I CAN'T BELIEVE IT CAN'T COUNT: VISION-LANGUAGE MODELS FAIL AT BASIC ENUMERATION BEYOND THE SUBITIZING RANGE

Amirhossein Afsharrad[1]*, **Seyed Shahabeddin Mousavi**[1], **Sanjay Lall**[1]

[1]Stanford University

ABSTRACT

Deep learning and vision-language models (VLMs) are increasingly deployed for real-world visual understanding tasks, yet their ability to perform the fundamental cognitive skill of basic enumeration remains poorly characterized. We present a systematic evaluation of three frontier VLMs (GPT-5, Gemini 3 Flash, and Claude Sonnet 4.6) on synthetic dot-counting images spanning 1–120 objects (360 trials each). Our results reveal dramatic performance differences: Gemini 3 Flash achieves 72.5% overall accuracy with a mean absolute error (MAE) of 1.26, demonstrating that accurate enumeration is achievable by current VLMs. By contrast, GPT-5 achieves only 15.6% accuracy (MAE 6.01) and Claude Sonnet 4.6 only 5.6% (MAE 25.04), with Claude exhibiting a severe positive bias of $+25.04$ reflecting systematic overcounting. For GPT-5, we observe two distinct behavioral regimes: a *counting regime* up to $\sim$70 objects characterized by consistent undercounting, and an *estimation regime* beyond $\sim$70 objects where predictions anchor strongly around 100. Chain-of-thought prompting and stimulus variation applied to Claude produced negligible improvement, pointing to architectural rather than prompt-level explanations. These findings demonstrate that enumeration failure is model-dependent rather than task-inherent, with implications for safety-critical applications including medical imaging, crowd monitoring, and scientific figure interpretation.

1 INTRODUCTION

Vision-language models (VLMs) have achieved impressive performance on complex visual reasoning tasks, from image captioning to visual question answering (Tong et al., 2024; Fu et al., 2024). This success has led to their deployment in increasingly consequential applications: analyzing scientific charts (Masry et al., 2022), counting cells in medical images (Falk et al., 2019), monitoring crowd density for public safety (Bai et al., 2020), and detecting objects in aerial imagery (Xia et al., 2018). These applications share a common requirement: *accurate enumeration*.

Yet recent work suggests VLMs struggle with precisely this capability. Guo et al. (2025) found that even state-of-the-art models "can't count to 20," while Zanchi et al. (2024) demonstrated that GPT-4V behaves as a "4-knower," reliably enumerating only up to 4 items, consistent with human subitizing limits but failing to develop true counting skills. These findings echo broader concerns about VLM perceptual limitations: Kamath et al. (2023) showed VLMs fail at basic spatial reasoning, and the BLINK benchmark revealed that humans outperform GPT-4V by over 40% on tasks solvable "within a blink" (Fu et al., 2024).

We contribute a rigorous empirical investigation of enumeration across three frontier VLMs using a controlled synthetic setup, and further test chain-of-thought prompting and stimulus variation to assess whether failures can be mitigated. Our key findings are:

- **Enumeration failure is model-dependent, not task-inherent**: Gemini 3 Flash achieves 72.5% accuracy (MAE 1.26), while GPT-5 (15.6%, MAE 6.01) and Claude Sonnet 4.6 (5.6%, MAE 25.04) fail dramatically.

*Corresponding author: afsharrad@stanford.edu

- **Qualitatively distinct failure modes**: GPT-5 exhibits a "100-anchor" phenomenon where predictions gravitate toward 100 for counts in the 85–105 range, while Claude Sonnet 4.6 displays severe systematic overcounting (bias +25.04) that grows with the true count.
- **Two distinct regimes in GPT-5**: A *counting regime* (up to ∼70 objects) characterized by consistent undercounting, and an *estimation regime* (beyond ∼70 objects) with haphazard, high-variance outputs.
- **Interventions fail to rescue poor models**: Chain-of-thought prompting and stimulus variation applied to Claude produce negligible accuracy improvements (5.6% to 6.7% at best), pointing to architectural rather than prompt-level explanations.

