# Learning from Primitive: Probing Visual Reasoning of LVLMs via Counting

Liwei Che[1*], Zhiyu Xue[2*], Yihao Quan[1*], Benlin Liu[3,4], Zeru Shi[1], Ruixiang Tang[1†], Ranjay Krishna[3,4], and Vladimir Pavlovic[1†]

[1] Rutgers University, Piscataway NJ 08854, USA
[2] University of California, Santa Barbara, CA 93106, USA
[3] University of Washington, Seattle, WA 98195, USA
[4] Allen Institute of AI, Seattle, WA 98105, USA

**Abstract.** Counting serves as a simple but powerful test of a Large Vision-Language Model's (LVLM) reasoning; it forces the model to identify each individual object and then add them all up. In this study, we investigate how LVLMs implement counting using controlled synthetic and real-world benchmarks, combined with mechanistic analyses. Our results show that LVLMs display a human-like counting behavior, with precise performance on small numerosities and noisy estimation for larger quantities. We introduce two novel interpretability methods, Visual Activation Patching and HeadLens, and use them to uncover a structured "counting circuit" that is largely shared across a variety of visual reasoning tasks. Building on these insights, we propose a lightweight intervention strategy that exploits simple and abundantly available synthetic images to fine-tune arbitrary pretrained LVLMs exclusively on counting. Despite the narrow scope of this fine-tuning, the intervention not only enhances counting accuracy on in-distribution synthetic data, but also yields an average improvement of +8.36% on out-of-distribution counting benchmarks and an average gain of +1.54% on complex, general visual reasoning tasks for Qwen2.5-VL. These findings highlight the central, influential role of counting in visual reasoning and suggest a potential pathway for improving overall visual reasoning capabilities through targeted enhancement of counting mechanisms.

**Keywords:** Vision Language Model · Mechanistic Interpretability

## 1 Introduction

Counting is one of the most fundamental yet revealing capacities of visual intelligence. Cognitive science reveals that human numerosity perception is shaped by severe information-processing constraints, resulting in near-perfect accuracy for small sets (subitizing) and noisy estimation for larger quantities [6]. Recent work further demonstrates that these discontinuities arise not from separate cognitive

* Equal contribution.
† Corresponding author.

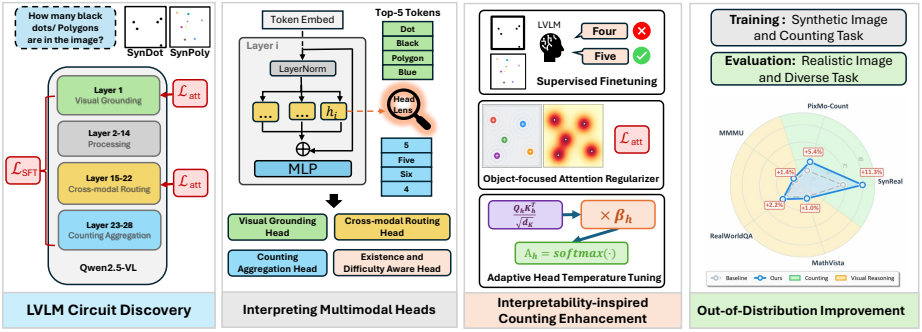



**Fig. 1:** Overview of Main Contributions

modules, but from resource-rational trade-offs [20] between precision, exposure time, and environmental statistics, suggesting that counting reflects a core organizing principle of visual reasoning rather than a task-specific heuristic [7].

Recent LVLMs [3, 22] have demonstrated remarkable generalization across diverse visual reasoning tasks [23, 34, 37, 39]. However, our benchmarks (Tab. 1) reveal a surprising phenomenon: these models struggle to count even fewer than ten simple black dots, a trivial task for human vision. This discrepancy raises fundamental questions regarding the nature of counting ability in LVLMs and its relation to general visual reasoning. Unlike standard recognition, counting cannot be bypassed through semantic memorization or dataset priors. It strictly requires models to individuate discrete entities, maintain intermediate representations, and aggregate information under architectural constraints. Therefore, in this work, we utilize the counting task as a minimalist probe to investigate the internal visual reasoning process of LVLMs and explore how these mechanism discoveries can support more complex, high-level visual reasoning tasks.

Existing works often treat counting as an isolated task, attempting to improve the counting performance of LVLMs with methods like image preprocessing [29], fine-tuning [28], and attention intervention [31]. However, as specialized computer vision methods [15,33] undoubtedly yield better performance, we argue that understanding and evaluating the visual reasoning process of LVLMs via counting should be the primary focus. Closer work such as CountScope [13] and [De|Re] [1] heavily rely on probing methods to understand the counting pattern of LVLMs, which is unreliable due to the simplicity of the representation of their synthetic data. More importantly, they fail to understand the circuit-level mechanism of counting. Therefore, in this work, we investigate how LVLMs execute counting tasks from the dual perspectives of human cognitive alignment and mechanistic interpretability. First, we leverage the controllability of synthetic images to extend activation patching [14] to multimodal inputs. By tracing the flow of counting-related information across model layers, we reveal the cross-modal information routing occurring in the middle layers and the emergence of counting answers in the later layers. To further isolate and verify the roles of

different model components, we propose HeadLens, a novel tool that interprets the output of attention heads into semantic tokens. Through this, we identify four critical categories of attention heads: (1) *Visual Grounding Heads* (Early Layers): Extract foundational visual features (color, shape) directly from image tokens. (2) *Cross-Modal Routing Heads* (Middle Layers): Translate visual information into abstract numerical concepts. (3) *Counting Aggregation Heads* (Late Layers): Attend heavily to text prompts; their top-10 HeadLens decoded tokens highly correlate with the final prediction. (4) *Awareness Heads* (Late Layers): Two specialized heads encoding object existence and difficulty estimations.

