECOVERED: *refrigerator*; KEPT: *bottle, oven*.

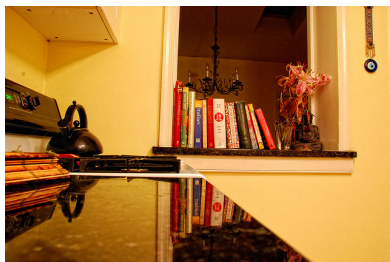


**Baseline.** The image shows a kitchen setup with a stainless steel gas **stove** and an adjacent metal shelving unit. The **stove** has four burners on top, a control panel with digital buttons, and a storage compartment below for warming or storing items. The **oven** door is closed, and the brand name "Bake Warm Store" is visible at the bottom of the **oven**. To the right of the **stove**, there is a metal shelving unit with three levels. On the top shelf, there is a large silver pot, a smaller pot, and some kitchen utensils such as spatulas and a pair of **scissors**. There is also a **bottle** of oil or sauce near the pots. The middle shelf holds a small red appliance (possibly a blender or food processor) and a white kettle. The bottom shelf appears to be empty. The background wall is painted yellow, and the floor is tiled with terracotta-colored tiles. The overall setting suggests a compact kitchen space with efficient use of vertical space.

**QK Product Steering.** The image shows a kitchen setup with a stainless steel gas **stove** on the left and a small rolling cart to its right. Here's a detailed description:

- Stove:** - The **stove** is a gas model with four burners on top. - It has a digital control panel at the front, which appears to be displaying settings or times. - Below the burners is an **oven** with a black glass door, indicating it is a combination **oven** (with both a conventional **oven** and a broiler). - The **stove** has a sleek, modern design with a brushed metal finish.
- Rolling Cart:** - The cart is made of metal and has three shelves. - On the top shelf: - There is a large silver pot, likely used for cooking or boiling water. - Next to the pot, there is a **bottle** that could contain oil or another cooking ingredient. - On the middle shelf: - A small red appliance, possibly a blender or food processor, is visible. - A white kettle or teapot is also present. - On the bottom shelf: - A round object, possibly a lid or a lid cover, is placed on the shelf. - A white corded device, potentially a small appliance like a food processor or mixer, is partially visible.
- Background:** - The wall behind the **stove** is painted yellow, adding warmth to the scene. - To the right of the cart, part of a **refrigerator** is visible, suggesting this is a compact kitchen space.
- Flooring:** - The floor appears to be tiled, with a terracotta or similar color, complementing the warm tones of the room. This setup suggests a functional and organized kitchen area, with appliances and tools neatly arranged for easy access.

**Example 5 (image 571893): kitchen pass-through with cookbooks.** REMOVED: *potted plant*; RECOVERED: *vase*; KEPT: *book, oven*.



**Baseline.** This image shows a cozy and well-organized corner of a room, likely a kitchen or study area. The focal point is a built-in shelf with a dark marble-like finish, which holds an assortment of items. On the left side of the shelf, there's a black kettle on what appears to be a **stovetop**, suggesting that this is part of a kitchen setup. On the right side of the shelf, there is a collection of **books** neatly arranged vertically. The spines of the **books** display various colors and designs, indicating a diverse range of topics or genres. Next to the **books**, there is a small **potted plant** with vibrant pink flowers, adding a touch of nature and color to the scene. Above the shelf, a chandelier hangs from the ceiling, providing warm lighting for the space. To the far right, there is a decorative item resembling an evil eye amulet, often used as a talisman for protection. The walls are painted a soft yellow, complementing the overall warm and inviting atmosphere of the room. The reflection on the shiny surface below the shelf adds depth to the image, creating a mirrored effect that enhances the visual appeal.