2 Experimental Setup

Stimulus Generation. We generated synthetic images of black dots (radius 10 pixels) on white backgrounds (800×800 pixels). Dot positions were randomized with minimum spacing of 2.5× radius to prevent overlap, isolating pure enumeration ability from confounds such as object recognition or occlusion.

Models and Protocol. We evaluated GPT-5, Gemini 3 Flash, and Claude Sonnet 4.6 via their respective APIs using structured JSON output with two fields: `count` (integer) and `confidence` (high/medium/low). We tested counts from 1 to 120 with 3 repetitions per count (360 trials per model) at default temperature (1.0).

Prompting Conditions. All three models were evaluated under an original prompt asking for precise, systematic counting of black dots. To test whether failures could be mitigated, we additionally evaluated Claude Sonnet 4.6 under two intervention conditions: a *chain-of-thought* (CoT) prompt instructing the model to divide the image into sections and sum the per-section counts, and a *varied objects* condition using images with mixed shapes and colors. Full prompt text is provided in Appendix A.

Metrics. We report accuracy (exact match), mean absolute error (MAE), signed bias (mean signed error), and prediction variance across repetitions.

3 Results

Figure 1 presents the cross-model comparison. Per-model detailed figures are in Appendix B; Claude intervention results are in Appendix C.

3.1 Dramatic Cross-Model Differences

The most striking finding is the range of performance across models evaluated under identical conditions. Gemini 3 Flash achieves 72.5% accuracy (MAE 1.26), demonstrating that accurate enumeration of synthetic stimuli is achievable by current VLMs. GPT-5 achieves 15.6% accuracy (MAE 6.01) with a modest positive bias of +0.72. Claude Sonnet 4.6 performs worst at 5.6% accuracy (MAE 25.04) with a severe positive bias of +25.04, reflecting systematic large-scale overcounting that grows with the true count. These differences, observed under identical stimuli and prompts, confirm that enumeration failure is model-dependent.

3.2 The Accuracy Cliff

Performance degrades precipitously with count for both GPT-5 and Claude. For GPT-5, accuracy is 93% for 1–9 dots, drops to 53% for 10–19, and falls below 10% for counts ≥30. Gemini maintains substantially higher accuracy throughout. This cliff around 10–20 objects aligns with the human subitizing boundary (Trick & Pylyshyn, 1994; Kaufman et al., 1949), though Gemini's performance demonstrates that models need not be bound by it.