Building on the interpretability findings, we design targeted intervention methods tailored to specific layers and heads via parameter tuning, attention regularization, and head temperature tuning. Utilizing 8,000 highly simplified synthetic images (e.g., black dots or colored polygons on a white canvas) that can be generated in minutes on a standard computer, our approach not only improves performance on synthetic data but also achieves up to an 8.36% average accuracy increase on out-of-distribution (OOD) real-world object counting. Furthermore, it yields a 1.54% average accuracy improvement across broader OOD visual tasks (e.g., general VQA, visual math). This demonstrates that counting serves as a vital foundational pillar of general visual intelligence and validates the existence of generalized visual reasoning mechanisms operating as internal circuits within LVLMs. Our main contributions are summarized as follows:

**Cognitive Alignment:** Our investigation reveals that LVLMs exhibit a striking discontinuity in visual counting that mirrors human behavior, characterized by precise subitizing for small sets and noisy estimation for larger quantities. We ground the model's behavioral failures in the topological entanglement of hidden state manifolds.

**Methodological Innovation:** We propose Visual Activation Patching (VAP) and HeadLens, two novel mechanistic interpretability techniques designed to isolate head-specific functions and perform circuit discovery for LVLMs.

**Circuit Discovery:** We identify four distinct functional categories of attention heads essential for visual counting: *visual grounding*, *cross-modal routing*, *counting aggregation*, and *difficulty/existence detection*. We further demonstrate that a great proportion of these specialized heads also serve as foundational components for general visual reasoning.

**Performance Enhancement:** We develop an interpretability-inspired intervention strategy. Using only the simplest synthetic images, we significantly enhance the model's OOD counting robustness and its performance on complex, general visual reasoning benchmarks.

**Related Work.** Recent efforts to improve LVLM counting include image preprocessing pipelines [29], CLIP-based supervision [28], and attention intervention [31]. Concurrent analyses such as CountScope [13] and [De|Re] [1] probe how counting information accumulates across layers, yet remain at the behavioral or probing level without isolating circuit-level mechanisms. On the interpretability side, Logit Lens [27] and Tuned Lens [4] decode layerwise hidden states into vocabulary space, while activation patching [14] enables causal tracing at the

token and layer level. However, these techniques have been largely confined to text-only LLMs; our work extends them to the visual modality via controlled synthetic pairs and further proposes HeadLens for lightweight head-level semantic decoding.

## 2 Problem Formulation

To validate the intrinsic counting capabilities of LVLMs, we employ synthetic datasets that isolate numerical reasoning from confounding variables such as background clutter and complex textures. As in Fig. 1, we use PIL [8] generated **SynDot** and **SynPoly**, comprising randomly distributed black dots or multi-colored polygons on a white canvas; and **SynReal**, which uses FLUX.1-dev [18] to generate photorealistic objects across six categories. Object counts range in $[1, 10]$; details are in Sec. C. We evaluate counting performance on four axes : **Accuracy (ACC)** for exact-match precision; **MAE** and **RMSE** for deviation magnitude and stability; and **Off-by-one Accuracy (OBO)** for near-miss reliability.

**Table 1:** Performance comparison on synthetic benchmarks.

| Model | Dataset | Acc ↑ | MAE ↓ | RMSE ↓ | OBO ↑ |
|---|---|---|---|---|---|
| Qwen2.5-VL-7B | SynDot | 76.12 | 0.25 | 0.52 | 96.21 |
| | SynPoly | 53.74 | 0.94 | 1.75 | 82.27 |
| | SynReal | 73.48 | 0.79 | 2.13 | 91.01 |
| Qwen3-VL-8B | SynDot | 73.22 | 0.32 | 0.66 | 95.21 |
| | SynPoly | 70.34 | 0.52 | 1.49 | 94.02 |
| | SynReal | 88.79 | 0.15 | 0.47 | 98.88 |
| LLaVA-1.5-7B | SynDot | 24.07 | 4.46 | 6.58 | 39.22 |
| | SynPoly | 33.30 | 1.65 | 2.34 | 56.71 |
| | SynReal | 54.27 | 3.59 | 5.05 | 66.34 |

We evaluate the counting performance of three popular LVLMs, with the prompt "*What is the number of the {object} in the image? Answer with number only.*". As shown in Tab. 1, we surprisingly find that these models fail significantly across all three benchmarks compared to humans, who can achieve flawless accuracy. A more counterintuitive observation is that LVLMs perform better on the more realistic SynReal, while struggling with the visually simpler SynDot and SynPoly. This discrepancy clearly exposes an inherent bias rooted in their training data distribution. This naturally raises a fundamental question: **Does the counting capability of LVLMs signify an emergent reasoning mechanism, or is it merely a statistical hallucination predicated on correlational visual cues and language priors?**

## 3 Do LVLMs know how to count? Or do they memorize?

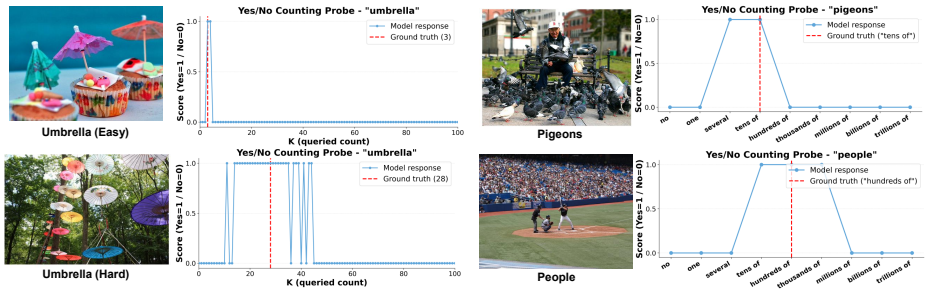



**Fig. 2:** Counting Uncertainty Curve by Yes/No Answer of Qwen2.5 VL 7B.

In this section, we attempt to address the problem proposed in the initial evaluation. We observe that models often hesitate to provide precise numerical counts when faced with occlusion, blurriness, or high object density. To circumvent this, we reformulate the counting task into a binary 'Yes/No' verification framework. Specifically, for a given image, we employ the prompt: "*There are <K> <Object Name> in the image, Yes or No?*". For images with countable instances, we iteratively vary $K$ from 0 to 100. For complex scenes (e.g., swarms of pigeons or dense crowds), we substitute $K$ with quantitative descriptors such as 'tens of' or 'hundreds of'. We assign scores of 1 for 'Yes', 0 for 'No'.

