UNCOVERING SEPARABLE AND COMPOSABLE SKILL MECHANISMS IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work has sought to localize the internal mechanisms underlying visual reasoning skills in vision-language models. We posit two important desiderata, separability and composability, for skill localization. Separability requires identified mechanisms to contribute primarily to their target skill while having limited effects on other skills, enabling precise diagnosis and targeted intervention. Composability requires skill mechanisms to be combinable to support complex tasks, allowing a small set of localized skill mechanisms to account for broader capabilities. However, existing methods overlook separability and composability in both mechanism identification and evaluation. To address these limitations, we propose a multi-skill interpretability framework **DIS**entangled and **CO**mposable Skill Mechanisms (DISCO) for identifying and evaluating separable and composable attention heads. For mechanism identification, we propose a novel technique called Separable Head Selection, which selects skill-specific heads based on their contributions across skills. For evaluation, we are the first to design metrics for evaluating separability and composability of identified heads. Experiments on three VLMs and four fundamental visual reasoning skills show that the heads identified by DISCO achieve the highest overall separability on single-skill tasks, outperforming the strongest baseline by 22.75, 10.95, and 9.96 points across the three models, while also achieving the highest average composability on multi-skill tasks across all models.

1 INTRODUCTION

The identification (Tiong et al., 2024; Kang et al., 2026), enhancement (Chen et al., 2023; Zhu et al., 2025), and mechanistic interpretation (Shi et al., 2025; Schaumlöffel et al., 2026; Che et al., 2026; Chen et al., 2026) of *skills* in Large Language Models (LLMs) and Vision-Language Models (VLMs) have emerged as an important research area. A skill can be defined as a unit of computational capability that serves a single purpose and can be combined to form more complex capabilities. For example, determining the number of apples in an image requires the skill of counting objects. When combined with the positioning skill, which determines the relative positions between objects, they allow the model to answer the question "how many apples are on the left of the box in the image?" With the understanding of skills, we can efficiently select training data (Chen et al., 2023; Zhao et al., 2025), steer skill-related internal activations (Kang et al., 2026; Che et al., 2026), and conduct fine-grained evaluation and diagnosis (Xu et al., 2024; Zhang et al., 2025).

Recent work has sought to localize the network components that implement skills, such as attention heads (Jiang et al., 2025; Schaumlöffel et al., 2026), layers (Shi et al., 2025), and circuits (Che et al., 2026). We posit that two requirements, separability and composability, are critical to skill localization. **Separability** requires that the effects of identified components be isolated to the target skill, with minimal impact on other skills; achieving separability would enable precise diagnosis of model errors and targeted skill enhancement with minimal spillover. **Composability** requires that different skills can be combined to support complex tasks, allowing a small set of skills to parsimoniously account for a broad range of capabilities. However, existing methods do not consider the effects of mechanisms on skills other than the target skill or the joint effects of multiple skill mechanisms, overlooking separability and composability in both mechanism identification and evaluation.

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

To address these limitations, we propose a multi-skill interpretability framework, **DIS**entangled and **CO**mposable Skill Mechanisms (DISCO), for identifying and evaluating separable and composable attention heads. We study four fundamental visual reasoning skills: counting, positioning, size comparison, and distance comparison, using corresponding VQA tasks. We propose a new mechanism identification technique called Separable Head Selection, which selects skill-specific heads based on their contributions across skills. We also design novel evaluation metrics that assess the separability and composability of the identified heads.

Separable Head Selection assigns an attention head to Skill A only if its contribution to Skill A exceeds its contributions to any other skill by a separation threshold. We estimate the contributions of heads through *Patching* and *Erasure*, which disrupt head activations in different directions and measure the resulting changes in model outputs. Instead of tuning the separation threshold manually, we find its value by maximizing a separability metric defined on the set of heads selected. This separability is measured as the contribution of these heads to Skill A minus their maximum contribution to any other skill. In plain English, we aim to select a group of heads whose contribution to Skill A exceeds their contribution to any other skill the most.

We are the first to propose metrics for evaluating the separability and composability of identified heads. Separability follows the definition above. Composability is defined using tasks requiring two skills jointly. For a task requiring Skills A and B, we separately disrupt the activations of the heads for Skill A, the heads for Skill B, and their union, and measure the resulting effects on the task. We define composability as the effect of disrupting the heads for Skills A and B together minus the effect of disrupting the heads for either skill alone. Disrupting the heads for one skill alone may have limited effects rather than causing the task to fail completely. For example, some heads contribute to Skill A but primarily contribute to other skills. They will not be selected for Skill A, but may still serve as weak substitutes, partially supporting the skill when its identified heads are disrupted.

To provide reliable contribution scores of each head to each skill, we introduce Surface-Form-Agnostic Head Scoring. Using skill data with a single surface form, such as a fixed answer format or visual style, may induce spurious correlations between the skill and the surface form. To mitigate this issue, we use data with multiple surface forms to estimate head contributions and average them across surface forms. Experiments on three VLMs show that the heads identified by our method achieve the highest overall separability on single-skill tasks, outperforming the strongest baseline by 22.75, 10.95, and 9.96 points on Qwen2.5-VL, Qwen3-VL, and Gemma-4, respectively, while also achieving the highest average composability on multi-skill tasks for all three models.

We summarize our main contributions as follows:

• We propose a multi-skill interpretability framework, DISCO, to identify and evaluate separable and composable attention heads in VLMs. The framework is formulated around two desiderata: **separability**, requiring identified heads to primarily contribute to their target skill while having limited contributions to other skills, and **composability**, requiring identified heads for skills to be combinable to support complex tasks. We are also the first to propose metrics for evaluating both the separability and composability of identified heads.

• We propose two modules for identifying skill-specific heads. **Separable Head Selection** optimizes a separation threshold to select the set of heads whose contribution to the target skill most exceeds their contribution to any other skill. To provide reliable contribution scores of each head to each skill, we introduce **Surface-Form-Agnostic Head Scoring**, which averages head contributions across multiple surface forms to mitigate spurious correlations between skills and surface forms.

2 Related Work

2.1 Skills in Vision-Language Models

Existing studies have investigated skills in VLMs by conducting skill-oriented benchmarking (Johnson et al., 2017; Hudson & Manning, 2019; Zhao et al., 2022; Yuksekgonul et al., 2022; Liu et al., 2023; Li et al., 2025; 2026), identifying the skills underlying different tasks (Tiong et al., 2024), and localizing skill mechanisms (Shi et al., 2025; Che et al., 2026; Schaumlöffel et al., 2026). Existing localization work typically identifies skill-related components based only on how strongly they contribute to the target skill, whether focusing on a single skill such as counting (Che et al.,

Table 1: Comparison of representative VLM mechanistic interpretation methods. Targeting Skills indicates whether the method localizes skill mechanisms. Separability-Aware Selection indicates whether the method considers component contributions across studied capabilities during selection.

Method	Granularity	Patching Source	Targeting Skills	Separability-Aware Selection	Composability Evaluation	Surface-Form Agnostic
NOTICE	Head	Counterfactual	✗	✗	✗	✗
V-SEAM	Head	Layer Average	✗	✗	✗	✗
VFL	Layer	Counterfactual	✓	✗	✗	✗
Ours	Head	Counterfactual	✓	✓	✓	✓

2026) or object localization (Schaumlöffel et al., 2026), or on multiple skills as in VFL (Shi et al., 2025). Such localization does not consider the contribution of a component to non-target skills and therefore does not ensure the separability of identified components. Moreover, prior work has not examined whether components identified for individual skills can compose to support complex tasks. In this work, we identify and evaluate separable and composable attention heads for skills. We further design novel metrics to evaluate the separability and composability of identified heads.

2.2 Mechanistic Interpretation in Vision-Language Models

Mechanistic interpretation of VLMs seeks to localize model components underlying specific capabilities (Jiang et al., 2025; Che et al., 2026). Activation patching (Vig et al., 2020; Meng et al., 2022; Wang et al., 2022; Conmy et al., 2023; Zhang & Nanda, 2024; Heimersheim & Nanda, 2024) is widely used for this purpose by replacing the activation of a component and measuring the resulting change in model output. NOTICE replaces activation of each attention head with its counterpart from a counterfactual input to identify modality-related heads (Golovanevsky et al., 2025); the same strategy has also been used to identify heads associated with object localization (Schaumlöffel et al., 2026) and counting (Che et al., 2026). VFL applies this strategy at the layer level to localize diverse vision-language capabilities (Shi et al., 2025). In contrast, V-SEAM uses the mean activation of the other heads in the same layer for patching to identify heads associated with different semantic levels (Wang et al., 2025). Beyond activation patching, some methods also identify heads through visual-token attention patterns (Bi et al., 2025; Kang et al., 2025; Ji et al., 2026) and head-wise probing (Jiang et al., 2025). Despite these differences, current methods typically rely on data with a single surface form, although VLM behavior is sensitive to surface-form variations (Ismithdeen et al., 2025; Atabuzzaman et al., 2025; Rosenthal et al., 2025). Consequently, the identified heads may capture surface-form-specific behavior rather than the underlying capability. In contrast, our method incorporates multiple surface forms to identify surface-form-agnostic skill mechanisms. Tab. 1 compares our method with representative mechanistic interpretation methods.

