The Halo Effect and Language Takeover: Spatiotemporal Attention Decay Explains Vision-Language Model Failures in Simple Visual Counting

Anonymous ACL submission

Abstract

Despite their remarkable capabilities in complex multimodal reasoning, Vision Language Models (VLMs) exhibit a perplexing inability to perform elementary visual counting tasks reliably. Existing hypotheses, often centering on input resolution or patch tokenization, fail to fully explain the stochastic nature of these errors, particularly in multi-digit generation. In this work, we investigate the internal decision-making dynamics of VLMs (e.g., Qwen3-VL, Gemma3) through the lens of attention mechanisms. By leveraging a controlled synthetic dataset and introducing novel metrics for Visual *Sparsity* and *Entropy*, we discover a novel phenomenon: **Spatiotemporal Attention Decay**. Our analysis reveals two distinct failure modes. Spatially, models exhibit a **Halo Effect**, where attention focuses on the peripheral convex hull of object clusters rather than penetrating the geometric centers of individual instances. Temporally, we observe a phenomenon of **Language Takeover**: during auto-regressive decoding, visual grounding decays rapidly after the initial token. Quantitative analysis confirms that as attention sparsity drops and entropy rises, the generation of subsequent digits degenerates from visual perception into hallucination driven by language priors. These findings suggest that counting failures stem from the model's inability to maintain spatiotemporal focus, highlighting the need for mechanisms that enforce persistent visual grounding.

1 Introduction

Vision Language Models (VLMs) have demonstrated exceptional proficiency in complex multimodal reasoning. Paradoxically, these systems frequently stumble on elementary visual counting tasks—a fundamental capability that serves as a litmus test for the fidelity of visual grounding. Despite their ability to describe intricate scenes, models like Qwen2-VL and Gemma often fail to enumerate simple objects correctly, exhibiting errors that appear stochastic and resistant to standard instruction tuning.

Recent scholarship have begun to investigate this numeracy gap, some studies have started to probe the attention mechanisms underlying counting (Sengupta et al., 2025; Behrens et al., 2025). However, these works largely treat attention as a *static* feature map, focusing on where the model looks generally. They overlook the critical *temporal dynamics* of generation, particularly in multi-digit scenarios where the model must maintain visual focus across multiple decoding steps. We argue that the root cause of counting failures is not merely *spatial* misalignment, but the **decoupling of visual grounding from the temporal generation process**.

By dissecting the decision-making process of SOTA open-weights models on controlled synthetic datasets, we identify a dual failure mode that bridges the gap between static artifacts and dynamic generation. We term the underlying phenomenon responsible for this failure **Spatiotemporal Attention Decay**. Spatially, we observe the **Halo Effect** by designing an attention extraction mechanism. As shown in Figure 1, visual attention consistently fails to penetrate the geometric centers of objects. Instead, it aggregates along the peripheral convex hull of object clusters (resembling "register tokens" (Darcet et al., 2024)), suggesting the model relies on heuristic area estimation rather than precise instance enumeration. Temporally, we identify the **Language Takeover** phenomenon by quantifying a rapid decay in attention sparsity as well as growth in entropy during auto-regressive decoding. As the sequence progresses from the tens to units digit, the model's reliance on visual tokens diminishes, and generation becomes dominated by language priors—effectively transitioning the model from a "viewer" to a "guesser." By uncovering this decay, researchers can better understand the mechanistic origins of model hallucinations,

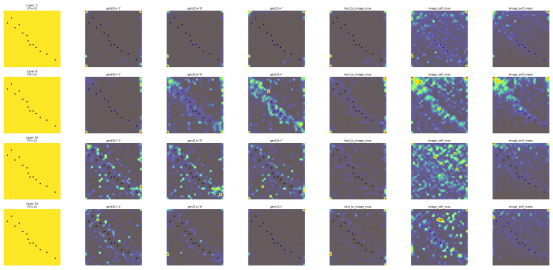

Figure 1: **Visualizing Spatiotemporal Attention Decay.** Cross-modal attention heatmaps for Qwen3-VL-2B (GT: 12 → Pred: 10). The top rows (Layer -1, 8, 16, 24) illustrate the **Halo Effect**, where the model focuses on the edges of the dot distribution rather than individual instances. Crucially, compare the attention during the generation of the first token "1" versus the second token "0": the visual focus becomes increasingly diffuse or misaligned in later steps. This **Spatiotemporal Decay** decoupling vision from generation explains the model's systematic under-counting bias.

enabling the development of targeted interventions to maintain visual fidelity throughout generation.