As shown in Fig. 2, counting uncertainty strongly correlates with the stability of the model's "Yes Band" (a contiguous interval of positive responses). Simple scenarios (e.g., few, unoccluded umbrellas) yield a concentrated, stable Yes Band, whereas challenging cases blur decision boundaries, causing the band to expand and oscillate. Nonetheless, LVLMs retain robust coarse-grained estimation, reliably distinguishing orders of magnitude (e.g., "tens" vs. "millions" of pigeons) despite lacking precision. These properties validate that: *LVLMs possess a visual subitizing and estimation capability similar to human behavior [6]. Instead of relying on discrete tokens, the model encodes quantity within a continuous latent space, enabling seamless cross-modal alignment between visual stimuli and linguistic quantifiers.*

**Cognitive Alignment from the Representation Aspect.** To further verify the counting ability alignment between LVLMs and human cognition, we project the hidden states of the SynDot dataset into a 2D space using t-SNE (Fig. 3). As features propagate from the first layer to the final layer into deeper layers, *class 0* (background) and *class 1 - 4* (black dots number) form distinct, isolated clusters separated from other numbers. This topological isolation provides a mechanistic explanation for the model's subitizing capability, the accurate recognition of small quantities. However, as the object count increases, the corresponding clusters become progressively entangled and heavily overlapped. This spatial

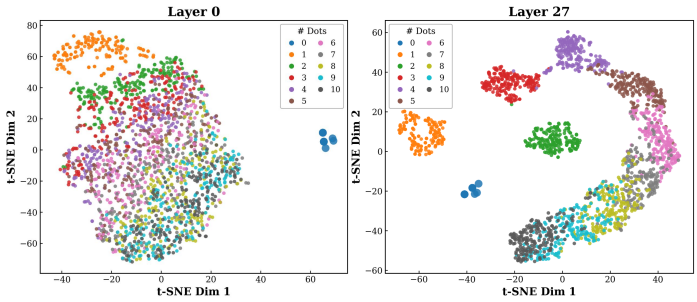



**Fig. 3:** T-SNE visualization of model hidden states based on black dots data.

compression in the latent space directly accounts for the performance degradation on larger numbers, reflecting a natural cognitive shift from precise subitizing to approximate magnitude estimation.

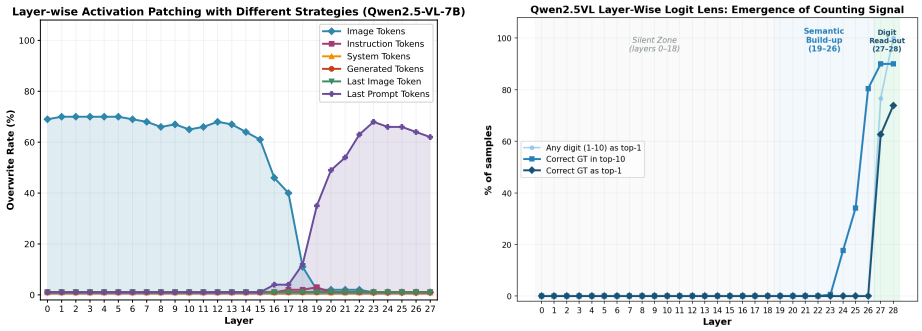
## 4 Understand Layerwise Behavior of Counting

In this section, we investigate the underlying mechanisms that execute visual counting tasks inside LVLMs. We choose Qwen2.5-VL-7B as the default model for our interpretability studies due to its balance between architectural simplicity and good counting performance.

### 4.1 Information Flow and Cross-Modal Routing

To trace how counting information propagates from visual input to numerical output, we adapt activation patching [14, 40], a technique traditionally applied to text-only LLMs [11, 38], for LVLM analysis. By fixing the random seed in our synthetic datasets, we construct tightly controlled image pairs. Each pair consists of a clean image (e.g., three dots) and a corrupted counterpart with an altered object count (e.g., five dots) that strictly preserves the original spatial layout. This minimal perturbation extends activation patching to the visual modality, allowing us to precisely isolate how the model's internal states respond to quantitative changes.

**Visual Activation Patching.** In an $L$-layer LVLM, the hidden state $s_i^l$ encodes token $x_i$ after being processed through layers 1 to $l$. We first record hidden states from a forward pass on the corrupted image, then re-run inference on the clean image while replacing $s_i^l$ of a target token set at layer $l$ with its corrupted counterpart. Whether the model's prediction flips from the clean to the corrupted label gives the *overwrite rate*, measuring the causal efficacy of the patched tokens. We partition input tokens into six groups: System Prompt, Image Tokens, User Instruction, Generated Tokens, Last Image Token ("|img_end|"), and Last Prompt Token (the "Assistant:" role tag), and use the prompt "*What is the number of the black dots in the image? Answer with the number only.*" so that the first generated token is the counting answer.




**(a)** **(b)**

**Fig. 4:** Left: Layer-wise Overwrite Rate of Different Input Tokens Patching Strategy; Right: Layer-wise Logit Lens Tracking Curve for Counting Number.