3 Framework

Our framework, DISCO, first identifies skill-specific attention heads and then evaluates their separability and composability. As illustrated in Fig. 1, the identification stage consists of two modules: (1) **Surface-Form-Agnostic Head Scoring** quantifies the contribution of each head to each skill across multiple surface forms, and (2) **Separable Head Selection** uses a skill-adaptive separation threshold to select heads for each skill based on their contribution scores. We then design novel separability and composability metrics for evaluating identified heads.

3.1 Data and Notation

Our data are built upon the VQA dataset SPHERE (Zhang et al., 2025). The single-skill data cover Counting, Position, Size, and Distance, each requiring one skill, while the multi-skill data cover Position + Size, Counting + Position, and Distance + Size, each requiring two skills jointly. For single-skill data, we construct counterfactual image pairs and multiple surface forms, as illustrated in Fig. 1. The surface forms include two answer formats (Option Content/Option Label) and three visual styles (Natural/Cartoon/Oil Painting). We then split the data into disjoint identification and

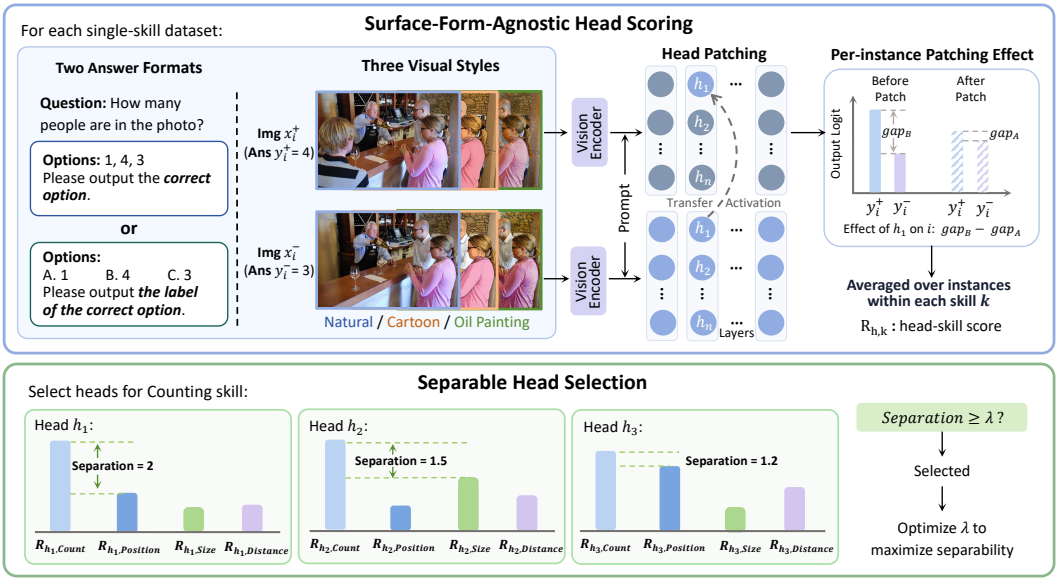

Figure 1: Overview of our head identification pipeline. Surface-Form-Agnostic Head Scoring measures the contribution score of each head to each skill across multiple surface forms. Separable Head Selection then optimizes a separation threshold for each skill to select skill-specific heads.

evaluation sets. Multi-skill data are used only for evaluation and contain neither counterfactual pairs nor surface-form variants. More details are provided in App. A.

We use H_k to denote the attention heads identified for skill k. Each single-skill instance i is instantiated with one answer format and one visual style, and consists of a question prompt q_i and a counterfactual image pair (x_i^+, x_i^-) with corresponding correct answers (y_i^+, y_i^-). The prompt q_i contains the question, answer options, and response instruction.

3.2 Surface-Form-Agnostic Head Scoring

To obtain the contribution scores of each head to each skill, we first measure the intervention effect of each head on individual instances through counterfactual activation patching, and then average the effects over all instances of a skill. Since these instances span multiple surface forms, the resulting scores favor heads that consistently contribute to the skill across surface forms. The procedure of head scoring is summarized in Alg. 1 in App. B.

Position-aware counterfactual activation patching. To distinguish head contributions to context processing and next-token prediction, we perform activation patching separately over the preceding token positions and the final token position. This is inspired by prior mechanistic work showing that attention heads can perform complementary computations involving different token positions to support a function (Olsson et al., 2022). For a multimodal input sequence of T tokens, we consider two token-position sets: the preceding positions $\mathcal{T}_{\text{preceding}} = \{1, \ldots, T-1\}$ and the final position $\mathcal{T}_{\text{final}} = \{T\}$. We use $p \in \{\text{preceding}, \text{final}\}$ to index these position sets.

For each skill k, we first retain the instances (q_i, x_i^+, x_i^-) for which the VLM correctly predicts y_i^+ from (q_i, x_i^+), and denote the resulting set by $\mathcal{I}_k$, so that intervention effects are measured from initially correct predictions. For each instance $i \in \mathcal{I}_k$ and attention head h, we run the model on (q_i, x_i^+) and (q_i, x_i^-) separately and patch the head outputs from the x_i^- run into the x_i^+ run over token positions $\mathcal{T}_p$. We focus on the first generated token, as it is sufficient for distinguishing the candidate answers under both the Option Content and Option Label formats. Let ℓ_y denote the output logit for answer token y. We define $M_i = \ell_{y_i^-} - \ell_{y_i^+}$ as the original logit margin of the counterfactual answer over the original answer, and let $M_{i,h}^p$ denote the margin after patching head

h over $\mathcal{T}_p$. We quantify the effect of head h as the change in this margin:

$$\Delta^p_{i,h} = M^p_{i,h} - M_i. \tag{1}$$

A positive $\Delta^p_{i,h}$ indicates that patching head h shifts the prediction from the original answer y_i^+ toward the counterfactual answer y_i^-. We treat it as evidence that head h contributes to the skill k on its instance i over token positions $\mathcal{T}_p$. A larger $\Delta^p_{i,h}$ indicates a stronger contribution. Finally, we average the effects across instances in $\mathcal{I}_k$ to obtain the head contribution score:

$$S^p_{h,k} = \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \Delta^p_{i,h}, \qquad p \in \{\text{preceding}, \text{final}\}. \tag{2}$$

Coverage weighting and within-skill normalization. A large contribution score can be dominated by a few instances on which a head exhibits unusually strong effects, even if it has little or opposite effect on most other instances. Such outlier-driven scores may overestimate the contribution of heads. To mitigate this issue, we define the *coverage* of each head as the fraction of instances for which patching the head produces a positive effect, and use it to weight the score:

$$C^p_{h,k} = \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \mathbb{I}\left[\Delta^p_{i,h} > 0\right], \qquad \widetilde{S}^p_{h,k} = S^p_{h,k} C^p_{h,k}. \tag{3}$$

The resulting coverage-weighted score $\widetilde{S}^p_{h,k}$ better reflects the contribution of head h to skill k by accounting for both effect strength and consistency across instances.

To enable cross-skill score comparison, we further normalize the scores over all heads for each skill. Specifically, we compute $R^p_{h,k} = (\widetilde{S}^p_{h,k} - \mu^p_k)/(\sigma^p_k + \epsilon)$, where μ^p_k and σ^p_k are the mean and standard deviation of $\widetilde{S}^p_{h,k}$ over all heads for skill k. The resulting $R^p_{h,k}$ is the final head-skill contribution score used for following head selection.

3.3 Separable Head Selection

Building on the head-skill contribution scores, we conduct Separable Head Selection for each skill. We introduce a separation threshold and select a head only when its score for target skill exceeds its score for any other skill by the threshold. We optimize this threshold by maximizing a separability metric defined on the set of heads selected. In this way, we can select a group of heads whose contribution to the target skill exceeds their contribution to any other skill the most. The procedure of head selection is summarized in Alg. 1 in App. B.

Candidate heads. For skill k, we identify its candidate heads based on both their target-skill score and how much it exceeds their scores for any other skill. Specifically, we first require a head to have a target-skill score above a positive fixed threshold z_{th}. We then require its target-skill score to exceed its largest non-target-skill score according to a separation threshold λ:

$$H^p_k(\lambda) = \left\{ h : R^p_{h,k} > z_{\text{th}} \text{ and } R^p_{h,k} \geq \lambda \max_{j \neq k} R^p_{h,j} \right\}. \tag{4}$$

When the largest non-target-skill score is positive, a larger λ requires greater target-skill dominance; when it is non-positive, the separation condition is automatically satisfied once the target-skill score exceeds the positive fixed threshold. Overall, λ controls the precision-recall trade-off in head selection, with larger values yielding fewer but more skill-specific heads. We collect the position-specific candidate heads as $H_k(\lambda) = \{H^p_k(\lambda)\}_{p \in \{\text{preceding,final}\}}$. To obtain the appropriate threshold, we consider $\Lambda = \{\lambda_1, \ldots, \lambda_L\}$ and optimize λ through grid search in the next step.

Skill-adaptive threshold Selection. Since the affinity with other skills differs across skills, we optimize the separation threshold separately for each skill. For skill k, we select the optimal separation threshold λ^*_k by maximizing the separability of $H_k(\lambda)$ on single-skill identification data. For each $\lambda \in \Lambda$, we jointly intervene on $H_k(\lambda)$, with each subset $H^p_k(\lambda)$ applied at its corresponding token positions $\mathcal{T}_p$. Separability is defined as the effect on skill k minus the largest effect on any non-target skill $j \neq k$. We measure these effects using two types of interventions, Patching and

Erasure, which disrupt the head activations in different directions. For a skill j, $\mathrm{Patch}(H, j)$ is the fraction of instances in $\mathcal{I}_j$ whose prediction changes to the counterfactual answer after replacing the outputs of H with counterfactual activations, while $\mathrm{Eras}(H, j)$ is the fraction of originally correct predictions that become incorrect after zeroing out or reversing the outputs of H.