In summary, our contributions are threefold: (1) We formally define the **Halo Effect**, linking counting failures to attention artifacts; (2) We propose dynamic metrics (Visual Entropy/Sparsity) to quantify the **Language Takeover** phenomenon; (3) We provide empirical evidence that VLMs fail because visual grounding cannot persist through the temporal span of auto-regressive generation.

2 Related Work

Visual Counting and Numeracy in VLMs. Despite general multimodal success, precise counting remains a bottleneck for VLMs. Early works attributed this to instance-level information loss in contrastive objectives (Paiss et al., 2023). Recent studies have shifted to analyzing attention allocation: Sengupta et al. (2025) demonstrated that attention interventions can improve counting, while Behrens et al. (2025) distinguished between "relation-based" and "inventory-based" strategies. However, these works primarily focus on static snapshots or theoretical bounds. We extend this analysis to the *spatiotemporal* domain, characterizing the **Language Takeover** phenomenon where counting strategies degrade into linguistic probability matching during auto-regressive decoding.

Mechanistic Interpretability and Artifacts. Understanding internal ViT representations is crucial for diagnosing failures. Our observed **Halo Effect**—where attention clings to object peripheries—aligns with discoveries of "register tokens" (Darcet et al., 2024) and "visual attention sinks" (Kang et al., 2025), where models allocate disproportionate weight to background or high-norm tokens. We reframe these artifacts as a structural limitation in dense retrieval tasks, where the model registers an object cluster's presence (convex hull) but fails to resolve its internal instances.

Language Priors and Hallucination. Hallucination in VLMs is often driven by *Language Bias*, where learned linguistic correlations override visual evidence (Cao et al., 2025; Li et al., 2025). While prior work identifies this bias conceptually, we provide a mechanistic explanation via **Spatiotemporal Attention Decay**. We quantify the precise moment the model abandons visual evidence for language priors, effectively transitioning the generation process from perceptual counting to probabilistic guessing.

3 Methodology

3.1 Controlled Probing Dataset

To rigorously diagnose the internal attention dynamics during counting, we constructed a controlled probing dataset designed to isolate geometric perception from background semantic noise.

Image Specifications: The dataset consists of 256 samples. Each sample is a 1024×1024 grayscale image rendered on a clean white background ($I = 255$).

Object Definition: Targets are dark circles (intensity $v \in [0, 20]$) with a fixed radius $r = 10$ pixels. This high-contrast setup minimizes low-level detection errors, forcing the model to focus on the enumeration task.

Count Distribution (The Two-Digit Trap): The number of objects N is sampled uniformly from the interval $[10, 16]$. This range is strategically chosen to require *multi-token generation* (e.g., generating "1" then "2" for "12"), allowing us to analyze the shift in attention dynamics between the tens and units digits.

Spatial Layout: We employ a uniform rejection sampling strategy. Centroids are sampled uniformly across the image plane with a strict constraint of minimum edge-to-edge separation ($d_{min} \geq 2r$), ensuring no occlusion occurs. This guarantees that any counting failure is due to the model's internal mechanism, not visual ambiguity.

3.2 Attention Extraction Mechanism

To diagnose the model's visual grounding during generation, we extract the cross-modal attention maps at each decoding step. Let x_t denote the text token generated at step t, and $V = \{v_1, ..., v_M\}$ be the sequence of image patch tokens. We focus on the attention weights from the current token x_t (serving as the query) to the visual tokens V (serving as keys). For a model with H attention heads at layer L, the aggregated attention map $A_t \in R^M$ is computed by averaging across heads:

$$A_t = \frac{1}{H} \sum_{h=1}^{H} \text{Softmax}\left(\frac{Q_{t,h}^{(L)} (K_{V,h}^{(L)})^\top}{\sqrt{d_k}} \right) \quad (1)$$

where $Q_{t,h}^{(L)}$ is the query vector of token x_t. We specifically analyze the attention maps from the final Transformer layer, as it most directly influences the immediate next-token prediction. The 1D sequence A_t is then reshaped into a 2D grid $A_t^{2D} \in R^{h \times w}$ matching the vision encoder's grid size (e.g., 16×16 for Gemma3) for spatial analysis.