We run layer-wise activation patching on 100 black dot image pairs. As shown in Fig. 4a, in early-to-middle layers ($L < 15$), counting information is predominantly anchored in *All Image Tokens* with high, stable causal influence. From Layer 15 onward, image token importance drops sharply while the *Last Prompt Token* ("Assistant:" tag) surges, peaking at Layer 23, revealing a clear "handover" mechanism. Layers 15–22 thus form the critical bottleneck for cross-modal routing, where spatially distributed visual features are compressed into the linguistic stream. The negligible impact of user instruction and system prompt tokens confirms that the numerical signal flows directly from visual representations to the response-triggering tokens.

To corroborate the emergence of numerical representations subsequent to the cross-modal routing phase, we employ a logit lens on the hidden state $s^l_{-1}$ immediately preceding the generation of the final counting token. By projecting these intermediate representations through the model's terminal unembedding layer, we map $s^l_{-1}$ directly into the vocabulary space to decode its latent semantic trajectory. As depicted in the Fig. 4b, correct number tokens begin to surface within the top-10 logit predictions between layers 19 and 26, gaining pronounced significance from layer 23 onwards. By layers 27 and 28, the ground-truth counting token converges to the rank-1 position.

This layer-wise progression captures the internal patterns of how the model aggregates visual numerosity, translates it into a discrete linguistic space, and ultimately solidifies the final numerical output. To further investigate the mechanistic underpinnings of these patterns, the subsequent section aims to pinpoint the specific attention heads responsible for executing the counting task and characterize their precise functionalities.

## 5 Mechanistic Analysis on Attention Head Function

Isolating complete end-to-end circuits in LVLMs is intractable due to their massive scale and redundant backup heads [35]. Therefore, we focus on identifying and revealing the critical attention heads that dominate the counting task.

### 5.1 Important Attention Heads for Counting

We conduct head-level visual activation patching to identify the important attention heads for counting tasks. Specifically, we leverage 100 SynDot image pairs as the layerwise VAP but replace the target head activation. To prevent confounding effects from the residual stream, we define head activation strictly as the output of the projection layer. The importance score of each head is defined as the answer token logit difference change before and after the VAP. A positive importance score greater than 0.05 indicates that the target head plays a significant role in the counting task. There are $\frac{43}{784} = 5.5\%$ heads noted as important heads. We visualize the heads with positive importance score as a heatmap in Fig. 5 (left). Next, we introduce a new interpretability tool named HeadLens and use it to reveal the functions of representative important heads based on their attention distribution and the semantic meaning of their output.

### 5.2 HeadLens: Decoding Individual Attention Heads

To isolate the semantic contribution of individual attention heads, we introduce a novel interpretability tool termed **HeadLens**. Utilizing the linear additivity of the multi-head self-attention (MHSA) projection, HeadLens enables granular, token-level semantic decoding without structural modifications.

Let $h_i(x) \in \mathbb{R}^{d_{\text{head}}}$ denote the output of the $i$-th attention head, where $H$ is the total number of heads and $d_{\text{model}} = H \times d_{\text{head}}$. We first isolate the $i$-th head's contribution by constructing a zero-padded representation $\tilde{h}_i(x) \in \mathbb{R}^{d_{\text{model}}}$:

$$\tilde{h}_i(x) = [\ \underbrace{0, \ldots, 0}_{(i-1)\times d_{\text{head}}}\ ,\ h_i(x),\ \underbrace{0, \ldots, 0}_{(H-i)\times d_{\text{head}}}\ ] \tag{1}$$

Due to the properties of block matrices, the standard concatenated output of the MHSA mechanism can be equivalently formulated as the sum of these sparse representations. The final output of MHSA $\hat{x}$ is then obtained by applying the output projector matrix $W_O$:

$$\hat{x} = \left( \sum_{i=1}^{H} \tilde{h}_i(x) \right) W_O = \sum_{i=1}^{H} \left( \tilde{h}_i(x) W_O \right) \tag{2}$$

Rather than interpreting the aggregated hidden state $\hat{x}$, HeadLens operates directly on the isolated projection $\tilde{h}_i(x)W_O$. To translate this independent vector into human-interpretable concepts, we employ the learned affine translator [4] $T : \mathbb{R}^{d_{\text{model}}} \to \mathbb{R}^{d_{\text{model}}}$ (e.g., $T(z) = Az+b$) to map it into the final residual stream

space. Utilizing the model's unembedding matrix $U \in \mathbb{R}^{|\mathcal{V}| \times d_{\text{model}}}$ and bias $c$, we formulate the head-specific residual transformation $r_i(x)$, its corresponding logits $\ell_i(x)$, and the token probability distribution $p_i(\cdot \mid x)$ as follows:

$$r_i(x) = T\big(\tilde{h}_i(x) W_O\big) \qquad \ell_i(x) = U\, r_i(x) + c \qquad p_i(\cdot \mid x) = \text{softmax}\big(\ell_i(x)\big)$$

By treating each $\tilde{h}_i(x) W_O$ as an independent semantic component, HeadLens effectively decodes the head activation into semantic tokens. We are particularly interested in tokens related to the visual features (e.g., color and shape) and counting (e.g., digits and numbers). We define the ratio of visual feature tokens and counting tokens in the top-10 HeadLens results as the Visual Grounding Score (VGS) and Counting Token Emergence Rate (CTER), respectively.

### 5.3 Revealing Attention Head Functionalities

In this section, we reveal the functionalities of four representative attention heads, including two counting heads (Cross-modal Routing Heads and Counting Aggregation Heads) and two special functional heads (Visual Grounding Heads and Awareness Heads) for the counting task. We provide the experiment details of head discovery in Sec. E.1.