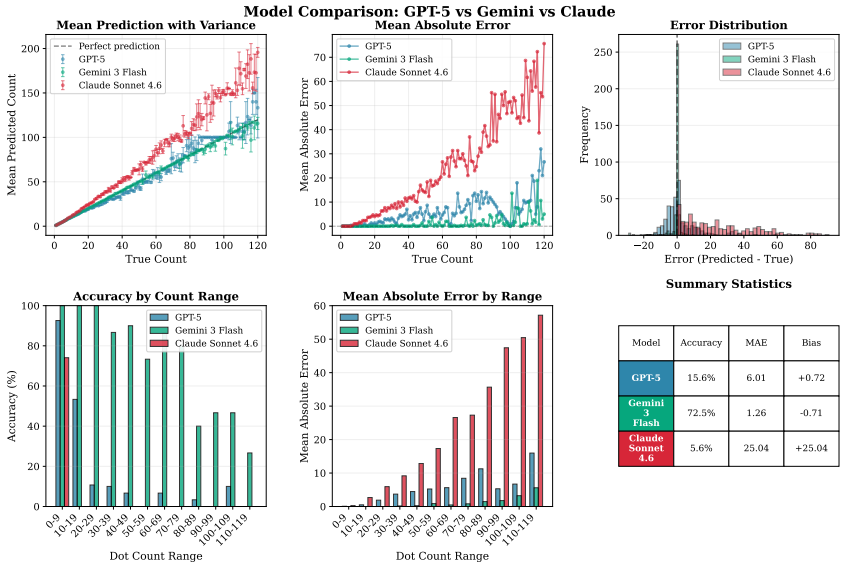

Figure 1: Cross-model enumeration performance across 360 trials each. *Top row*: (Left) Mean predictions: Gemini 3 Flash tracks the ideal closely, GPT-5 anchors around 100 at high counts, Claude Sonnet 4.6 overcounts severely; (Center) MAE by true count; (Right) Error distributions revealing qualitatively different failure modes. *Bottom row*: Accuracy and MAE by count range, with summary table.

3.3 The 100-Anchor Phenomenon

The top-left panel of Figure 1 reveals a pattern specific to GPT-5: predictions track true counts reasonably until approximately 70 dots, then gravitate toward 100. Values both below and above 100 are pulled toward this anchor, suggesting either a learned prior toward a salient round number, training data biases, or representational limitations in encoding large quantities. Gemini shows no such anchoring. The anchor creates a non-monotonic MAE pattern in GPT-5: error peaks in the 70–89 range, *decreases* for 90–109 where the anchor is closer to truth, then spikes again at 110–119.

3.4 Qualitatively Distinct Failure Modes

The error distributions (Figure 1, top-right) show that the three models fail differently. Gemini's distribution is tightly centered on zero. GPT-5 exhibits a moderately asymmetric distribution, with the negative tail truncating around -20 while the positive tail extends further. Claude's distribution is dramatically right-skewed: the model almost never undercounts and produces large positive errors throughout the range, consistent with its bias of $+25.04$.

3.5 Counting vs. Estimation Regimes in GPT-5

GPT-5 data reveal two behavioral regimes. From approximately 10 to 70 objects, the model exhibits consistent undercounting, suggesting continued enumeration effort. Beyond $\sim$70 objects, outputs become haphazard with high variance across repetitions (std up to 34), combined with prominent 100-anchoring. This transition aligns with human cognitive limits where subitizing gives way to serial counting, which GPT-5 appears unable to perform reliably (Dehaene, 2011).

3.6 Interventions Fail to Rescue Claude

Chain-of-thought prompting (6.1%, MAE 25.10, bias $+14.98$) and varied-object stimuli (6.7%, MAE 32.12, bias -7.33) produced negligible accuracy improvements over Claude's baseline (5.6%, MAE 25.04). The varied-objects condition increased MAE despite a marginal accuracy gain and

flipped the bias sign, suggesting a shift in estimation strategy without genuine improvement in enumeration.

4 Why Does Counting Fail?

Subitizing Without Counting. Humans possess two distinct numerical systems: the Object Tracking System enabling parallel individuation of 1–4 items (subitizing), and serial counting requiring attention shifts (Trick & Pylyshyn, 1994). Failing VLMs may exploit subitizing-like parallel processing but lack mechanisms for reliable serial enumeration. GPT-5 functions as a "10-knower," an improvement over GPT-4V's "4-knower" status (Zanchi et al., 2024), but still fundamentally limited. Gemini's success suggests some models have developed more robust enumeration strategies.

Architectural Constraints. Vision encoders like CLIP compress images into fixed-dimensional representations optimized for semantic similarity, not precise spatial structure (Tong et al., 2024). Object individuation, i.e., tracking "this dot" vs. "that dot," may be fundamentally difficult in such representations (Sengupta et al., 2025). The failure of CoT prompting to improve Claude's accuracy, despite explicitly instructing a divide-and-count strategy, suggests the limitation lies in visual feature extraction rather than reasoning.