**QK Product Steering.** This image shows a cozy and well-decorated corner of a room, likely a kitchen or dining area. Here's a detailed description: 1. Foreground (Mirror Reflection): - The left side of the image features a mirror reflecting part of the scene. - A black kettle is visible on what appears to be a **stovetop** or countertop. - A small black object, possibly a kitchen appliance or utensil, is placed near the kettle. 2. Background (Window Shelf): - A window with a white frame is positioned above the mirror reflection. - On the windowsill, there is an organized collection of **books** arranged vertically. The spines of the **books** are visible, suggesting they might be cookbooks or reference **books** due to their size and color. - To the right of the **books**, there is a decorative **vase** holding pink flowers, adding a touch of color and freshness to the space. - A small chandelier hangs from the ceiling, providing soft lighting for the area below. - A hanging ornament resembling an evil eye amulet is attached to the wall to the right of the window, adding a cultural or decorative element. 3. Wall and Color Scheme: - The walls are painted a warm yellow tone, which gives the space a welcoming and inviting feel. - The window frame and trim are white, creating a contrast against the yellow wall. 4. Lighting: - The lighting in the room is soft and warm, likely coming from the chandelier and possibly additional light sources not visible in the image. Overall, the image conveys a sense of warmth, functionality, and personalization, with a mix of practical elements like the kettle and **books**, and decorative touches like the flowers and evil eye ornament.

**Example 6 (image 58350): two people on a couch eating pizza.** REMOVED: *chair*; RECOVERED: *couch*; KEPT: *person, pizza*.



**Baseline.** The image shows two women sitting indoors, likely in a casual setting such as a living room or a similar space. The **woman** in the foreground is seated on a patterned **chair** with a red and beige design. She has blonde hair styled up and is wearing a black top with a light-colored strap visible underneath. She appears to be eating a slice of **pizza**, which she holds close to her mouth. Her expression seems neutral or slightly focused. The **woman** in the background is wearing a green top and is partially out of focus. She is holding what looks like a yellow object, possibly a utensil or part of a game, and is looking toward the camera with a slight smile. The background is plain and light-colored, suggesting an indoor environment with minimal decoration. The overall atmosphere of the photo feels informal and relaxed.

**QK Product Steering.** The image shows two women sitting on a **couch** in what appears to be an indoor setting, possibly a living room or a casual gathering space. The **woman** in the foreground is holding a large slice of **pizza** with a visible topping that looks like cheese and possibly tomato sauce. She has blonde hair styled up and is wearing a black top with a light-colored strap visible underneath. Her expression seems neutral as she holds the **pizza**. The **woman** in the background is wearing a green top and is partially out of focus. She appears to be holding a yellow object, possibly a utensil or tool, and her expression suggests she might be engaged in conversation or reacting to something happening off-camera. The **couch** they are sitting on has a patterned design with a mix of red and beige tones. The overall atmosphere of the image feels casual and informal.

**Example 7 (image 470924): outdoor pizza lunch.** REMOVED: *umbrella*; RECOVERED: *fork, knife*; KEPT: *car, chair, dining table, person, pizza*.



**Baseline.** The image shows a group of four **people** sitting at an outdoor **dining table**, enjoying a meal together. The setting appears to be a casual restaurant or café with wooden **tables** and **chairs**. The **table** is set with plates of **pizza**, glasses of beer, and cutlery. The **pizzas** look freshly made, with visible toppings such as cheese, herbs, and possibly vegetables. The individuals are dressed casually, with one **person** wearing a red jacket, another in a black hoodie, and the others in darker clothing. They seem relaxed and engaged in conversation, with some looking directly at the camera while others appear to be mid-conversation. The background reveals other diners seated at similar **tables**, and the area is shaded by large **umbrellas**, suggesting it's a sunny day. Beyond the **umbrellas**, there is a street view with parked **cars** and a sidewalk, indicating that the location is likely in an urban or suburban area. The atmosphere seems lively yet laid-back, typical of a social gathering over food.

**QK Product Steering.** This image shows four individuals seated at an outdoor **dining table**, enjoying a meal together. The setting appears to be a casual restaurant or café with other patrons and **tables** visible in the background. The **table** is set with plates of **pizza**, glasses of beer, and cutlery. The individuals seem relaxed and engaged in conversation.