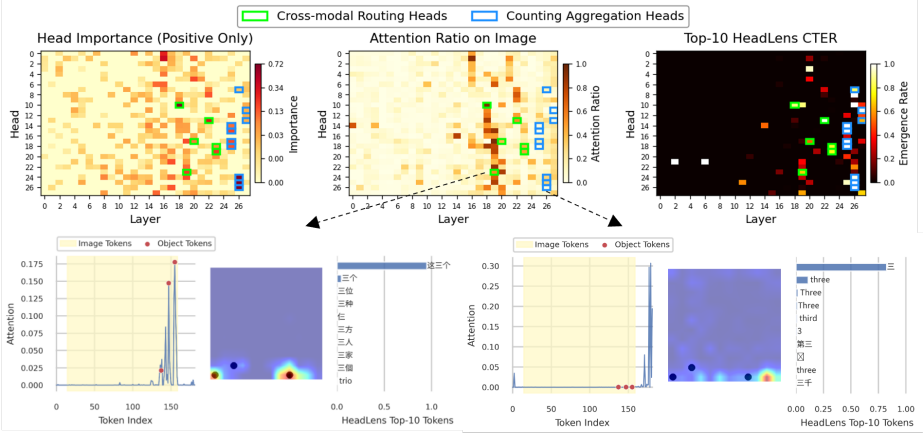



**Fig. 5: Visualization of two functional groups of counting heads**. Left to right: Head importance (positive-only), attention ratio on image tokens, and top-10 HeadLens CTER across heads. We present the typical attention distribution and top-10 HeadLens results for Cross-modal Routing Heads (green boxes) and Counting Aggregation Heads (blue boxes), using L19H23(left) and L26H26(right). Both have "three" (三) in Chinese as the top-1 HeadLens token.

**Counting Heads.** We discuss two functionally distinct attention head groups for the counting task, Cross-modal Routing Heads and Counting Aggregation

Heads. **Cross-modal Routing Heads** are defined as heads with high attention ratio on image, high head importance, and high top-10 HeadLens CTER. These heads directly extract information from image tokens and transfer visual information into the language related to the counting task. As shown in Fig. 5, these heads mainly lie in Layers 18–24 noted with green boxes. We visualize L19H23 as a typical example. It attends to object-related image regions and top-10 HeadLens tokens are relevant to number (e.g., three), further confirming its role in bridging image and text. **Counting Aggregation Heads** are defined as heads with low image attention, high importance, and high top-10 HeadLens CTER. Despite extracting information from image tokens, these heads aggregate counting-relevant information already deposited in the residual stream by earlier layers rather than extracting visual information directly. We visualize L26H26, and its top-10 decoded token exhibit a remarkably high alignment with the model's final counting prediction while paying little attention to the image, confirming their role in counting aggregation.

**Visual Grounding Heads.** Concentrated primarily in the early stages of the network, specifically within Layer 1, these attention heads are dedicated to fundamental visual processing. They predominantly attend to object-specific image regions, serving as the primary mechanism for extracting basic, low-level visual attributes such as color, boundaries, and shape. We collect the HeadLens top-10 tokens of each head and compute the VGS accordingly. As shown in Fig. 6, most visual grounding heads are located at layer 1, which has the highest average visual grounding score.

**Existence and difficulty Awareness Heads.** We also identify two deep-layer attention heads that show awareness of object existence and the difficulty estimation of the counting.**L26H8** acts as a universal *existence detector*, outputting "1" whenever any object is present. **L23H19** is a difficulty-aware head. L23H19 is one of the best counting aggregation heads (32.9% top-1 accuracy). Besides number tokens, its top-10 tokens are mixed with "no" and "unnecessary" for $GT = 1$, indicating counting 1 dot is trivially easy. For $GT \geq 2$, the head switches to digit predictions with reasonable accuracy, and for higher counts begins mixing in Chinese difficulty-related tokens like difficult and impossible, alongside the numeric predictions.

## 6 Enhancing LVLMs' counting ability

Our findings reveal a step-by-step counting mechanism: early layers extract visual features, middle layers route information into a semantic space, and later layers aggregate the answer. Guided by these findings, we introduce two targeted interventions.

### 6.1 Object-Focused Attention Regularizer

To improve the grounding of counting, we supervise the model's visual attention to be more "object-centric" using SynDot and SynPoly. Each image naturally

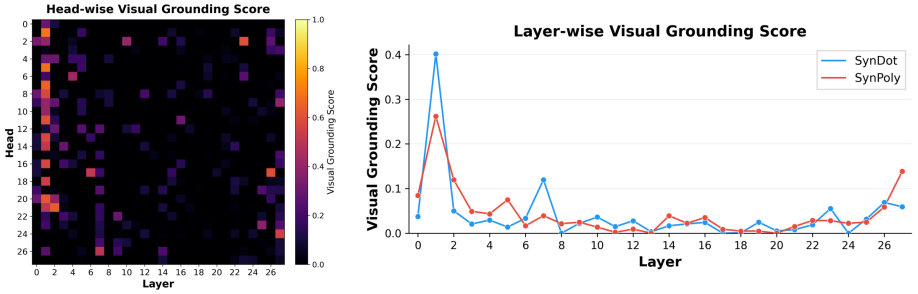



**Fig. 6:** Visual Grounding Heads in Early Layers. Left: Head-wise heatmap based on Visual Grounding Score; Right: Layer-wise Visual Grounding Score.

provides object centers $\{c_k\}_{k=1}^{N}$, which we convert into a *soft instance prior* on the $H \times W$ patch grid. We define an unnormalized grid score $u(p)$ at patch $p$ using a Gaussian kernel:$u(p) = \sum_{k=1}^{N} \exp\left(-\frac{\|p-\pi(c_k)\|_2^2}{2\sigma^2}\right)$, where $\pi(\cdot)$ maps pixels to patch coordinates and $\sigma = 1$ controls the supervision spread. We normalize $u(p)$ to obtain a target distribution $g \in \Delta^{|\mathcal{V}|-1}$ over the visual tokens $\mathcal{V}$.