3.3 Visual Sparsity and Entropy

To quantify the "Language Takeover" phenomenon, we propose two metrics to measure the concentration and uncertainty of the visual attention distribution A_t.

Attention Sparsity (Visual Focus). We utilize the Gini Coefficient to measure how focused the model's attention is on specific image regions. For a flattened probability distribution $p = \text{vec}(A_t)$ sorted in ascending order:

$$\text{Sparsity}(A_t) = 1 - \frac{2}{M} \sum_{i=1}^{M} (M + 1 - i) p_i \quad (2)$$

A high Gini coefficient (≈ 1) indicates that attention is highly concentrated (sparse), suggesting strong visual grounding. A low coefficient indicates diffuse attention, implying the model is not focusing on any specific visual feature.

Visual Entropy (Uncertainty). To measure the uncertainty of the visual connection, we compute the Normalized Shannon Entropy:

$$\text{Entropy}(A_t) = -\frac{1}{\log M} \sum_{i=1}^{M} p_i \log p_i \quad (3)$$

We normalize by $\log M$ to ensure the metric is invariant to the varying visual token counts of different architectures (e.g., Qwen3-VL vs. Gemma3). An increase in entropy during generation signals a loss of visual informativeness.

4 Experimental Setup

Models. We evaluate two state-of-the-art open-weights VLM families: **Qwen3-VL** (2B, 4B, 8B) (Bai et al., 2025) and **Gemma3** (4B, 12B) (Team et al., 2025). This selection covers a wide range of parameter scales and distinct vision encoding strategies (e.g., Qwen's dynamic resolution vs. Gemma's fixed patch grid).

Inference Protocol. To eliminate randomness and ensure reproducibility for mechanistic analysis, we employ **Greedy Decoding** (temperature=0) for all experiments. The prompt is standardized as: *"How many dots are there in the image? Answer directly. Answer: "*, forcing the model to output the count immediately.

Implementation. We extract attention weights from the last Transformer layer. For Qwen3-VL, we dynamically resolve the 2D grid size based on the '<|vision_start|>' and '<|vision_end|>' delimiters. For Gemma3, we utilize its fixed 256-token distinct visual representation.

5 Results and Analysis

In this section, we dissect the performance of VLMs on the controlled counting dataset. We structure our analysis in three layers: (1) quantification of error magnitudes and biases; (2) the correlation between visual attention metrics and task performance; and (3) the spatiotemporal mechanisms driving these failures.

5.1 Performance and Error Bias

We first evaluate the baseline performance across the five models. As illustrated in Figure 2, counting accuracy is generally low, with small-scale models (Qwen3-VL-2B, Gemma3-4B) failing almost completely ($< 6\%$), while larger models (Qwen3-VL-4B/8B) reach a plateau around 50%.

Beyond accuracy, the **Mean Bias Error (MBE)** analysis (Figure 2) reveals distinct failure modes:

- **Systematic Under-counting:** Most models (Gemma3-4B/12B, Qwen3-VL-2B/8B) exhibit a negative bias (MBE < 0). This aligns with the *Halo Effect* hypothesis, where multiple adjacent points are perceptually merged into a single "blob" within the attention halo, leading to under-estimation.

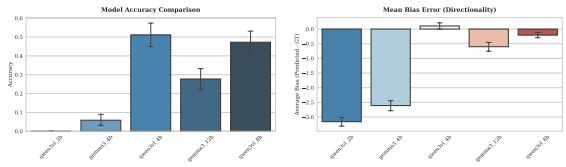

Figure 2: **Left:** Counting accuracy across models. **Right:** Mean Bias Error (MBE). Most models tend to under-count (negative bias), while Qwen3-VL-4B slightly over-counts.

- **Over-counting Exception:** Qwen3-VL-4B is the notable outlier, showing a slight positive bias. This suggests that for this specific model, the *Language Takeover* mechanism (hallucinating extra tokens) may overpower the visual perceptual limit, leading to generated sequences longer than the ground truth.