Because different skills respond differently to random-head interventions, raw intervention effects may conflate the contribution of the selected heads with skill-specific intervention sensitivity. We therefore subtract a matched random baseline to isolate the effect attributable to the selected heads. For $\mathrm{a} \in \{\mathrm{Patch}, \mathrm{Eras}\}$, we define the adjusted intervention effect as

$$\mathrm{Adjust}_{\mathrm{a}}(H, j) = \mathrm{a}(H, j) - \mathrm{Random}_{\mathrm{a}}(H, j), \tag{5}$$

where $\mathrm{Rand}_{\mathrm{a}}(H, j)$ is the effect of randomly selected heads matched to H in number, layer range and intervention token positions. Then, we define the separability of $H_k(\lambda)$ as

$$\mathrm{Sep}_{\mathrm{a}}\left(H_k(\lambda)\right) = \mathrm{Adjust}_{\mathrm{a}}\left(H_k(\lambda), k\right) - \max_{j \neq k} \mathrm{Adjust}_{\mathrm{a}}\left(H_k(\lambda), j\right). \tag{6}$$

A high separability score indicates that the identified heads have a strong effect on the target skill beyond random intervention while having limited effects on the other skills. Finally, we select the optimal threshold via grid search by maximizing the average of the two separability scores:

$$\lambda_k^* = \arg\max_{\lambda \in \Lambda} \frac{1}{2}(\mathrm{Sep}_{\mathrm{Patch}}(H_k(\lambda)) + \mathrm{Sep}_{\mathrm{Eras}}(H_k(\lambda))). \tag{7}$$

The resulting $H_k(\lambda_k^*)$ constitutes the final heads for skill k.

3.4 Evaluation Metrics of Separability and Composability

Beyond identifying skill-specific heads, we design novel metrics to evaluate the two desiderata of separability and composability. Separability assesses whether the identified heads contribute strongly to the target skill but weakly to other skills, while composability assesses whether heads of individual skills can be combined to support multi-skill tasks.

Separability. Beyond the single-skill separability defined in Eq. (6), we also evaluate separability on multi-skill tasks. Since their data do not contain counterfactual image pairs, we only use Erasure intervention. For a task requiring skills A and B jointly, we define the relevant head combination as $H_{\mathrm{rel}} = H_A \cup H_B$ and the irrelevant combination as $H_{\mathrm{irr}} = H_C \cup H_D$, where C and D are the two remaining skills. Since the two combinations may contain different numbers of heads, we use the adjusted Erasure effects in Eq. (5); subtracting the matched random baselines also mitigates differences in intervention effects caused by head count. For a fixed task $A + B$, we abbreviate $\mathrm{Adjust}_{\mathrm{Eras}}(H, A + B)$ as $\mathrm{Adjust}_{\mathrm{Eras}}(H)$. We define multi-skill separability as

$$\mathrm{Sep}(A + B) = \mathrm{Adjust}_{\mathrm{Eras}}(H_{\mathrm{rel}}) - \mathrm{Adjust}_{\mathrm{Eras}}(H_{\mathrm{irr}}). \tag{8}$$

A high score indicates that the identified heads have strong effects when their associated skills are required by the task but weak effects when they are irrelevant.

Composability. On multi-skill tasks, if the heads identified for the required skills are composable, jointly intervening on them should produce a larger effect than intervening on the heads of either skill alone. For the multi-skill task $A + B$ defined above, we first quantify absolute composability as the average difference between the joint and individual Erasure effects:

$$\mathrm{Comp}_{\mathrm{Abs}}(A + B) = \frac{1}{2}\left[(\mathrm{Eras}(H_{\mathrm{rel}}) - \mathrm{Eras}(H_A)) + (\mathrm{Eras}(H_{\mathrm{rel}}) - \mathrm{Eras}(H_B))\right]. \tag{9}$$

A larger value indicates that the heads identified for the required skills can better combine to support the corresponding multi-skill task.

We further introduce a stricter measure, relative composability, to control for the effect of head count. Let R_A and R_B denote randomly selected heads matched in number to H_A and H_B, respectively. We define relative composability as

$$\mathrm{Comp}_{\mathrm{Rel}}(A + B) = \frac{1}{2}\left[(\mathrm{Eras}(H_{\mathrm{rel}}) - \mathrm{Eras}(H_A \cup R_B)) + (\mathrm{Eras}(H_{\mathrm{rel}}) - \mathrm{Eras}(H_B \cup R_A))\right]. \tag{10}$$

Table 2: Separability evaluation results on single-skill tasks. Each skill block reports the separability of the heads identified for that skill. *Avg.* is the average of the Patch- and Erasure-based separability scores, while *Overall* averages the corresponding scores across the four skills. Best and second-best results within each model are shown in **bold** and underline, respectively.

Model	Method	Counting Separability			Position Separability			Size Separability			Distance Separability			Overall		
		Patch↑	Erasure↑	Avg.↑	Patch↑	Erasure↑	Avg.↑	Patch↑	Erasure↑	Avg.↑	Patch↑	Erasure↑	Avg.↑	Patch↑	Erasure↑	Avg. ↑
Qwen2.5-VL 7B-Instruct	VFL	-2.54	-0.56	-1.55	-1.74	-0.05	-0.89	-0.60	2.32	0.86	-0.67	-3.51	-2.09	-1.39	-0.45	-0.92
	V-SEAM	21.81	-1.80	10.00	34.31	-0.86	16.73	-33.29	0.07	-16.61	-22.57	-7.37	-14.97	0.07	-2.49	-1.21
	NOTICE	-0.65	-1.79	-1.22	1.02	**8.99**	5.01	6.95	1.33	4.14	21.09	8.28	14.68	7.10	4.20	5.65
	Ours	**55.03**	**28.87**	**41.95**	**41.70**	6.28	**23.99**	**26.82**	**13.29**	**20.05**	**34.38**	**20.84**	**27.61**	**39.48**	**17.32**	**28.40**
Qwen3-VL 8B-Instruct	VFL	-1.55	-2.95	-2.25	-0.30	-0.49	-0.39	-0.20	-1.21	-0.71	-0.18	-0.22	-0.20	-0.56	-1.22	-0.89
	V-SEAM	4.58	34.35	19.47	**50.93**	**9.76**	**30.34**	-40.61	7.14	-16.74	11.40	-9.81	0.80	6.58	10.36	8.47
	NOTICE	38.78	6.94	22.86	0.00	-0.20	-0.10	5.83	1.78	3.81	**65.82**	**54.06**	**59.94**	27.61	15.65	21.63
	Ours	**64.22**	**43.85**	**54.03**	21.68	-4.09	8.80	**26.27**	**23.56**	**24.91**	60.44	24.72	42.58	**43.15**	**22.01**	**32.58**
Gemma-4 E4B-it	VFL	1.84	-2.79	-0.47	-2.79	-10.80	-6.79	-1.65	0.71	-0.47	-0.34	-3.69	-2.01	-0.73	-4.14	-2.44
	V-SEAM	-8.36	-9.57	-8.96	22.97	2.79	**12.88**	-27.98	-4.68	-16.33	-43.28	0.28	-21.50	-14.16	-2.79	-8.48
	NOTICE	9.48	0.96	5.22	1.51	**13.64**	7.57	21.39	-1.03	10.18	9.73	**24.88**	17.30	10.53	9.61	10.07
	Ours	**31.89**	**18.83**	**25.36**	**23.18**	-3.81	9.69	**35.03**	**17.17**	**26.10**	**24.85**	13.11	**18.98**	**28.74**	**11.33**	**20.03**

A large value indicates that combining the heads of the two required skills produces a greater effect than combining the heads of either skill with random heads. We define the final composability score as the average of these two measures:

$$\mathrm{Comp}(A+B) = \frac{1}{2}(\mathrm{Comp}_{\mathrm{Abs}}(A+B) + \mathrm{Comp}_{\mathrm{Rel}}(A+B)). \tag{11}$$

4 Experiments

4.1 Experimental Setup

Models. We conduct experiments on three instruction-tuned VLMs: Qwen/Qwen2.5-VL-7B-Instruct (Bai et al., 2025b), Qwen/Qwen3-VL-8B-Instruct (Bai et al., 2025a), and google/gemma-4-E4B-it (Team et al., 2026). Our experiments cover models from different families and generations.

Data. We use the single-skill and multi-skill data described in Sec. 3.1. The single-skill identification split is used for head scoring and selection, while the single-skill evaluation split and all multi-skill data are used for evaluation. More details on data construction, statistics, and examples are provided in App. A.

Implementation Details. We set the target-skill score threshold to $z_{\mathrm{th}} = 0.3$ for Qwen models and $z_{\mathrm{th}} = 0.1$ for the Gemma model. We consider separation threshold values $\Lambda = \{0, 0.5, 1.0, \ldots, 5.0\}$. Each random baseline is computed by averaging over 10 trials. Random heads are sampled from the continuous layer range between the minimum and maximum layers of the corresponding identified heads. For Erasure, we scale the selected head outputs by 0, -0.25, and -0.5, and report the average effect across the three intervention strengths.