We encourage the attention heads to allocate their weights on each image token $t \in \mathcal{T}$ to align with the instance prior $g$. For each layer $l$, we compute the average attention weights across all $H$ heads and renormalize it over $\mathcal{V}$ to obtain the predicted distribution $q_t^l$:

$$q_t^l(j) = \frac{\sum_{h=1}^{H} a_h^l(t,j)}{\sum_{j' \in \mathcal{V}} \sum_{h=1}^{H} a_h^l(t,j')}, \quad j \in \mathcal{V}. \tag{3}$$

The focus loss is defined as the cross-entropy (equivalent to KL divergence) between $g$ and $q_t^l$, where $\epsilon$ is a small number for stability:

$$\mathcal{L}_{\text{focus}} = \frac{1}{|\mathcal{L}||\mathcal{T}|} \sum_{l \in \mathcal{L}} \sum_{t \in \mathcal{T}} \left( -\sum_{j \in \mathcal{V}} g(j) \, \log\big(q_t^l(j) + \varepsilon\big) \right). \tag{4}$$

Guided by our circuit discovery (Sec. 5), we explore applying attention supervision to the visual grounding heads at early layers and cross-modal routing heads with a high image attention ratio at late layers to verify if the model's counting mechanism can be precisely localized and regularized.

### 6.2 Adaptive Head Temperature Tuning

Our analysis reveals that cross-modal routing and counting aggregation heads exhibit highly targeted attention patterns. To amplify this intrinsic circuitry, we introduce Adaptive Head Temperature Tuning, which reduces the attention entropy of these critical heads to sharpen their focus on relevant tokens.

Instead of applying uniform temperature scaling, we adaptively modulate the pre-softmax logits of the identified target heads. For a given head $h$, we define an inverse temperature multiplier $\beta_h = \alpha \times \gamma_h$, where $\alpha \geq 1$ provides a baseline entropy reduction, and $\gamma_h \geq 0$ is the head's intrinsic importance score derived from our circuit analysis. The modified attention matrix is thus computed as:

$$A_h = \text{softmax}\left(\beta_h \frac{Q_h K_h^T}{\sqrt{d_k}}\right) \tag{5}$$

Applying this training-free technique exclusively to the routing and aggregation heads enhances the signal-to-noise ratio of the counting information flow.

### 6.3 Joint Optimization Objective

To integrate our targeted interventions, we combine standard Supervised Fine-Tuning (SFT) with our focus regularizer. The SFT process exclusively utilizes the SynDot and SynPoly datasets, restricting the autoregressive loss calculation strictly to the target counting tokens. The overall training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{SFT}} + \lambda \mathcal{L}_{\text{focus}}, \tag{6}$$

where $\lambda$ controls the regularization strength (set to 1 by default).

## 7 Experiments

**Model and Data.** We evaluate the proposed method on three different LVLMs: Qwen2.5-VL-7B, Qwen3-VL-8B and LLaVA-1.5-7B. The training data comprises 8000 synthetic images equally from SynDot and SynPoly (classes 1–10, evenly per class). For OOD counting, we evaluate on SynReal and PixMo-Count [9]. For visual reasoning, we use MMMU [39], RealWorldQA [37], and MathVista [23].
**Training.** We optimize with the joint loss ( Eq. (6)). For SFT, we fine-tune models using LoRA ($r = 64$) on the attention projection of all layers. We apply an attention regularizer to layers 2, 18-22 for Qwen2.5-VL-7B, layers 17-19 for Qwen3-VL-8B, and layers 0, 14 for LLaVA-1.5-7B, which have the most visual grounding heads and cross-modal routing heads. All models are trained for 2 epochs on a single NVIDIA H200 GPU. We use AdamW, BF16 precision, a batch size of 2, and a learning rate of $2 \times 10^{-5}$ with linear decay and 3% warmup. We take the mean results with 3 random seeds. For head tuning, we use a global value $\alpha = 1.2$ across different models.

### 7.1 Evaluation Results

**Counting Evaluation.** Our method consistently improves OOD counting across all three backbones on both SynReal and PixMo-Count. As shown in Table 2, the gains are consistent across all metrics, including higher accuracy and OBO,

| Backbone | Method | SynReal | | | | PixMo-Count | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Acc | MAE↓ | RMSE↓ | OBO↑ | Acc | MAE↓ | RMSE↓ | OBO↑ |
| Qwen2.5-VL-7B | Baseline | 73.48 | 0.79 | 2.13 | 91.01 | 58.79 | 0.84 | 1.79 | 84.00 |
| | Ours | **84.83** | **0.74** | **1.47** | **97.75** | **64.15** | **0.62** | **1.46** | **87.10** |
| Qwen3-VL-8B | Baseline | 88.79 | 0.15 | 0.47 | 98.88 | 58.75 | 0.73 | 1.35 | 88.20 |
| | Ours | **91.21** | **0.11** | **0.34** | **99.41** | **66.98** | **0.60** | **1.25** | **91.65** |
| LLaVA-1.5-7B | Baseline | 54.27 | 3.59 | 5.05 | 66.34 | 31.88 | 1.81 | 3.09 | 59.77 |
| | Ours | **60.11** | **1.56** | **2.08** | **91.01** | **32.64** | **1.61** | **2.59** | **62.43** |

**Table 2:** Evaluation on Out-of-distribution Counting Benchmarks.

| Backbone | Method | MMMU | RealWorldQA | MathVista | $\Delta$ (avg.) |
|---|---|---|---|---|---|
| Qwen2.5-VL-7B | Baseline | 54.89 ± 0.00 | 61.96 ± 0.00 | 57.30 ± 0.00 | – |
| | Ours | **56.33** ± 0.13 | **64.14** ± 0.26 | **58.30** ± 0.15 | +1.54 |
| Qwen3-VL-8B | Baseline | 58.33 ± 0.00 | 70.05 ± 0.00 | 66.30 ± 0.00 | – |
| | Ours | **60.74** ± 0.17 | **71.83** ± 0.15 | **67.90** ± 0.19 | +1.93 |
| LLaVA-1.5-7B | Baseline | 44.44 ± 0.00 | 56.21 ± 0.00 | 23.90 ± 0.00 | – |
| | Ours | **44.56** ± 0.18 | **56.48** ± 0.23 | **26.30** ± 0.31 | +0.93 |

**Table 3:** Evaluation on general capability benchmarks. Results are reported as mean ± standard deviation.

as well as lower MAE and RMSE. Notably, Qwen3-VL-8B, despite already being strong on SynReal, still shows clear gains, and LLaVA-1.5-7B also benefits substantially. These results show that our method transfers effectively to OOD counting rather than only fitting the training distribution.