The 100-Anchor as Learned Prior. GPT-5's gravitation to 100 may reflect training data statistics, since 100 appears frequently as a salient round number in text corpora. This is consistent with findings that deep learning struggles to abstract natural numbers from visual representations (Chen et al., 2019). Its absence in Gemini and replacement by a different failure mode in Claude suggest model-specific rather than universal behavior.

5 Implications and Limitations

Practical Recommendations. Model choice matters enormously for enumeration tasks. Gemini 3 Flash substantially outperforms the other models on this benchmark, so practitioners should evaluate specific models rather than assuming uniform VLM limitations. For models that perform poorly, enumeration beyond $\sim$10 objects should not be deployed without human oversight in safety-critical settings such as medical imaging, crowd monitoring, and scientific chart interpretation.

Limitations and Future Work. We evaluated models on synthetic stimuli; real-world images with occlusion, varying sizes, and semantic content may reveal different failure modes. Interventions were tested only on Claude; it remains unclear whether CoT prompting could further improve Gemini's performance or would reveal an upper bound. The architectural or training differences that explain Gemini's superiority remain an open question. Promising future directions include testing specialized counting architectures (Hou et al., 2025), extended count ranges beyond 120, and targeted fine-tuning to mitigate anchoring.

6 Conclusion

We presented a systematic cross-model evaluation showing that enumeration failure in VLMs is neither universal nor inevitable. Gemini 3 Flash achieves 72.5% accuracy across 1–120 objects, while GPT-5 and Claude Sonnet 4.6 fail through qualitatively different mechanisms: a "100-anchor" heuristic with two distinct regimes in GPT-5, and systematic large-scale overcounting in Claude. Prompting interventions fail to rescue Claude, pointing to architectural explanations. As VLMs are deployed in safety-critical applications requiring accurate enumeration, these findings urge both caution and careful model selection.

Reproducibility Statement

All stimuli were synthetically generated with parameters fully specified (800$\times$800 pixels, 10-pixel radius dots, 2.5$\times$ radius minimum spacing). Models were evaluated via their respective APIs with default temperature and structured JSON output. Prompts for all conditions are in Appendix A. Code and data will be released upon publication.

REFERENCES

Haoyue Bai, Jiageng Mao, and S.-H. Gary Chan. A survey on deep learning-based single image crowd counting: Network design, loss function and supervisory signal. *arXiv preprint arXiv:2012.15685*, 2020.

Shijie Chen, Wei Luo, and Jiangping Liu. Cognitive deficit of deep learning in numerosity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 2630–2637, 2019.

Stanislas Dehaene. *The Number Sense: How the Mind Creates Mathematics*. Oxford University Press, revised and updated edition, 2011.

Thorsten Falk, Dominic Mai, Robert Bensch, et al. U-Net: deep learning for cell counting, detection, and morphometry. *Nature Methods*, 16(1):67–70, 2019.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Xuyang Guo, Zekai Huang, Zhenmei Shi, Zhao Song, and Jiahao Zhang. Your vision-language model can't even count to 20: Exposing the failures of VLMs in compositional counting. *arXiv preprint arXiv:2510.04401*, 2025.

Kuinan Hou, Jing Mi, Marco Zorzi, Lamberto Ballan, and Alberto Testolin. Assessing the visual enumeration abilities of specialized counting architectures and vision-language models. *arXiv preprint arXiv:2512.15254*, 2025.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? Investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9161–9175. Association for Computational Linguistics, 2023.

Edna L. Kaufman, Mary W. Lord, Thomas W. Reese, and John Volkmann. The discrimination of visual number. *American Journal of Psychology*, 62(4):498–525, 1949.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279. Association for Computational Linguistics, 2022.

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

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024.

Lana M. Trick and Zenon W. Pylyshyn. Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychological Review*, 101(1):80–102, 1994.

Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. DOTA: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3974–3983, 2018.

Alberto Zanchi, Alberto Testolin, and Marco Zorzi. Visual enumeration remains challenging for multimodal generative AI. *arXiv preprint arXiv:2402.03328*, 2024.