Baselines. We compare with three representative mechanistic localization methods for VLMs: VFL (Shi et al., 2025), V-SEAM (Wang et al., 2025), and NOTICE (Golovanevsky et al., 2025). Because they do not consider multiple surface forms, we apply them using only Option Content with natural photographs for both identification and evaluation, whereas our method uses all surface forms. For V-SEAM, we select the top-20 scoring heads per skill; for NOTICE, we select heads scoring more than two standard deviations above the skill mean and remove heads selected for multiple skills; for VFL, we select the highest-scoring layer per skill and intervene on all its heads.

4.2 Main Results

Evaluating separability on single-skill tasks. For each skill, we evaluate the separability of its identified heads on the four single-skill tasks, as defined in Eq. (6). Results are shown in Tab. 2. Our

Table 3: Separability and composability evaluation results on multi-skill tasks. *Sep.* denotes the separability score, and *Comp.* is the composability score. *Rank* denotes the harmonic mean of the task-wise method ranks across the three multi-skill tasks, computed separately for Sep. and Comp. Best and second-best results within each model are shown in **bold** and underline, respectively.

Model	Method	Position + Size		Counting + Position		Distance + Size		Average			
		Sep. ↑	Comp. ↑	Sep. ↑	Comp. ↑	Sep. ↑	Comp. ↑	Sep. ↑	Rank ↓	Comp. ↑	Rank ↓
Qwen2.5-VL 7B-Instruct	VFL	-0.52	-0.24	-3.09	-1.36	5.68	1.23	0.69	2.57	-0.12	3.60
	V-SEAM	-3.11	0.00	-4.08	0.16	**7.35**	0.00	0.05	2.00	0.05	3.27
	NOTICE	0.26	0.38	5.18	1.54	4.12	**6.29**	3.19	2.25	2.74	1.50
	Ours	**28.97**	**22.04**	**29.26**	**16.74**	-7.62	1.74	**16.87**	**1.33**	**13.50**	**1.20**
Qwen3-VL 8B-Instruct	VFL	0.00	0.00	0.42	-2.14	-0.57	-0.77	-0.05	3.00	-0.97	3.60
	V-SEAM	12.45	4.22	-6.67	-0.84	4.84	0.23	3.54	2.40	1.20	2.57
	NOTICE	-0.50	0.00	4.38	4.07	**18.68**	12.55	7.52	1.71	5.54	2.25
	Ours	**32.70**	**30.73**	**35.10**	**7.72**	-2.52	**16.09**	**21.76**	**1.33**	**18.18**	**1.00**
Gemma-4 E4B-it	VFL	12.20	5.00	**27.22**	**5.77**	-11.73	-3.04	**9.23**	1.71	2.58	1.71
	V-SEAM	-3.86	-1.36	25.41	1.60	-0.65	-0.72	6.97	2.77	-0.16	3.60
	NOTICE	6.66	0.82	-9.16	5.63	**27.78**	**12.86**	8.43	1.89	6.44	**1.64**
	Ours	**14.81**	**5.16**	5.37	4.99	-0.31	11.86	6.62	**1.64**	**7.34**	**1.64**

method achieves the best overall separability on all three VLMs, with average scores of 28.40, 32.58, and 20.03 on Qwen2.5-VL, Qwen3-VL, and Gemma-4, respectively, exceeding the strongest baseline by 22.75, 10.95, and 9.96 points. The gains are broadly consistent across both intervention types and the four skills, rather than being driven by a particular setting. On Qwen2.5-VL, our method ranks first on 11 of the 12 skill-level metrics and second on the remaining Position Erasure separability. These results show that our method, consisting of Surface-Form-Agnostic Head Scoring and Separable Head Selection, can effectively identify separable skill mechanisms in VLMs. The identified skill-specific heads for all models are provided in App. C.1. More detailed single-skill separability results and an analysis of separability differences among skill mechanisms are reported in App. C.3 and App. C.2, respectively.

Evaluating separability and composability on multi-skill tasks. We further evaluate the separability and composability of identified heads on three multi-skill tasks. The metrics are defined in Eq. (8) and Eq. (11), respectively. To measure consistency across tasks, we also report the harmonic mean of the task-wise method ranks for each metric. Results are shown in Tab. 3. Our method achieves the strongest overall performance across all three VLMs. For composability, it achieves the highest average score on all models, outperforming the strongest baselines by 10.77, 12.64, and 0.90 points on Qwen2.5-VL, Qwen3-VL, and Gemma-4, respectively, while also obtaining the best rank on the two Qwen models and tying for the best on Gemma-4. For separability, our method achieves both the highest average score and the best rank on Qwen2.5-VL and Qwen3-VL, exceeding the second-best averages by 13.68 and 14.24 points, respectively. Although its average separability on Gemma-4 is not the highest, our method achieves the best rank, indicating more consistent strong separability across the three tasks rather than an average dominated by an exceptionally high score on a single task. We show more detailed results in App. C.4.

4.3 Ablation Studies

The effectiveness of separable head selection. We compare our method against three alternative head selection strategies on Qwen2.5-VL while keeping the head scoring procedure fixed. Top-K selects the top-20 scoring heads for each skill, following prior works (Bi et al., 2025; Wang et al., 2025). Top-K + Dedup. further removes heads selected for more than one skill. Global λ^* uses the same head-selection criterion as ours but optimizes a single separation threshold shared across all skills by maximizing their average separability, yielding $\lambda = 1.5$. Results are shown in Tab. 4. Top-K performs worst, with negative separability on both single-skill and multi-skill evaluations, showing that selecting heads solely by their target-skill scores can yield heads with stronger effects on other skills than on the target skill. It also achieves very limited composability. Top-K + Dedup. substantially improves performance by removing shared heads, but this hard separation ignores their relative contributions across skills and may discard heads that contribute much more to one skill than

Table 4: Evaluation results of different head selection strategies. $\mathrm{Sep}_{\mathrm{Patch}}$ and $\mathrm{Sep}_{\mathrm{Eras}}$ are separability averaged over the four skills, while multi-skill Sep. and Comp. are averaged over the three multi-skill tasks. Best and second-best results are shown in **bold** and underline, respectively.

Head Selection Method	Single-Skill			Multi-Skill	
	$\mathrm{Sep}_{\mathrm{Patch}}$	$\mathrm{Sep}_{\mathrm{Eras}}$	Avg.	Sep.	Comp.
Top-K ($K = 20$)	-7.94	-1.44	-4.69	-3.94	0.42
Top-K + Dedup. ($K = 20$)	13.04	3.76	8.40	0.26	3.79
Global λ^* ($\lambda = 1.5$)	30.08	16.99	23.54	16.20	13.28
Skill-specific λ^* (Ours)	**39.48**	**17.32**	**28.40**	**16.87**	**13.50**

Table 5: Separability evaluation results on single-skill tasks under different surface forms. $\mathrm{Sep}_{\mathrm{Patch}}$ and $\mathrm{Sep}_{\mathrm{Eras}}$ are the average separability over the four skills. Best and second-best results are shown in **bold** and underline, respectively.

Answer Format for Identification	Answer Format for Evaluation					
	Option Content			Option Label		
	$\mathrm{Sep}_{\mathrm{Patch}}$	$\mathrm{Sep}_{\mathrm{Eras}}$	Avg.	$\mathrm{Sep}_{\mathrm{Patch}}$	$\mathrm{Sep}_{\mathrm{Eras}}$	Avg.
Option Content	**48.93**	**31.19**	**40.06**	15.35	-2.01	6.67
Option Label	18.56	3.63	11.10	**25.31**	**11.68**	**18.49**
Content + Label (Ours)	42.97	22.09	32.53	23.96	5.86	14.91
Visual Style for Identification	**Visual Style for Evaluation**					
	Natural			Artificial		
	$\mathrm{Sep}_{\mathrm{Patch}}$	$\mathrm{Sep}_{\mathrm{Eras}}$	Avg.	$\mathrm{Sep}_{\mathrm{Patch}}$	$\mathrm{Sep}_{\mathrm{Eras}}$	Avg.
Natural	**40.62**	16.34	28.48	38.20	16.49	27.34
Artificial	36.83	17.93	27.38	38.90	18.43	28.66
Natural + Artificial (Ours)	39.58	**19.14**	**29.36**	**39.05**	**18.70**	**28.87**

to the others. Global λ^* further improves performance by considering relative contributions across skills, but a shared separation threshold cannot account for differences in affinity between each target skill and the other skills. In contrast, our method optimizes a skill-specific λ^* and achieves the best results across all metrics, reaching 28.40 average single-skill separability, 16.87 multi-skill separability, and 13.50 composability. These results show the importance of considering both relative cross-skill contributions and skill-specific separation thresholds for identifying separable and composable skill heads.

