PLOS ONE

RESEARCH ARTICLE

Visual enumeration remains challenging for multimodal generative AI

Alberto Testolin[1]*, **Kuinan Hou**[2], **Marco Zorzi**[3,4]

1 Department of General Psychology and Department of Mathematics, University of Padova, Padova, Italy, **2** Department of General Psychology, University of Padova, Padova, Italy, **3** Department of General Psychology and Padova Neuroscience Center University of Padova, Padova, Italy, **4** IRCSS San Camillo Hospital, Venice-Lido, Italy

* alberto.testolin@unipd.it

Check for updates

OPEN ACCESS

Citation: Testolin A, Hou K, Zorzi M (2025) Visual enumeration remains challenging for multimodal generative AI. PLoS One 20(9): e0331566. https://doi.org/10.1371/journal.pone.0331566

Editor: Yih-Kuen Jan, University of Illinois Urbana-Champaign, UNITED STATES OF AMERICA

Received: April 7, 2025

Accepted: August 18, 2025

Published: September 12, 2025

Copyright: © 2025 Testolin et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: The data used in the current study and the complete code of our benchmark are freely available on GitHub at: https://github.com/CCNL-UniPD/Numbersense-AI.

Funding: We are grateful to OpenAI for granting research access to the GPT-4V and DALL-E APIs. This work was supported by the European Union Next-Generation EU grant (grant n.PE13 - BAC FAIR SP.10 to M.Z.) and by the Italian Ministry of University and Research (PRIN grant

Abstract

Many animal species can approximately judge the number of objects in a visual scene at a single glance, and humans can further determine the exact cardinality of a set by deploying systematic counting procedures. In contrast, it has been observed that even state-of-the-art AI systems have very limited enumeration skills. In this work, we propose two benchmark tasks inspired by cognitive science that allow to precisely evaluate the visual enumeration capabilities of multimodal foundation models, thereby providing an objective measure of their number sense and counting level. We consider popular visual question answering models (BLIP, LLaVA and ViLT) as well as advanced image-to-text (Gemini, GPT and Qwen) and text-to-image (DALL-E, FLUX and Stable Diffusion) AI systems. Our analyses show that even the most advanced models cannot reliably name the number of objects in simple visual stimuli or generate images containing a target number of items, as indexed by their low accuracy in both types of tasks. Especially for numbers outside the subitizing range, their responses are often far from the target numerosity, and, in stark contrast with human behavior, in many cases the distribution of errors depends on the object category. We also observe some striking mistakes with small numbers. Our findings demonstrate that developing an intuitive visual understanding of number remains challenging for AI models and that merely increasing model size might not be a viable strategy to promote the emergence of systematic counting skills. We release the full code of our benchmark to facilitate the evaluation of enumeration skills in future AI systems.

1 Introduction

Artificial Intelligence (AI) is progressing rapidly, with deep learning models approaching or even surpassing human performance in a variety of domains, including perceptual judgements [1] and natural language processing [2]. However, even the most advanced multimodal AI systems still struggle in judging the numerosity of visual sets, a core cognitive capability that humans share with many animal species [3]. Even infants are sensitive to numerosity [4] and toddlers can generate sets that contain a target number of items [5], suggesting that a pre-verbal understanding of numerical quantities develops well before language development

n. C53D23004110006 to M.Z.). K.H. acknowledges the support of the China Scholarship Council (ID: 202307820031). There was no additional external funding received for this study.

Competing interests: The authors have declared that no competing interests exist.

and formal education. Small numerosities in the "subitizing range" (up to 4) are perceived in an exact manner (i.e., enumeration is error-free), while the numerosity of larger sets is only approximately estimated when counting is precluded [6]. In the latter case, responses follow Weber's law, so that variability increases proportionally to the mean estimate [3]. Another key signature of our number sense is its abstract nature: similar response patterns are observed for items of all categories, despite variation in object features (such as color or shape). Indeed, numerosity is spontaneously extracted by our visual system [7] and it is encoded independently from object category, location, or presentation modality [4]. Moreover, during childhood, humans (but no other species) learn counting algorithms that allow them to establish the exact cardinality of any set of objects by performing a one-to-one mapping between visual or auditory items and the list of number words [8]. Importantly, there is a broad consensus that number sense and mastery of counting principles are foundational for the development of numeracy and the acquisition of higher-level mathematical competence [3,9–11].