Visual Acuity is Necessary but Insufficient. A critical divergence is observed in Gemma3-12B. Despite recording the highest attention sparsity and lowest entropy—indicating that it successfully maintains sharp visual focus throughout generation—its accuracy (27.73%) remains suboptimal. This stark contrast to Qwen3-VL (which fails due to attention decay) isolates a secondary failure mode: *Reasoning Collapse*. While Qwen models struggle to "see" the instances (Perceptual Failure), Gemma3-12B likely "sees" distinct entities but fails to map this visual register to the correct numerical sequence (Reasoning Failure). This dissociation confirms that robust visual grounding is a prerequisite for numeracy, but does not guarantee it without corresponding arithmetic reasoning capabilities.

5.2 Spatiotemporal Mechanisms of Failure

Finally, we explain the root causes of these errors through the lens of *Spatiotemporal Attention Decay*.

Spatial: The Halo Effect. Qualitative analysis of attention heatmaps confirms that even in models with high sparsity, attention peaks often do not align with object centroids but form a "halo" around the cluster periphery. This explains the prevalent under-counting bias observed in Section 5.1: the model enumerates the "area" or "edges" rather than discrete instances.

Temporal: Language Takeover. Figure 3 quantifies the temporal dynamics during multi-digit generation. We consistently observe a **Sparsity Drop-off**: the attention sparsity for the second digit (units) is significantly lower than for the first digit (tens).

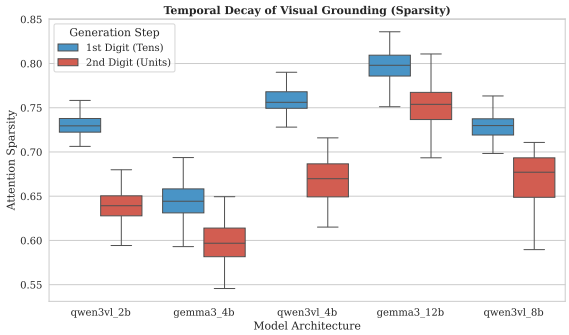

Figure 3: **The Language Takeover.** Box plots showing the distribution of attention sparsity for the 1st vs. 2nd generated digits. A consistent decay in visual grounding is observed across all models as generation progresses.

- For **models with higher accuracy**, the model maintains a relatively high sparsity across steps.

- For **models with lower accuracy**, we observe a sharp spike in entropy at the second step.

This confirms that counting errors are often driven by *Language Takeover*: once the initial visual estimate (often flawed by the Halo Effect) is tokenized, the model abandons the image and predicts the subsequent digit based on linguistic probability distributions, decoupling generation from visual reality.

6 Conclusion

We present a mechanistic framework, **Spatiotemporal Attention Decay**, to explain the fragility of VLM visual counting. Our analysis reveals that failures stem from a dual collapse in grounding: spatially, a **Halo Effect** restricts attention to object peripheries (convex hulls) rather than individual instances; temporally, **Language Takeover** causes visual focus to vanish during auto-regressive decoding, degrading multi-digit generation into hallucination driven by language priors. These findings indicate that standard scaling alone is insufficient to resolve numeracy deficits. Instead, future work must prioritize mechanisms that enforce persistent visual attention throughout the decoding timeline, such as object-centric objectives or region-anchored chain-of-thought reasoning.

Limitations

While our work offers a unified mechanistic framework for VLM counting failures, we acknowledge several limitations that define the scope of our findings:

Model and Architecture Diversity. Our analysis is currently restricted to state-of-the-art open-weights models (Qwen3-VL and Gemma3 families). While these cover a range of parameter scales (2B to 12B) and vision encoding strategies, we have not evaluated proprietary closed-source models (e.g., GPT, Gemini) due to the inaccessibility of their internal attention weights. Consequently, it remains an open question whether the observed *Halo Effect* is a universal artifact of Vision Transformers or a specific pathology of the architectures studied.

Sequence Length and Task Complexity. Our experiments primarily focus on two-digit counting tasks ($N \in [10, 16]$) with direct answer templates. This setting serves as a minimal viable testbed for *Language Takeover*, but it does not capture the dynamics of longer generation sequences. We did not investigate whether explicitly prompting the model for Chain-of-Thought (CoT) reasoning could mitigate the *Spatiotemporal Attention Decay* by "refreshing" visual attention, nor did we test larger counts (e.g., three-digit numbers) where the decay might be more severe.