A PROMPTS USED IN ALL CONDITIONS

Original Prompt (all three models). "Count the exact number of black dots in this image. Be precise and systematic in your counting."

Chain-of-Thought Prompt (Claude only). "Count the objects in this image. First, divide the image into sections (e.g., quadrants or rows), count the objects in each section, then sum them up to get the total count. Be systematic and precise."

Varied Objects Prompt (Claude only). "Count the total number of objects in this image, regardless of their shape or color. Be precise and systematic in your counting."

B Per-Model Detailed Results

Figures 2–6 show full six-panel result figures for each experimental condition, following the same layout as Figure 1: mean predictions with variance, MAE by true count, error distribution, prediction variance, accuracy by count range, and MAE by count range.

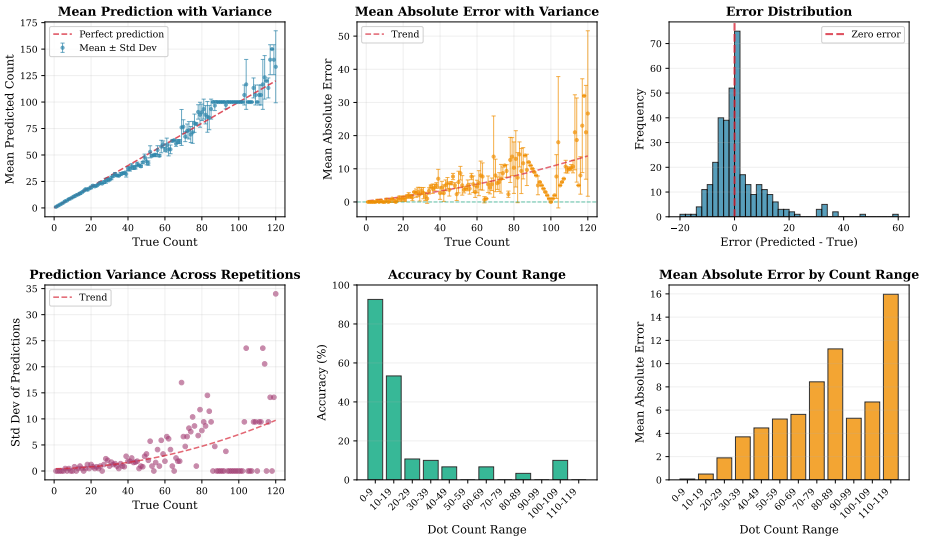

Figure 2: GPT-5 (baseline) individual results across 358 trials. Predictions anchor around 100 for true counts above approximately 85, and accuracy collapses sharply beyond 20 objects.

C Claude Intervention Comparison

Figure 7 shows all three Claude conditions overlaid for direct comparison. No intervention produces meaningful accuracy improvement, confirming that Claude's failures are not addressable through prompting or stimulus modification alone.

Figure 3: Gemini 3 Flash (baseline) individual results across 360 trials. The model tracks the ideal prediction line closely throughout the full 1–120 range, with low MAE and near-zero bias.