AI researchers have engineered a variety of specialized computer vision architectures to count objects in visual scenes, often tailored to specific categories such as animals [12], crowds [13] or common objects encountered in a specific domain of interest [14]. The most popular framework consists of running an object detector to first segment the target items in the image and then explicitly counting the resulting bounding boxes or object proposals [15,16], in some cases summing fractional counts estimated from different sections of the image [17]. However, in these approaches numerosity representations do not emerge within the model itself because the encoding of numbers is delegated to an external, hard-wired (and often category-specific) mechanism. A radically different perspective considers the possibility that numerosity representations might spontaneously emerge in neural systems as a high-order statistical feature of the sensory signal [18]. Indeed, a rudimentary visual number sense has been shown to emerge in small-scale generative models trained with the goal of reconstructing images with a varying number of items [19,20] and number-selective neurons have been observed in generic convolutional networks trained for object recognition [21,22]. Thus, one might wonder whether a similar capacity could emerge in modern multimodal foundation models [23], which are large-scale generative architectures trained on huge data sets that exhibit emergent abilities [24] and can readily solve a wide range of downstream tasks [25,26]. Unlike domain-specific architectures engineered for visual counting [27,28], foundation models are domain-general systems that can be used out-of-the-box without the need of fine-tuning on numerical tasks. However, despite their flexibility and their remarkable performance in a variety of domains [29], even the most advanced foundation models fall short in tasks that require the manipulation of numerical information [30–32], calling for a systematic investigation of their basic visual enumeration skills.

In line with the proposal of using methods from cognitive science to test AI models [33], in this work we address this problem by introducing two benchmark tasks that are commonly used to evaluate enumeration skills in humans: numerosity naming [34], which requires establishing how many items are present in a given stimulus, and numerosity production [5,35], which requires generating a set containing a target number of items. The former task can be used to probe *image-to-text* architectures, while the latter can be used to probe *text-to-image* architectures. Our benchmark allows to characterize the distribution of model responses using a variety of object categories, providing aggregate scores indicating the overall model performance and additional metrics that measure whether the distribution of responses follows a human-like pattern. Perfect accuracy across the entire numerical range would suggest the emergence of systematic counting skills, while error-free responses with only small numbers would either indicate subitizing capabilities, or that counting is only partially developed as in children who do not fully master the counting principles [36,37].

Error-prone responses centered on the target number would instead suggest that the AI model relies on numerosity estimation, which may follow Weber's law (as in humans) or not.

We perform our evaluation across a wide range of models of different sizes and complexities, considering the most powerful multimodal AI systems and visual question answering models available at the time of the research. In the image-to-text domain, we consider AI models that can provide written answers to non-trivial questions about the content of an image or accurate descriptions of complex visual scenes. In particular, we test three popular architectures used in visual question answering: ViLT (vision-and-language transformer) [38], BLIP (bootstrapping language-image pre-training model) [39] and LLaVA (large language and vision assistant model) [40]. We further test Qwen2.5-VL [41], which is one of the latest open-source multimodal models available, as well as two proprietary models that are considered among the most advanced multimodal systems currently available: GPT-4V [42] and Gemini Pro Vision [43]. In the text-to-image domain, we instead consider foundation models that can produce high-quality visual content following detailed user prompts provided in natural language. We test two popular open-source generative architectures for images, Stable Diffusion [44,45] and FLUX [46], as well as DALL-E [47,48], which is regarded among the most powerful proprietary systems. For Qwen and DALL-E, we also compare different versions of the same architecture to investigate whether increasing model size supports more refined visual enumeration skills.