**General Capabilities Evaluation.** Our method consistently improves general visual capabilities across three backbones, though trained on the counting task only. As shown in Table 3, all models achieve positive gains on MMMU, RealWorldQA, and MathVista, suggesting our intervention does not merely improve counting in isolation. Instead, it enhances the counting-related circuit which transfers to broader visual capabilities, indicating that counting may serve as a useful primitive for general visual reasoning. Notably, this improvement cannot be simply attributed to insufficient counting-related pretraining in the backbone. Qwen3-VL [2] emphasized stronger visual grounding and reasoning in its model and training design than Qwen2.5-VL [3], especially with additional counting training data.

### 7.2 Ablation Study

We decompose our full method into its individual components: LoRA-based SFT, Object-Focused Attention Regularizer ($\mathcal{L}_{\text{focus}}$), and Adaptive Head Temperature Tuning ($\beta_h$). As shown in Tab. 4, each component provides complementary gains.

**Table 4:** Component ablation on Qwen2.5-VL-7B across specialized and general benchmarks.

| SFT | $\mathcal{L}_{\text{focus}}$ | $\beta_h$ | Synreal | PixMo-Count | MMMU | RealWorldQA | MathVista |
|---|---|---|---|---|---|---|---|
| Baseline | | | 73.48 | 58.79 | 54.89 | 61.96 | 57.30 |
| ✓ | | | 79.78 | 61.67 | 56.00 | 63.14 | 58.10 |
| ✓ | ✓ | | 82.34 | 63.32 | 56.11 | 64.05 | 58.20 |
| ✓ | ✓ | ✓ | **84.83** | **64.15** | **56.33** | **64.14** | **58.30** |

### 7.3 Mechanistic Overlap: Why Does Counting Enhancement Generalize?

Enhancing counting performance via our synthetic datasets (SynDot and SynPoly) unexpectedly improved OOD performance on broader tasks, including general VQA and mathematical reasoning. To investigate the origins of this transferability, we analyze five representative visual tasks: (1) **Counting** (SynDot and Pixmo-Count), (2) **Visual Attribution** (a custom color and shape diagnostic), (3) **Spatial Relations** (CLEVR [16]), (4) **General VQA**, and (5) **Math Reasoning** (MathVista).

To identify the underlying circuit mechanisms, we perform mean ablation at the head level. Specifically, we substitute each head's activation with its dataset-level mean and identify the Top-20 important heads based on the resulting logit degradation. We then use the Jaccard similarity $J$ to quantify the mechanistic overlap across tasks. We use $HS_A$ and $HS_B$ to denote two arbitrary important head sets.

$$J(HS_A, HS_B) = \frac{|HS_A \cap HS_B|}{|HS_A \cup HS_B|} \tag{7}$$

As shown in Fig. 7, *Attribution Color* exhibits minimal overlap with other tasks, indicating reliance on pure perception. In contrast, *Counting* shares $> 33\%$ of its critical heads with reasoning-heavy tasks, explaining the OOD transferability. The high similarity between synthetic and real-world counting further confirms that our method targets fundamental enumeration circuits rather than dataset-specific artifacts. In comparing, applying the same intervention to visual attribution (color/shape recognition) yields no OOD reasoning gains , reinforcing counting's critical role.

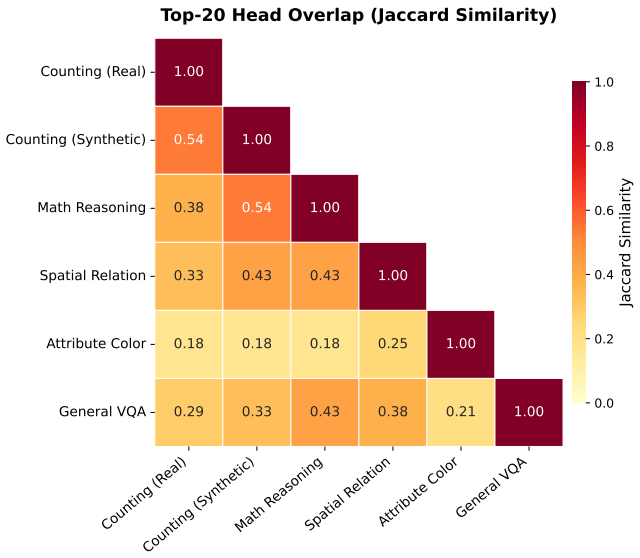



**Fig. 7:** Jaccard Similarity for the top-20 most important heads of different tasks.

## 8 Conclusion

In this work, we show that counting is a meaningful probe of visual reasoning in LVLMs. Through controlled benchmarks and mechanistic analysis, we identify structured counting-related circuits and demonstrate that improving this primitive skill with lightweight synthetic-data intervention can yield gains beyond counting itself. These findings suggest that counting is not merely a narrow task, but an important building block of broader visual reasoning. Looking forward, an important direction is to examine whether these insights generalize to larger LVLMs and to other primitive visual skills, such as spatial reasoning and attribute comparison, toward a unified understanding of visual reasoning.

## 9 Acknowledgment

We are deeply grateful to Prof. Michelle Hurst and Prof. Jacob Feldman from Rutgers University for their substantial intellectual contributions to this work. Their insights from the perspective of cognitive science were instrumental in shaping the conceptual framing and interpretation of our findings. We also thank them for their thoughtful discussions and careful review of the manuscript.