The effectiveness of surface-form-agnostic head scoring. To examine whether incorporating multiple surface forms helps break the spurious correlation between skills and surface forms, we conduct ablation experiments on Qwen2.5-VL. For experiments on answer formats, we fix the visual style to natural photographs and identify heads using data with Option Content, Option Label, or both formats, followed by evaluation under each format. For experiments on visual styles, we consider both answer formats and identify heads from Natural Photographs, Artificial Images (Cartoon and Oil Painting), or their combination, followed by evaluation under each style. We evaluate the separability of identified heads on single-skill tasks, with results shown in Tab. 5. We can observe that heads identified from a restricted set of surface forms perform much better on matched surface forms than on different ones, indicating strong surface-form dependence. For example, identification using Option Content yields an average separability of 40.06 on Option Content but only 6.67 on Option Label. In contrast, jointly using both answer formats to identify heads yields strong separability on both evaluation formats. For visual styles, heads identified jointly from Natural and Artificial images achieve the highest average separability across the two evaluation styles. These results demonstrate that Surface-Form-Agnostic Head Scoring produces head scores that reflect the contribution to the skill function rather than to particular surface forms. Additional ablations on the token-position strategy are provided in App. C.5.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic datasets, assist with the design of and provide feedback on research methodology and experiments, support method implementation and translation, and clean and reformat datasets. We did not use generative AI tools to develop theoretical models or conceptual frameworks, support qualitative or thematic data analysis, or interpret experimental results. The following uses are not applicable to this work: formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, and proposing or refining hypotheses. We have reviewed all AI-assisted work. In particular, we manually verified the soundness of the research methodology and experimental design, assessed the reliability of the generated data, and verified and tested LLM-generated code for correctness. We take responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Md Atabuzzaman, Ali Asgarov, and Chris Thomas. Benchmarking and mitigating mcqa selection bias of large vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 33536–33550, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo 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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Jing Bi, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, Bingjie Wang, and Chenliang Xu. Unveiling visual perception in language models: An attention head analysis approach. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4135–4144. IEEE, 2025.

Liwei Che, Zhiyu Xue, Yihao Quan, Benlin Liu, Zeru Shi, Michelle Hurst, Jacob Feldman, Ruixiang Tang, Ranjay Krishna, and Vladimir Pavlovic. Counting circuits: Mechanistic interpretability of visual reasoning in large vision-language models. *arXiv preprint arXiv:2603.18523*, 2026.

Hang Chen, Xinyu Yang, Jiaying Zhu, and Wenya Wang. Skill path: Unveiling language skills from circuit graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30210–30217, 2026.

Mayee Chen, Nicholas Roberts, Kush Bhatia, Jue Wang, Ce Zhang, Frederic Sala, and Christopher Ré. Skill-it! a data-driven skills framework for understanding and training language models. *Advances in Neural Information Processing Systems*, 36:36000–36040, 2023.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Michal Golovanevsky, William Rudman, Vedant Palit, Carsten Eickhoff, and Ritambhara Singh. What do VLMs NOTICE? a mechanistic interpretability pipeline for Gaussian-noise-free text-image corruption and evaluation. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association*

for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 11462–11482, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.571. URL https://aclanthology.org/2025.naacl-long.571/.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702. IEEE, 2019.

Mohamed Insaf Ismithdeen, Muhammad Uzair Khattak, and Salman Khan. Promptception: How sensitive are large multimodal models to prompts?, 2025. URL https://arxiv.org/abs/2509.03986.

Junyang Ji, Qifan Liu, Wenming Yang, and Zhihai He. Causallens: Sensitivity-guided multi-head causal intervention for hallucination mitigation in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4199–4209, 2026.

Yanbei Jiang, Xueqi Ma, Shu Liu, Sarah Monazam Erfani, Tongliang Liu, James Bailey, Jey Han Lau, and Krista A Ehinger. Investigating the functional roles of attention heads in vision language models: Evidence for reasoning modules. *arXiv preprint arXiv:2512.10300*, 2025.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Feiyang Kang, Mahavir Dabas, Myeongseob Ko, and Ruoxi Jia. Characterizing model-native skills. *arXiv preprint arXiv:2604.17614*, 2026.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9339–9350. IEEE, 2025.

Miaoyu Li, Qin Chao, and Boyang Li. Two causally related needles in a video haystack. *Advances in Neural Information Processing Systems*, 38, 2025.

Tianle Li, Jihai Zhang, Yongming Rao, and Yu Cheng. Unveiling the compositional ability gap in vision-language reasoning model. *Advances in Neural Information Processing Systems*, 38: 126574–126592, 2026.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Fabio Rosenthal, Sebastian Schmidt, Thorsten Graf, Thorsten Bagodonat, Stephan Günnemann, and Leo Schwinn. Unexplored flaws in multiple-choice vqa evaluations. *arXiv preprint arXiv:2511.22341*, 2025.

Timothy Schaumlöffel, Martina G Vilas, and Gemma Roig. Mechanisms of object localization in vision-language models. *arXiv preprint arXiv:2605.19792*, 2026.

Cheng Shi, Yizhou Yu, and Sibei Yang. Vision function layer in multimodal llms. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 27510–27537. Curran Associates, Inc., 2025. doi: 10.52202/085713-0926. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/27c4e15d9af120d7fef04432c7db577f-Paper-Conference.pdf.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Juyeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adamek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, Arthur Bražinskas, Ryan Burnell, Robert Busa-Fekete, Mu Cai, Daniele Calandriello, Glenn Cameron, Charlotte Caucheteux, Rahma Chaabouni, Garima Chadha, Jetha Chan, Blake Jianhang Chen, Jesse Chen, Lin Chen, Xu Chen, Derek Cheng, Tzu hsiang Chien, Nikolai Chinaev, Yi Chou, Zhaohui Chu, Benjamin Coleman, Pooja Consul, Sam Conway-Rahman, Scott Crowell, Dylan Cutler, Vivek Dani, Samira Daruki, Anil Das, Daniel Deutsch, Nishanth Dikkala, Li Ding, Qiuhan Ding, Shenil Dodhia, Konstantin Donhauser, Tulsee Doshi, Anca Dragan, Alex Druinsky, Sahil Dua, Zoltan Egyed, Danielle Eisenbud, Daniel Eppens, Cindy Fan, Bahare Fatemi, Yassir Fathullah, Vlad Feinberg, Milen Ferev, Sebastian Flennerhag, Takumi Fujimoto, João Gabriel Oliveira, Isaac Galatzer-Levy, João Gante, Simon Geisler, Soham Ghosal, Antonious M. Girgis, Tamara von Glehn, Alec Go, Alhaad Gokhale, Alex Grills, Yiming Gu, Mayank Gupta, Pramod Gupta, Guru Guruganesh, Raia Hadsell, Hamza Harkous, Jitendra Harlalka, Demis Hassabis, Anja Hauth, Joe Heyward, Arian Hosseini, Chih-Yang Hsia, I-Hung Hsu, Xiaopeng Huang, Yangsibo Huang, Kevin Hui, Adrian Hutter, Te I, Fotis Iliopoulos, Advait Jain, Ganesh Jawahar, Ziwei Ji, Qilin Jin, Melvin Johnson, Kandarp Joshi, Arun Kandoor, Wang-Cheng Kang, Koray Kavukcuoglu, Mehran Kazemi, Kathleen Kenealy, Amr Khalifa, Phoebe Kirk, Ivan Korotkov, Suraj Kothawade, Vitaly Kovalev, Neel Kovelamudi, Adam Kraft, Ravin Kumar, Vivek Kumar, Harish Kuppam, Justin Lannin, Chen-Yu Lee, Seungji Lee, Dmitry Lepikhin, Alon Levkovitch, Dongdong Li, Qiujia Li, Valentin Liévin, Ethan Lin, Ziqian Lin, Casper Liu, Tianlin Liu, Tianqi Liu, Xin Liu, Ivan Lobov, Mayank Lunayach, Min Ma, Gagan Madan, Andrii Maksai, Eric Malmi, Michal Matuszak, Daniel McDuff, Gaurav Menghani, Maciej Mikuła, Daniil Mirylenka, Karolis Misiunas, Vedant Misra, Andreea Mitran, Kareem Mohamed, Maksim Mukha, Eric Noland, James O'Donnell, Brendan O'Donoghue, Kate Olszewska, Bernett Orlando, Wanqiong Pan, Rina Panigrahy, Unnati Parekh, Nicolas Perez-Nieves, Chunjong Park, Eric Paskie, Liqian Peng, Bryce Petrini, Slav Petrov, Jonas Pfeiffer, Bilal Piot, Martyna Plomecka, Siim Poder, Octavio Ponce, Arijit Pramanik, David Racz, Anish Rajan, Michelle Ramanovich, Anand Rao, Marvin Ritter, Vitor Rodrigues, Evan Rosen, Mikołaj Rybiński, Noveen Sachdeva, Michaël E. Sander, Rohit Sathyanarayana, Sagar Savla, Samuel Schmidgall, Tal Schuster, George Scrivener, Benoit Seguin, Andrew Sellergren, Aliaksei Severyn, Izhak Shafran, Dhruv Shah, Bobak Shahriari, Yuan Shangguan, Ashish Shenoy, Pradeep Shenoy, Rakesh Shivanna, Pauline Sho, Lucas Spangher, Wojciech Stokowiec, Tim Strother, Yao Su, Yinghao Sun, Mukund Sundararajan, Andrea Tacchetti, Mor Hazan Taege, Pouya Tafti, Jean Tarbouriech, Chetan Tekur, Shantanu Thakoor, Rahul Thapa, Madeleine Traverse, Lenart Treven, Tao Tu, Chien Te Tung, Çağlar Ünlü, Petar Veličković, Malini Pooni Venkat, Sagar Gubbi Venkatesh, Vidya Venkiteswaran, Francesco Visin, Alex Vitvitskyi, Kiran Vodrahalli, Weiyi Wang, Xin Wang, Tris Warkentin, Jan Wassenberg, John Wieting,

Cindy Wu, Lechao Xiao, Hao Xu, Yuhui Xu, Fuzhao Xue, Arun Yadav, Jun Yan, Antoine Yang, Lin Yang, Ming-Hsuan Yang, Ziyu Ying, Jae Hyeon Yoo, Morteza Zadimoghaddam, Sajjad Zafar, Fred Zhang, Jiageng Zhang, Jianyi Zhang, Xiaofan Zhang, Chao Zhao, David Zhou, and Chen Zou. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Anthony Tiong, Junqi Zhao, Boyang Li, Junnan Li, Steven Hoi, and Caiming Xiong. What are we measuring when we evaluate large vision-language models? an analysis of latent factors and biases. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3427–3454, 2024.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *arXiv preprint arXiv:2004.12265*, 2020.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Qidong Wang, Junjie Hu, and Ming Jiang. V-seam: Visual semantic editing and attention modulating for causal interpretability of vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17407–17431, 2025.