The research contributions of our work are multifaceted. From a methodological perspective, we introduce a unified experimental procedure to evaluate the numerical skills of both image-to-text and text-to-image AI models, which allows us to quantitatively characterize their numerical competence across different types of object categories. We make our benchmarking pipeline publicly available to allow systematic evaluation of future AI models [https://github.com/CCNL-UniPD/Numbersense-AI]. From a theoretical perspective, we demonstrate that although larger models generally possess a better number sense, merely scaling-up the model size might not be the best way to spur the emergence of systematic counting skills. In this regard, we show that the weak performance in visual enumeration might be partially due to properties of the training corpora commonly used to build foundation models: for example, in popular training datasets the frequency of numerosities rapidly falls off according to a power law, implying that larger numbers are underrepresented, and textual captions often contain numerical information that is not related to the numerosity of the visual scene, thus injecting noise into the alignment of different input modalities.

The article is structured as follows: in the Materials and Methods section, we describe the problem setting, introducing the benchmark tasks and the evaluation metrics used to quantify numerical competence. We also provide details about the models considered and the relevant baselines. In the Results section, we present the quantitative results comparing different models, along with a detailed analysis of response errors and the statistical properties of two representative corpora commonly used to train large-scale multimodal AI systems. In the Discussion, we review and interpret our findings, while in the Conclusions we highlight their implications for AI research and discuss possible future directions.

2 Material and methods

2.1 Benchmark tasks

2.1.1 Image-to-text: Numerosity naming. In the numerosity naming task, the models are asked to establish how many objects are present in a set of simple images containing up to 10 objects. We created a new dataset of synthetic images, each including only items of the

same category sampled from 5 possible object types: apples, people, butterflies, colored dots, and "fast cards" depicting regularly placed clip-arts similar to those used to test number sense in young children [36]. Examples of stimuli are shown in Fig 1. For each object class and target number n, we created 50 high-resolution images (1024×1024 pixels) where the n