## References

1. Alghisi, S., Roccabruna, G., Rizzoli, M., Mousavi, S.M., Riccardi, G.: [de| re] constructing vlms' reasoning in counting. arXiv preprint arXiv:2510.19555 (2025)

2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Belrose, N., Furman, Z., Smith, L., Halawi, D., Ostrovsky, I., McKinney, L., Biderman, S., Steinhardt, J.: Eliciting latent predictions from transformers with the tuned lens. arXiv preprint arXiv:2303.08112 (2023)
5. Che, L., Liu, T.Q., Jia, J., Qin, W., Tang, R., Pavlovic, V.: Hallucinatory image tokens: A training-free eazy approach to detecting and mitigating object hallucinations in lvlms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 21635–21644 (October 2025)
6. Cheyette, S.J., Piantadosi, S.T.: A unified account of numerosity perception. Nature human behaviour **4**(12), 1265–1272 (2020)
7. Cheyette, S.J., Wu, S., Piantadosi, S.T.: Limited information-processing capacity in vision explains number psychophysics. Psychological Review **131**(4), 891 (2024)
8. Clark, A., Contributors: Pillow (pil fork) documentation (2015), https://pillow.readthedocs.io, accessed on <date of access>
9. Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J.S., Salehi, M., Muennighoff, N., Lo, K., Soldaini, L., et al.: Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 91–104 (2025)
10. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
11. Dumas, C., Wendler, C., Veselovsky, V., Monea, G., West, R.: Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 31822–31841 (2025)
12. Guo, X., Huang, Z., Shi, Z., Song, Z., Zhang, J.: Your vision-language model can't even count to 20: Exposing the failures of vlms in compositional counting. arXiv preprint arXiv:2510.04401 (2025)
13. Hasani, H., Izadi, A., Askari, F., Bagherian, M., Mohammadian, S., Izadi, M., Baghshah, M.S.: Understanding counting mechanisms in large language and vision-language models. arXiv preprint arXiv:2511.17699 (2025)
14. Heimersheim, S., Nanda, N.: How to use and interpret activation patching. arXiv preprint arXiv:2404.15255 (2024)
15. Huang, Z., Dai, M., Zhang, Y., Zhang, J., Shan, H.: Point segment and count: A generalized framework for object counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17067–17076 (June 2024)
16. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: 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)

17. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
18. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025), https://arxiv.org/abs/2506.15742
19. Li, Y., Salehi, S., Ungar, L., Kording, K.P.: Does object binding naturally emerge in large pretrained vision transformers? arXiv preprint arXiv:2510.24709 (2025)
20. Lieder, F., Griffiths, T.L.: Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. Behavioral and Brain Sciences **43**, e1 (2020). https://doi.org/10.1017/S0140525X1900061X, pMID: 30714890
21. Liu, B., Kamath, A., Grunde-McLaughlin, M., Han, W., Krishna, R.: Visual representations inside the language model. ArXiv **abs/2510.04819** (2025), https://api.semanticscholar.org/CorpusID:281843173
22. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
23. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255 (2023)
24. Meng, K., Bau, D., Andonian, A., Belinkov, Y.: Locating and editing factual associations in gpt. Advances in neural information processing systems **35**, 17359–17372 (2022)
25. Minderer, M., Gritsenko, A., Houlsby, N.: Scaling open-vocabulary object detection. Advances in Neural Information Processing Systems **36**, 72983–73007 (2023)
26. Neo, C., Ong, L., Torr, P., Geva, M., Krueger, D., Barez, F.: Towards interpreting visual information processing in vision-language models. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=chanJGoa7f
27. nostalgebraist: Interpreting gpt: The logit lens (August 2020), https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, accessed: 2024-12-21
28. Paiss, R., Ephrat, A., Tov, O., Zada, S., Mosseri, I., Irani, M., Dekel, T.: Teaching clip to count to ten. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3170–3180 (2023)
29. Qharabagh, M.F., Ghofrani, M., Fountoulakis, K.: Lvlm-count: Enhancing the counting ability of large vision-language models. arXiv preprint arXiv:2412.00686 (2024)
30. Sakarvadia, M., Khan, A., Ajith, A., Grzenda, D., Hudson, N., Bauer, A., Chard, K., Foster, I.: Attention lens: A tool for mechanistically interpreting the attention head information retrieval mechanism. arXiv preprint arXiv:2310.16270 (2023)
31. Sengupta, S., Moradinasab, N., Liu, J., Brown, D.E.: Can vision-language models count? a synthetic benchmark and analysis of attention-based interventions. arXiv preprint arXiv:2511.17722 (2025)
32. Treisman, A.: Feature binding, attention and object perception. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences **353**(1373), 1295–1306 (1998)

33. Đukić, N., Lukežič, A., Zavrtanik, V., Kristan, M.: A low-shot object counting network with iterative prototype adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 18872–18881 (October 2023)
34. Wang, K., Pan, J., Shi, W., Lu, Z., Ren, H., Zhou, A., Zhan, M., Li, H.: Measuring multimodal mathematical reasoning with math-vision dataset. Advances in Neural Information Processing Systems **37**, 95095–95169 (2024)
35. Wang, K., Variengien, A., Conmy, A., Shlegeris, B., Steinhardt, J.: Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. arXiv preprint arXiv:2211.00593 (2022)
36. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
37. xAI: Grok-1.5 vision preview. `https://x.ai/news/grok-1.5v` (April 2024), accessed: Dec 2025
38. Yeo, W.J., Satapathy, R., Cambria, E.: Towards faithful natural language explanations: A study using activation patching in large language models. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 10436–10458 (2025)
39. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9556–9567 (2024)
40. Zhang, F., Nanda, N.: Towards best practices of activation patching in language models: Metrics and methods. arXiv preprint arXiv:2309.16042 (2023)