- Foreground: - On the left, there's a **person** wearing a red jacket and sunglasses, holding a **fork** and **knife**, seemingly about to eat.
- In the center, another individual is wearing a black hoodie and sunglasses, looking directly at the camera with a slight smile.
- To the right, a **person** in a dark hoodie is also looking at the camera, appearing to be in mid-conversation.

- Background: - Other diners can be seen in the background, some seated at separate **tables**, while others are standing or walking around. - The outdoor seating area has wooden **tables** and **chairs**, and the ambiance suggests a sunny day with natural light illuminating the scene.

- Food and Drinks: - The **table** is laden with several **pizzas**, each with different toppings that appear to include cheese, vegetables, and possibly meat. - There are multiple glasses of beer on the **table**, indicating a casual dining experience.

- Environment: - The location seems to be a street-side café or restaurant with a view of a road and parked **cars** in the distance. - The weather looks pleasant, as suggested by the bright lighting and shadows cast by the **people** and objects. The overall mood of the image is friendly and convivial, capturing a moment of shared enjoyment among friends or acquaintances.

**Example 8 (image 340175): living and dining room with a bookshelf.** RECOVERED: *bench, potted plant*; KEPT: *book, chair, couch, dining table*.



**Baseline.** This image depicts a cozy and well-lit living room with a mix of modern and rustic elements. The room features a large window that allows natural light to flood the space, offering a view of trees outside. The walls are painted a neutral gray tone, which complements the wooden flooring. In the center of the room is a wooden **dining table** with a simple design, accompanied by red **chairs** with a modern aesthetic. The **table** appears to be made from reclaimed wood, adding a touch of rustic charm. A small decorative item, possibly a candle or ornament, sits on the **table** near the window. To the right of the dining area is a dark brown leather **sofa**, positioned against the wall. In front of the **sofa** lies a cowhide rug, adding texture and warmth to the space. A matching ottoman sits next to the **sofa**, providing additional seating or a place to rest one's feet. On the left side of the room, there is a tall bookshelf filled with **books**, suggesting an interest in reading or collecting literature. The bookshelf has a warm, wooden finish that matches the flooring. Above it, a decorative chandelier with a circular, flower-like design hangs from the ceiling, adding a unique and artistic touch to the room. The overall atmosphere of the room is inviting and comfortable, blending functionality with style. The combination of natural materials like wood and leather, along with the modern furniture and decor, creates a harmonious and stylish living space.

**QK Product Steering.** This image depicts a cozy and well-lit living room with a mix of modern and rustic elements. Here's a detailed description: 1. Furniture and Layout: - The room features a wooden parquet floor, which adds warmth to the space. - A dark red leather **sofa** is positioned against the wall on the right side, accompanied by a small white ottoman and a minimalist **chair** with a curved metal frame. - On the left side, there is a wooden **dining table** with a simple design, surrounded by bright red **chairs** that contrast with the neutral tones of the room. 2. Lighting and Decor: - Two large, decorative chandeliers hang from the ceiling, adding an artistic touch to the room. - Natural light streams in through a large window in the center of the back wall, which offers a view of trees outside, suggesting a peaceful setting. - The window is framed by a radiator below it, and the wall above the radiator has a few framed pictures or artwork. 3. Decorative Elements: - A bookshelf filled with **books** occupies the far left corner, indicating a love for reading or a scholarly interest. - A cowhide rug lies on the floor near the **sofa**, adding texture and a natural element to the room. - A small **potted plant** sits on the windowsill, bringing a touch of greenery indoors. - A colorful abstract painting hangs on the wall to the left of the window, adding vibrancy to the otherwise neutral palette. 4. Other Features: - A wooden **bench** with a simple design is placed near the **dining table**, offering additional seating. - The walls are painted in a muted gray tone, which complements the warm wood tones and red accents. - The overall aesthetic combines functionality with a touch of artistic flair, creating a comfortable and inviting atmosphere.