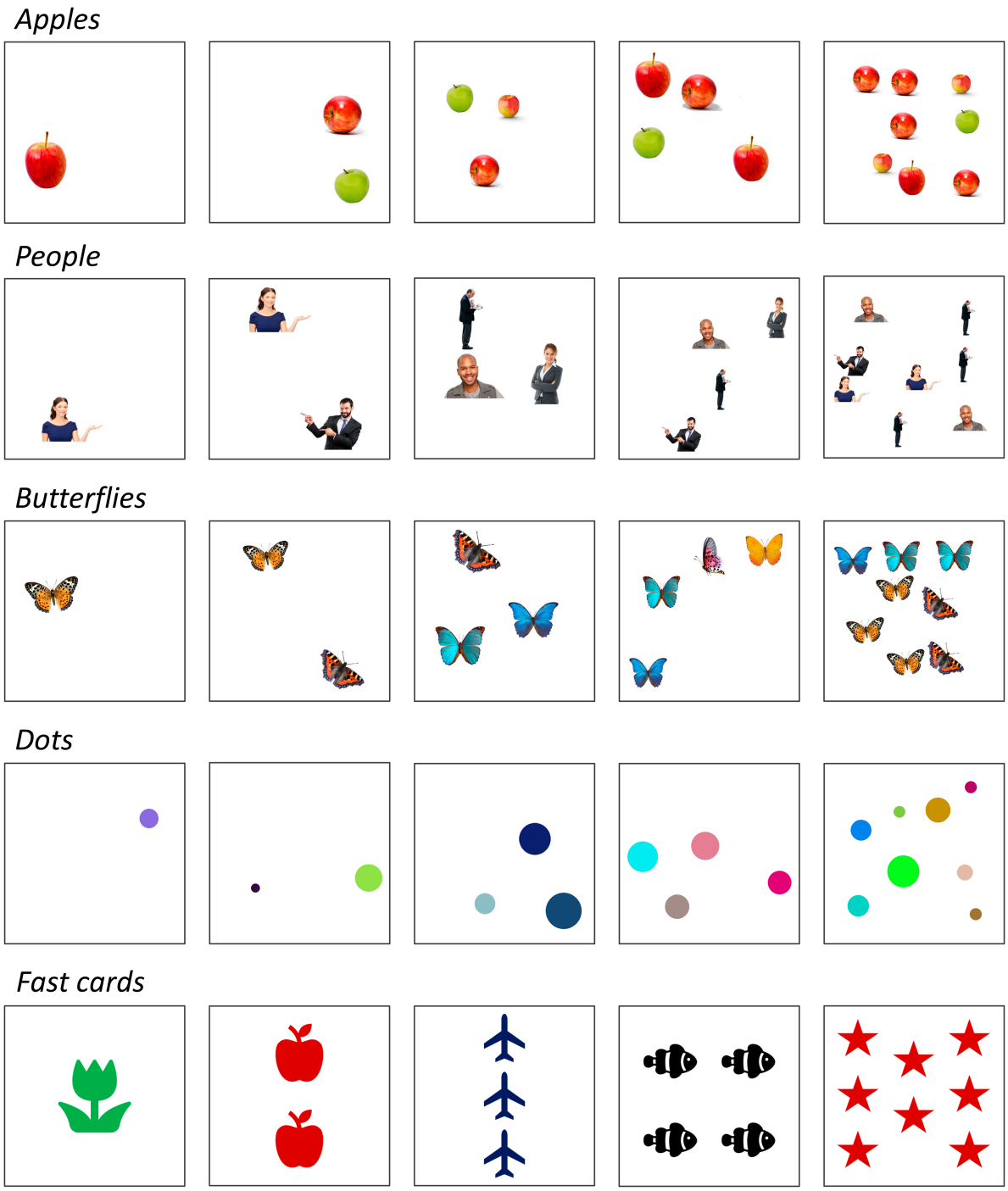


Fig 1. Samples from the numerosiy naming task. Each row contains samples from a different object category, while columns correspond to different numerosities: 1, 2, 3, 4 and 8.

https://doi.org/10.1371/journal.pone.0331566.g001

objects have variable size and are randomly placed on a uniform white background, with no overlap. Fast card stimuli are created using clip-arts of common objects (apples, bells, butterflies, candies, cars, fish, flowers, planes, stars) drawn in different colors (black, blue, green, orange, red).

For each model tested, we consider three different prompting methods and select the one leading to the best performance, measured as mean absolute distance from the target number. The first prompt requires to explicitly estimate the number of objects belonging to a specific category (i.e., *How many apples / butterflies / people / dots / shapes are there in the picture?*). The other two use the more general words "objects" or "things" to identify the items to count. To make sure that the models are prompted correctly, we also perform a control simulation related to a non-numerical task using the entire set of "apples" stimuli, probing the models with the following prompt: *What does the image represent?* and considering as correct the following answers: *apple(s)* and *fruit*. All models provided the correct answer for the entire set of stimuli in the control task, thus demonstrating a proper understanding of the image content and the prompt structure.

Model responses are automatically parsed: if present, number words are converted to numerical values using the word2number Python library, and responses are discarded if they contain multiple numbers or vague quantification terms (e.g., "a few", "a bunch of"). We verified that at least 20 eligible trials were recorded for each number / object category combination.

2.1.2 Text-to-image: Numerosity production. In the numerosity production task, each model is asked to generate 100 high-resolution images containing a target number of objects, in analogy with numerosity production tasks used in animal and human studies [5,35]. Target objects belong to the same classes used for the naming task, except for the "fast cards" category, which could be underrepresented in the corpora used to train foundation models.

All models were initially prompted with the following text: *An image with n apples / butterflies / people / dots* (where *n* varied between 1 and 10). When $n = 1$ the prompt was adjusted to the singular form. However, for the dots category this prompting method resulted in poor generations: we obtained better results when the models were prompted with a more specific description of the image: *n filled dots in white background*. For the people category, instead, we obtained better results with the prompt *An image with n persons*.

The generated images were automatically parsed using a computer vision pipeline [49] optimized on a set of 4,000 images that were manually labeled by one of the authors (K.H.) and independently verified by another author (A.T.). The pipeline employs Grounding DINO [50], a state-of-the-art object detection model, to identify objects within the generated images (see Fig 2). The process begins with parsing the prompts used for the text-to-image models to identify the object category for each image. Grounding DINO then detects all objects belonging to the specified category by providing bounding boxes for each detected object, along with a confidence score for each bounding box. Bounding boxes with low confidence scores are filtered out. To align the detections made by Grounding DINO with human annotations, a grid search was performed to optimize its confidence score threshold based on the NAE metric. The grid search explored confidence scores from 0.01 to 0.99 in increments of 0.01. The confidence score was iteratively adjusted to minimize the NAE between Grounding DINO's outputs and the human annotations. The optimal confidence score was determined to be 0.40, achieving a minimum NAE of 0.05. We provide the NAE as a function of different thresholds in S4 Fig. This calibration ensured that discrepancies between the object counts detected by Grounding DINO and the human annotations were minimized, thereby enhancing the accuracy and reliability of the automatic evaluation process. If the automatic pipeline did not find any countable object, the image was discarded and the model was prompted again.