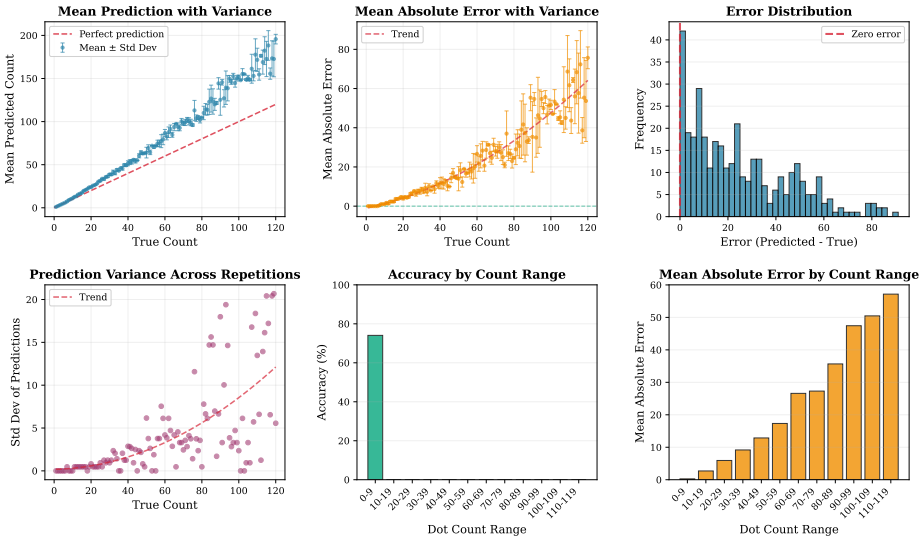

Figure 4: Claude Sonnet 4.6 (baseline, original prompt) individual results across 360 trials. The model systematically overcounts across the full range, with a strongly right-skewed error distribution and bias of +25.04.

Figure 5: Claude Sonnet 4.6 with chain-of-thought prompting, individual results across 360 trials. Accuracy (6.1%) and MAE (25.10) are nearly identical to the baseline, indicating CoT prompting does not address the underlying failure.

Figure 6: Claude Sonnet 4.6 with varied objects (mixed shapes and colors), individual results across 360 trials. Accuracy (6.7%) is marginally higher but MAE (32.12) worsens relative to baseline, and the bias flips to -7.33, indicating a change in estimation strategy rather than improved enumeration.

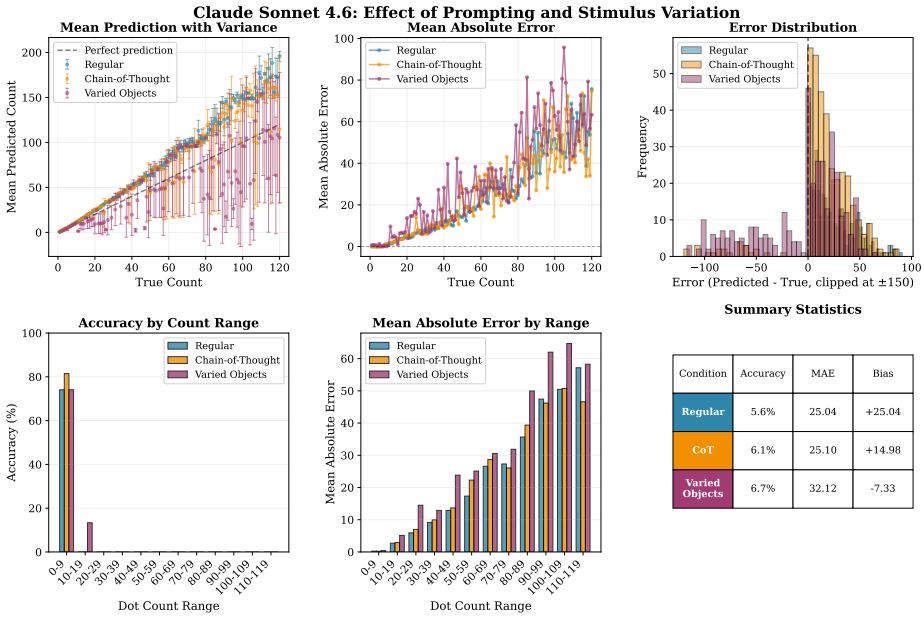

Figure 7: Claude Sonnet 4.6 across three conditions: regular prompt (blue), chain-of-thought (orange), and varied objects (purple). Summary statistics: Regular 5.6% accuracy, MAE 25.04, bias +25.04; CoT 6.1%, MAE 25.10, bias +14.98; Varied Objects 6.7%, MAE 32.12, bias −7.33.