Zhenran Xu, Senbao Shi, Baotian Hu, Longyue Wang, and Min Zhang. Multiskill: Evaluating large multimodal models for fine-grained alignment skills. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1506–1523, 2024.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024.

Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Junqi Zhao, Allison Koenecke, Boyang Li, and Lu Wang. Sphere: Unveiling spatial blind spots in vision-language models through hierarchical evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11591–11609, 2025.

Dezhi Zhao, Xin Liu, Xiaocheng Feng, Hui Wang, and Bing Qin. Probing and boosting large language models capabilities via attention heads. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28518–28532, 2025.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*, 2022.

Zhen Zhu, Yiming Gong, Yao Xiao, Yaoyao Liu, and Derek Hoiem. How to teach large multimodal models new skills. *arXiv preprint arXiv:2510.08564*, 2025.

Table 6: Dataset statistics for the single-skill and multi-skill tasks.

Task	Identification		Evaluation	
	# Questions	# Instances	# Questions	# Instances
Single-skill				
Counting	71	728	30	332
Position	127	1,168	54	488
Size	80	772	34	312
Distance	78	760	34	340
Multi-skill				
Position + Size	–	–	200	200
Counting + Position	–	–	48	48
Distance + Size	–	–	200	200

A Skill Data Construction

A.1 Single-Skill Dataset Construction

We derive four single-skill datasets from SPHERE (Zhang et al., 2025), covering counting, position, size, and distance. Because our analysis focuses on the first generated token, we retain questions whose answer options can either be represented by a single token or distinguished by their first tokens, rephrasing other questions where possible to satisfy this requirement and excluding those that remain incompatible.

By modifying only the question-relevant visual property while preserving the remaining scene content, counterfactual images are constructed separately for each skill. For position, where the retained questions focus on left-right relations, we horizontally flip the original image to reverse the correct answer. For the other skills, we use `GPT-5.4-2026-03-05` to construct image editing instructions and `GPT-image-1.5` to apply the edits. Specifically, for counting, the instruction specifies adding or removing target objects to change the count. For size, GPT-5.4 generates an instruction to scale one of the compared objects and reverse their size or height relation. For distance, GPT-5.4 identifies the relevant objects and specifies moving one of them to reverse their proximity relation. We use GPT-5.4 to verify the generated images and retain only those for which the model returns the expected answer.

We then generate cartoon and oil-painting versions of both original and counterfactual images using `GPT-image-1.5`. For the cartoon and oil-painting variants, we use the prompts *"Restyle this image as a clean cartoon illustration."* and *"Restyle this image as a detailed oil painting."*, respectively. Both prompts further instruct the model to *"Preserve the exact scene layout and semantic content. Do not add, remove, move, mirror, crop, or occlude any objects. Keep all object counts, relative positions, size relations, and distance relations unchanged."* We use GPT-5.4 to verify answer consistency for the generated images, and retain style variants only when the model returns the corresponding expected answer.

For these data, we use two answer formats: Option Content, where the model directly outputs an allowed answer, and Option Label, where three choices contain the original answer y^+, the counterfactual answer y^-, and one additional distractor. The choice order is randomized, and a single-token-compatible label scheme is randomly selected from {A, B, C}, {1, 2, 3}, and {First, Second, Third}. Combining the two answer formats with the three visual styles yields six surface-form combinations for each question. For each combination, either image in a counterfactual pair can serve as x_i^+, with the other serving as x_i^-. Thus, each original VQA question can yield up to $6 \times 2 = 12$ instances (q_i, x_i^+, x_i^-) for counterfactual activation patching. We split questions into identification and evaluation sets at a 70:30 ratio. All instances derived from the same question are kept in the same split. Dataset statistics and examples are provided in Tab. 6 and Fig. 2, respectively.



Figure 2: Examples of single-skill data.

A.2 Multi-Skill Data Construction

Our multi-skill data cover three tasks: Counting + Position (C+P), Position + Size (P+S), and Distance + Size (D+S), as shown in Fig. 3. Each question is designed such that answering it requires the joint application of the two designated skills, while the other two skills are not needed to determine the answer. For C+P, we use the corresponding questions from SPHERE (Zhang et al., 2025), excluding questions associated with images that appear in the single-skill identification set. These questions require identifying target objects according to their spatial relation and then counting them.

For P+S and D+S, it is difficult to isolate the designated skills in natural photographs because different visual properties are usually entangled. For example, due to perspective, the distance of an object from the camera affects its apparent size, making size comparisons dependent on both size and distance. We therefore construct controlled synthetic scenes in Blender, where object position, size, and distance can be manipulated independently. Each scene contains a reference object, two target objects, and several irrelevant objects instantiated as simple geometric shapes, with object

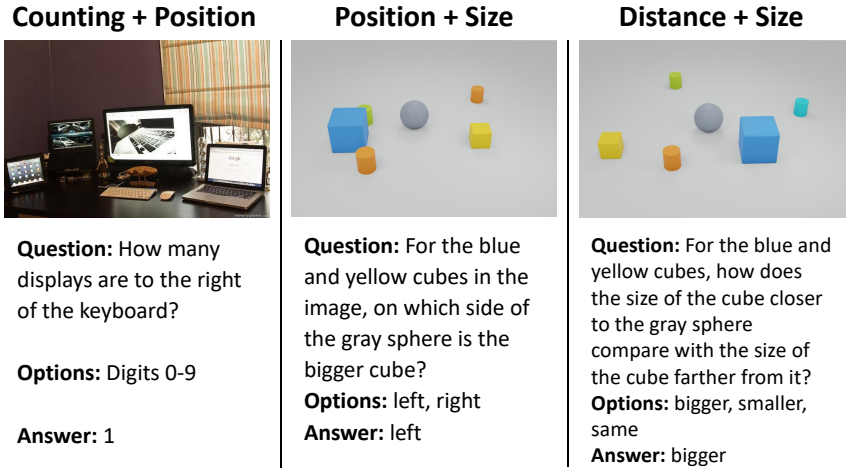

Figure 3: Examples of multi-skill data.

shapes randomly assigned to different roles. For P+S, the two target objects differ in size and are placed on different sides of the reference object. The question asks on which side of the reference object the bigger or smaller target lies, requiring both Position and Size skills. The two targets are placed at the same depth relative to the camera so that their size relation is not affected by perspective. For D+S, the two target objects differ in both their distance to the reference object and their size. The question first identifies the target that is closer to or farther from the reference object and then compares its size with the other target, requiring both Distance and Size skills. Distance is defined relative to the reference object rather than the camera, and the two targets are placed at the same camera depth to prevent perspective from affecting their size relation.

As with the single-skill data, we use GPT-5.4 to verify the constructed examples and retain only those for which the model returns the expected answer. Since the multi-skill data are used only for evaluation rather than head identification, we do not construct counterfactual image pairs or additional surface-form variants. All questions use the Option Content format. We show the dataset statistics and examples in Tab. 6 and Fig. 3, respectively.

B Algorithms

Alg. 1 summarizes the overall procedure for identifying skill-specific heads.

C Additional Experimental Results

C.1 Identified Skill-Specific Heads

Figs. 4-5 report the final skill-specific heads identified for the three VLMs. For each skill, we show the selected separation threshold λ^* together with the preceding- and final-position heads. The optimal separation thresholds differ across skills, supporting the use of skill-specific threshold optimization rather than a single shared threshold. We also observe substantial differences in both the number and composition of preceding- and final-position heads for each skill, indicating that our token-position-aware identification captures distinct heads that contribute to a skill at different stages of computation.

C.2 Analysis of Separability Differences Across Skills

To analyze separability differences among skill mechanisms, we extract the maximum Avg. separability achieved for each skill among all methods from Tab. 2. And we rank the four skills within each model, as shown in Tab. 7. We observe consistent patterns for Counting and Position across the three VLMs. Counting ranks first or second in all three models and achieves the best harmonic mean rank of 1.50, whereas Position ranks third or fourth and has the worst harmonic mean rank of 3.27.