Correlational vs. Causal Validation. Our study is primarily diagnostic and observational. While we establish a strong model-level correlation between attention sparsity and counting accuracy, we have not performed causal interventions to manipulate these internal states. We did not apply techniques such as activation steering or attention masking to manually suppress the Halo Effect or enforce visual focus during the generation of the second digit. Therefore, while the evidence for *Language Takeover* is compelling, strictly causal verification via mechanistic intervention remains a direction for future work.

References

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, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Freya Behrens, Luca Biggio, and Lenka Zdeborová. 2025. Counting in small transformers: The delicate interplay between attention and feed-forward layers. *Preprint*, arXiv:2407.11542.

Jinjin Cao, Zhiyang Chen, Zijun Wang, Liyuan Ma, Weijian Luo, and Guojun Qi. 2025. When images speak louder: Mitigating language bias-induced hallucinations in vlms through cross-modal guidance. *Preprint*, arXiv:2510.10466.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. 2024. Vision transformers need registers. *Preprint*, arXiv:2309.16588.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. See what you are told: Visual attention sink in large multimodal models. *Preprint*, arXiv:2503.03321.

Zongxia Li, Xiyang Wu, Guangyao Shi, Yubin Qin, Hongyang Du, Fuxiao Liu, Tianyi Zhou, Dinesh Manocha, and Jordan Lee Boyd-Graber. 2025. Videohallu: Evaluating and mitigating multi-modal hallucinations on synthetic video understanding. *Preprint*, arXiv:2505.01481.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. 2023. Teaching clip to count to ten. *Preprint*, arXiv:2302.12066.

Saurav Sengupta, Nazanin Moradinasab, Jiebei Liu, and Donald E. Brown. 2025. Can vision-language models count? a synthetic benchmark and analysis of attention-based interventions. *Preprint*, arXiv:2511.17722.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

A Appendix

A.1 Additional Spatiotemporal Analysis

In this section, we provide complementary qualitative and quantitative visualizations to further substantiate the *Spatiotemporal Attention Decay* framework proposed in the main text.

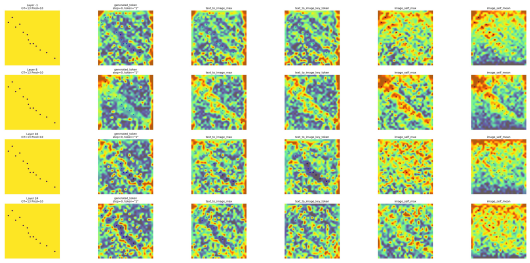

Figure 4: **Impact of Visual Enhancement on the Halo Effect.** Qualitative comparison of cross-modal attention maps before and after applying visual enhancement. After visual enhancement, the attention distribution becomes significantly sharper, penetrating the geometric centers of individual instances. This suggests that the Halo Effect is not an immutable property of the encoder but can be mitigated through targeted intervention.

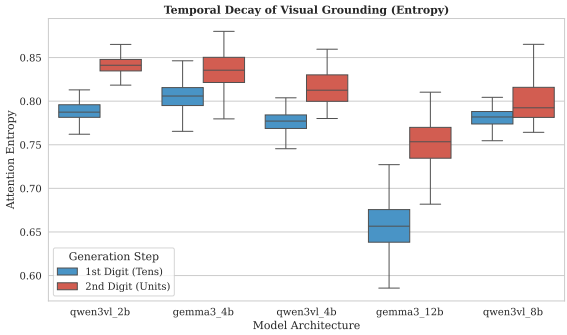

Figure 5: **Temporal Evolution of Visual Entropy (Uncertainty).** Complementing the Sparsity analysis in Figure 3, this plot illustrates the step-wise change in Visual Entropy (Eq. 3) during the auto-regressive decoding of multi-digit counts. We observe a *Phase Transition*: Entropy remains low (high certainty) during the initial token generation but spikes sharply at the transition to the second digit. This quantitative surge in uncertainty marks the precise moment of *Language Takeover*, where the model decouples from visual evidence.