Algorithm 1 Identification of Skill-Specific Heads

Require: Instances of each skill $\{\mathcal{I}_k\}_{k\in\mathcal{K}}$, attention heads $\mathcal{H}$, target-skill score threshold z_{th}, separation-threshold grid $\Lambda = \{\lambda_1, \ldots, \lambda_L\}$
Ensure: Skill-specific heads $\{H_k\}_{k\in\mathcal{K}}$

Stage 1: Surface-Form-Agnostic Head Scoring
for each skill $k \in \mathcal{K}$ **do**
  for each head $h \in \mathcal{H}$ **do**
    for each instance $i \in \mathcal{I}_k$ **do**
      Perform counterfactual activation patching on h over each token-position set $\mathcal{T}_p$, $p \in \{\text{preceding}, \text{final}\}$.
      Compute the per-instance effect $\Delta^p_{i,h}$ as the change in the logit margin using Eq. (1).
    end for
    Average the per-instance effects $\{\Delta^p_{i,h}\}_{i\in\mathcal{I}_k}$ to obtain $S^p_{h,k}$ using Eq. (2).
    Compute the coverage as the fraction of instances with positive effects, and use it to weight $S^p_{h,k}$ to obtain the coverage-weighted score $\widetilde{S}^p_{h,k}$ using Eq. (3).
  end for
  Normalize the coverage-weighted scores across heads to obtain the final head-skill contribution scores $R^p_{h,k}$.
end for
Stage 2: Separable Head Selection
for each skill $k \in \mathcal{K}$ **do**
  for each separation threshold $\lambda \in \Lambda$ **do**
    Select position-specific candidate heads $H^p_k(\lambda)$ satisfying $R^p_{h,k} > z_{\text{th}}$ and $R^p_{h,k} \geq \lambda \max_{j\neq k} R^p_{h,j}$ using Eq. (4).
    Collect the position-specific subsets as $H_k(\lambda)$, with each $H^p_k(\lambda)$ applied at its corresponding token positions $\mathcal{T}_p$.
    Intervene on $H_k(\lambda)$, measure the effect on every skill and subtract the matched random baseline to obtain adjusted intervention effects using Eq. (5).
    Compute separability as the adjusted effect on target skill k minus the largest adjusted effect on any non-target skill using Eq. (6).
  end for
  Select λ^*_k by grid search to maximize the separability using Eq. (7).
  Set $H_k \leftarrow H_k(\lambda^*_k)$.
end for
return $\{H_k\}_{k\in\mathcal{K}}$

Table 7: Maximum separability achieved for each skill on each VLM. The rank among the four skills within each model is shown in parentheses. For each skill, Avg. Rank denotes the harmonic mean of its ranks across the three VLMs. Among the four skills, the best and second-best results are shown in **bold** and underlined, respectively.

Model	Counting Separability	Position Separability	Size Separability	Distance Separability
Qwen2.5-VL	**41.95 (1)**	23.99 (3)	20.05 (4)	<u>27.61 (2)</u>
Qwen3-VL	<u>54.03 (2)</u>	30.34 (3)	24.91 (4)	**59.94 (1)**
Gemma-4	<u>25.36 (2)</u>	12.88 (4)	**26.10 (1)**	18.98 (3)
Avg. Rank ↓	**1.50**	3.27	2.00	<u>1.64</u>

This suggests that Counting may rely on a relatively distinct head-level mechanism that can be more readily separated from mechanisms associated with other skills. One possible reason for the low separability of Position is that performing other skills often requires first locating the target objects, which may lead to the reuse of Position-related mechanisms. Such reuse may cause Position-related heads to have non-negligible effects on other skills, thereby reducing their separability.

Qwen2.5-VL-7B-Instruct

Counting $\lambda^* = 1.00$

Preceding (27 heads): (11,11), (12,7), (12,24), (12,27), (13,25), (14,11), (14,12), (14,13), (14,14), (14,19), (15,7), (15,15), (16,2), (16,8), (16,20), (17,0), (17,1), (17,14), (17,16), (18,10), (18,14), (18,15), (18,16), (19,24), (20,1), (20,5), (22,13)

Final (8 heads): (18,10), (22,13), (25,14), (25,15), (25,17), (25,18), (26,24), (26,25)

Position $\lambda^* = 1.00$

Preceding (13 heads): (15,20), (15,27), (16,1), (16,5), (17,22), (17,24), (18,21), (19,13), (19,20), (19,23), (19,25), (19,26), (24,27)

Final (11 heads): (19,23), (19,25), (20,5), (20,23), (22,7), (22,8), (24,22), (24,24), (24,26), (26,26), (27,14)

Size $\lambda^* = 1.50$

Preceding (14 heads): (11,24), (12,14), (13,24), (14,15), (14,16), (14,18), (15,1), (15,3), (15,4), (16,3), (17,25), (18,20), (19,21), (20,19)

Final (8 heads): (18,15), (18,18), (20,19), (20,20), (22,21), (22,22), (23,26), (23,27)

Distance $\lambda^* = 1.50$

Preceding (12 heads): (14,0), (14,2), (15,22), (17,18), (17,23), (18,24), (19,15), (19,19), (19,22), (20,23), (20,26), (20,27)

Final (20 heads): (18,8), (18,13), (18,21), (18,24), (19,17), (19,19), (19,22), (19,24), (20,0), (20,26), (21,18), (21,19), (21,20), (22,1), (22,26), (22,27), (23,11), (24,23), (26,16), (27,24)

Qwen3-VL-8B-Instruct

Counting $\lambda^* = 2.00$

Preceding (30 heads): (14,12), (14,16), (15,29), (15,30), (16,4), (16,9), (16,14), (17,1), (17,3), (17,20), (17,21), (17,25), (18,5), (18,7), (18,12), (18,19), (19,10), (19,12), (19,13), (19,14), (21,3), (21,12), (21,14), (21,23), (22,26), (23,12), (23,13), (23,30), (28,16), (34,28)

Final (10 heads): (19,14), (23,12), (28,16), (28,19), (30,13), (31,9), (31,11), (33,7), (34,30), (35,26)

Position $\lambda^* = 1.00$

Preceding (13 heads): (14,30), (15,9), (15,11), (17,7), (18,14), (18,15), (19,24), (20,0), (20,29), (21,10), (21,16), (26,24), (26,25)

Final (11 heads): (21,16), (23,24), (26,24), (26,25), (26,26), (27,8), (29,19), (30,5), (34,17), (35,5), (35,19)

Size $\lambda^* = 2.00$

Preceding (13 heads): (13,19), (14,11), (15,13), (15,18), (16,2), (16,17), (17,14), (18,10), (18,27), (19,0), (19,30), (19,31), (22,5)

Final (11 heads): (22,4), (22,5), (22,7), (24,14), (27,19), (29,17), (30,3), (30,11), (30,25), (33,31), (34,15)

Distance $\lambda^* = 3.00$

Preceding (16 heads): (14,7), (16,18), (18,23), (20,14), (20,28), (20,30), (21,8), (21,11), (21,24), (21,25), (21,27), (21,30), (22,0), (22,10), (22,11), (23,26)

Final (14 heads): (19,21), (20,2), (21,18), (21,25), (21,27), (21,30), (22,0), (22,10), (22,11), (23,3), (23,26), (27,12), (27,17), (28,4)

Figure 4: Identified skill-specific heads in Qwen2.5-VL-7B-Instruct and Qwen3-VL-8B-Instruct. The number in parentheses denotes the size of each subset, and each tuple (A, B) denotes head B in layer A.

C.3 Detailed Single-Skill Separability Results

Tabs. 8-10 report the complete adjusted intervention effects underlying the single-skill separability results in Tab. 2. Each row corresponds to the heads identified for one skill, and the columns show

Gemma-4-E4B-it

Counting $\lambda^* = 3.00$
Preceding (6 heads): `(12,2), (14,7), (16,6), (17,0), (19,0), (21,2)`
Final (5 heads): `(26,0), (30,3), (34,1), (36,2), (39,6)`

Position $\lambda^* = 1.00$
Preceding (2 heads): `(16,4), (17,3)`
Final (8 heads): `(28,0), (28,3), (29,0), (29,2), (31,1), (31,2), (32,2), (33,4)`

Size $\lambda^* = 2.50$
Preceding (8 heads): `(13,0), (15,2), (15,5), (17,1), (18,5), (19,4), (19,6), (21,4)`
Final (5 heads): `(27,0), (29,6), (34,7), (35,0), (41,4)`

Distance $\lambda^* = 2.00$
Preceding (5 heads): `(16,1), (18,3), (19,7), (20,5), (20,7)`
Final (7 heads): `(23,2), (24,2), (27,1), (29,1), (32,1), (35,7), (41,1)`

Figure 5: Identified skill-specific heads in Gemma-4-E4B-it. The number in parentheses denotes the size of each subset, and each tuple (A, B) denotes head B in layer A.

their adjusted Patch and Erasure effects on all four single-skill tasks, as defined in Eq. (5). Separability is obtained by subtracting the largest non-target-skill effect from the target-skill effect, as defined in Eq. (6). The detailed results show that the target-skill effect of the identified heads is generally much larger than their effects on non-target skills, demonstrating the effectiveness of our identification method. The detailed results also reveal different patterns between Patch and Erasure. For example, as shown in Tab. 9, for the Position heads on Qwen3-VL, Patch produces the largest effect on the target skill (77.37), whereas under Erasure, the effects are more similar across skills and the effect on Distance (41.74) even exceeds that on Position (37.65). This shows that Patch and Erasure characterize different functional aspects of the identified heads, motivating the use of both interventions in our evaluation.

C.4 Detailed Multi-Skill Separability and Composability Results

Tabs. 11-13 report the Erasure effects used to compute the multi-skill separability and composability in Tab. 3. For each multi-skill task, we report the raw Erasure effects of the heads identified for each of the two required skills, their union, the union of heads identified for the two irrelevant skills, and head-count-matched random controls. The final separability and composability scores are computed from these quantities according to Eqs. (8)-(11). The detailed results reveal why our method yields negative separability on the Distance + Size task. As shown in Tab. 12, jointly intervening on the Distance and Size heads produces a strong Erasure effect of 42.22, but the irrelevant Counting and Position heads also produce a substantial effect of 37.04. After subtracting their matched random baselines, the latter effect is even slightly larger, leading to a separability of -2.52. This suggests that the negative separability is not caused by weak contributions from the heads of required skills, but by considerable effects from the irrelevant heads. This pattern is consistent with our earlier observation in App. C.2: Position heads may be reused across skills and therefore exert non-negligible effects on them.

C.5 The Effectiveness of Position-Aware Head Identification

To examine the effect of our token-position strategy, we compare variants that differ only in the token positions considered during head identification and evaluation, while keeping all other settings unchanged. Specifically, we compare our method with variants that use only the final position, only the preceding positions, or all positions together. We conduct experiments on Qwen2.5-VL-7B-Instruct for both single-skill and multi-skill evaluation. Results are shown in Tab. 14.

Compared with our strategy, using only the final or preceding positions consistently yields lower performance on both single-skill and multi-skill evaluation. This indicates that skill-related mechanisms are distributed across both position groups, and restricting localization to either one misses

Table 8: Adjusted intervention effects of the skill-specific heads on single-skill tasks for Qwen2.5-VL-7B-Instruct. **Bold** and underlined values denote the target-skill effect and the largest non-target-skill effect for identified heads of each skill, respectively.

Heads	Effects on Counting		Effects on Position		Effects on Size		Effects on Distance		Separability		
	Patch	Erasure	Patch	Erasure	Patch	Erasure	Patch	Erasure	Patch	Erasure	Avg.
Counting	**68.35**	**38.22**	3.93	7.96	13.32	9.35	12.60	7.97	55.03	28.87	41.95
Position	18.44	8.62	**77.82**	**27.51**	31.12	14.11	36.12	21.23	41.70	6.28	23.99
Size	-1.17	-4.67	6.70	5.13	**33.52**	**18.41**	0.34	1.09	26.82	13.29	20.05
Distance	9.36	3.46	11.31	-1.58	10.62	1.13	**45.69**	**24.30**	34.38	20.84	27.61
Overall									39.48	17.32	28.40

Table 9: Adjusted intervention effects of the skill-specific heads on single-skill tasks for Qwen3-VL-8B-Instruct. **Bold** and underlined values denote the target-skill effect and the largest non-target-skill effect for identified heads of each skill, respectively.

Heads	Effects on Counting		Effects on Position		Effects on Size		Effects on Distance		Separability		
	Patch	Erasure	Patch	Erasure	Patch	Erasure	Patch	Erasure	Patch	Erasure	Avg.
Counting	**71.34**	**50.65**	-2.95	4.68	7.12	6.80	-5.29	0.17	64.22	43.85	54.03
Position	14.96	33.92	**77.37**	**37.65**	29.78	34.73	55.69	41.74	21.68	-4.09	8.80
Size	-0.31	-5.67	6.65	5.74	**32.92**	**29.30**	0.17	-2.32	26.27	23.56	24.91
Distance	1.61	-4.99	-1.24	-3.17	-0.58	-7.60	**62.05**	**21.55**	60.44	24.72	42.58
Overall									43.15	22.01	32.58

useful contributions from the other. Importantly, simply including both position groups is not sufficient. When preceding and final positions are treated together, performance remains lower than our separate-position strategy, particularly for multi-skill separability, where the score drops from 16.87 to 3.66. These results show that the gain comes not only from covering both position groups, but also from distinguishing their contributions during head identification, which leads to more separable and composable skill mechanisms.

Table 10: Adjusted intervention effects of the skill-specific heads on single-skill tasks for Gemma-4-E4B-it. **Bold** and underlined values denote the target-skill effect and the largest non-target-skill effect for identified heads of each skill, respectively.

Heads	Effects on Counting		Effects on Position		Effects on Size		Effects on Distance		Separability		
	Patch	Erasure	Patch	Erasure	Patch	Erasure	Patch	Erasure	Patch	Erasure	Avg.
Counting	**15.49**	**10.89**	-16.40	-7.93	-16.53	-17.04	-18.58	-21.83	31.89	18.83	25.36
Position	12.84	11.66	**64.69**	**25.01**	34.61	28.82	41.51	25.48	23.18	-3.81	9.69
Size	-13.23	-11.39	-17.57	6.10	**21.80**	**23.26**	-22.57	-3.87	35.03	17.17	26.10
Distance	-12.07	4.06	-12.06	1.14	-13.42	0.15	**12.79**	**17.16**	24.85	13.11	18.98
Overall									28.74	11.33	20.03

Table 11: Multi-skill evaluation results of our method on Qwen2.5-VL-7B-Instruct. Values in parentheses are the matched-random baselines.

Position + Size		Counting + Position		Distance + Size	
Heads	Erasure	Heads	Erasure	Heads	Erasure
Position	25.33	Counting	32.10	Distance	19.84
Size	8.00	Position	14.81	Size	21.96
Position ∪ Size	39.67 (0.92)	Counting ∪ Position	41.98 (9.63)	Distance ∪ Size	25.66 (22.09)
Distance ∪ Counting	13.50 (3.72)	Distance ∪ Size	16.05 (12.96)	Counting ∪ Position	32.01 (20.82)
Position ∪ Rand(Size)	26.85	Counting ∪ Rand(Position)	33.70	Distance ∪ Rand(Size)	26.11
Size ∪ Rand(Position)	10.37	Position ∪ Rand(Counting)	20.37	Size ∪ Rand(Distance)	27.78
Separability	**28.97**	**Separability**	**29.26**	**Separability**	**-7.62**
Composability(Abs)	**23.01**	**Composability(Abs)**	**18.53**	**Composability(Abs)**	**4.76**
Composability(Rel)	**21.06**	**Composability(Rel)**	**14.95**	**Composability(Rel)**	**-1.29**
Composability	**22.04**	**Composability**	**16.74**	**Composability**	**1.74**

Table 12: Multi-skill evaluation results of our method on Qwen3-VL-8B-Instruct. Values in parentheses are the matched-random baselines.

Position + Size		Counting + Position		Distance + Size	
Heads	Erasure	Heads	Erasure	Heads	Erasure
Position	40.33	Counting	35.42	Distance	25.68
Size	0.50	Position	40.62	Size	17.04
Position ∪ Size	51.17 (0.62)	Counting ∪ Position	46.88 (14.69)	Distance ∪ Size	42.22 (22.69)
Distance ∪ Counting	19.50 (1.65)	Distance ∪ Size	22.92 (25.83)	Counting ∪ Position	37.04 (14.99)
Position ∪ Rand(Size)	37.52	Counting ∪ Rand(Position)	36.04	Distance ∪ Rand(Size)	33.75
Size ∪ Rand(Position)	3.45	Position ∪ Rand(Counting)	44.58	Size ∪ Rand(Distance)	28.07
Separability	**32.70**	**Separability**	**35.10**	**Separability**	**-2.52**
Composability(Abs)	**30.76**	**Composability(Abs)**	**8.86**	**Composability(Abs)**	**20.86**
Composability(Rel)	**30.69**	**Composability(Rel)**	**6.57**	**Composability(Rel)**	**11.31**
Composability	**30.73**	**Composability**	**7.72**	**Composability**	**16.09**

Table 13: Multi-skill evaluation results of our method on Gemma-4-E4B-it. Values in parentheses are the matched-random baselines.

Position + Size		**Counting + Position**		**Distance + Size**	
Heads	Erasure	Heads	Erasure	Heads	Erasure
Position	42.70	Counting	59.72	Distance	42.64
Size	39.46	Position	45.83	Size	40.05
Position ∪ Size	46.85 (12.43)	Counting ∪ Position	58.33 (27.27)	Distance ∪ Size	53.49 (33.28)
Distance ∪ Counting	36.58 (16.97)	Distance ∪ Size	58.33 (32.64)	Counting ∪ Position	46.51 (25.99)
Position ∪ Rand(Size)	44.43	Counting ∪ Rand(Position)	59.86	Distance ∪ Rand(Size)	44.57
Size ∪ Rand(Position)	40.20	Position ∪ Rand(Counting)	47.99	Size ∪ Rand(Distance)	39.28
Separability	**14.81**	**Separability**	**5.37**	**Separability**	**-0.31**
Composability(Abs)	**5.77**	**Composability(Abs)**	**5.56**	**Composability(Abs)**	**12.15**
Composability(Rel)	**4.54**	**Composability(Rel)**	**4.41**	**Composability(Rel)**	**11.57**
Composability	**5.16**	**Composability**	**4.99**	**Composability**	**11.86**

Table 14: Evaluation results of different token-position strategies. $\mathrm{Sep}_{\mathrm{Patch}}$ and $\mathrm{Sep}_{\mathrm{Erasure}}$ are the average separability over the four skills. Sep. and Comp. are the average separability and composability over the three multi-skill tasks. Best and second-best results are shown in **bold** and underline, respectively.

Token Positions	**Single-skill**			**Multi-skill**	
	$\mathrm{Sep}_{\mathrm{Patch}}$	$\mathrm{Sep}_{\mathrm{Erasure}}$	Avg.	Sep.	Comp.
Final Position Only	24.88	8.85	16.86	<u>9.30</u>	<u>12.74</u>
Preceding Positions Only	33.52	<u>13.86</u>	23.69	9.25	11.63
All Positions (Together)	<u>35.50</u>	12.47	<u>23.98</u>	3.66	12.01
All Positions (Separate, Ours)	**39.48**	**17.32**	**28.40**	**16.87**	**13.50**