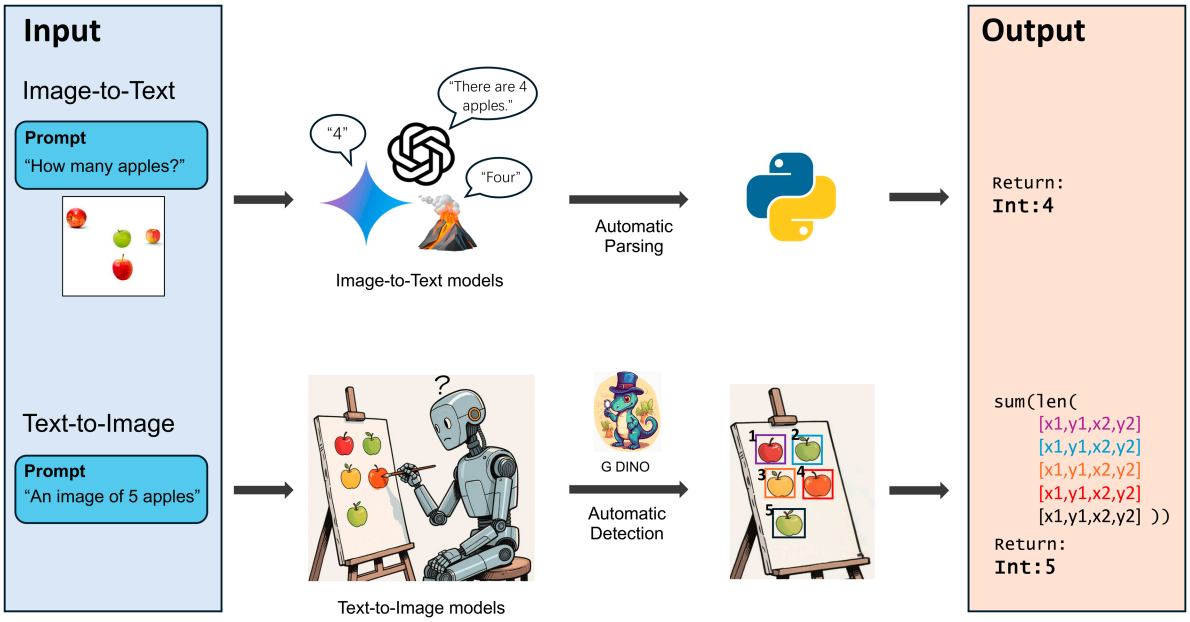


Fig 2. Graphical representation of our evaluation pipeline. Numerosity naming (image-to-text) is represented in the upper stream, while numerosity generation (text-to-image) is represented in the lower stream.

https://doi.org/10.1371/journal.pone.0331566.g002

2.2 Multimodal AI architectures

For the numerosity naming task, we consider several representative image-to-text foundation models. We first test three popular multimodal architectures used in visual question answering: ViLT [38], BLIP-2 [39], and LLaVA [40]. We further consider three large-scale multimodal language models representing the state-of-the-art in AI research: the open-source multimodal Qwen2.5-VL model recently developed by Alibaba [41], the multimodal GPT-4V model developed by OpenAI [42] and the multimodal Gemini Pro model developed by Google [43]. All these systems have remarkable visual reasoning abilities and can answer non-trivial questions related to image content (e.g., *What does the image represent? What are the feelings of the people in the scene and why?*). For ViLT, we test the vilt-b32-mlm version available through Hugging Face. This architecture incorporates text embeddings into a Vision Transformer, allowing it to have a minimal design for vision-and-language pre-training and thus speeding-up model training and inference phases. It has a total of 87.4 million parameters. For BLIP-2, we test the blip2-flan-t5-xl version developed by Salesforce, also available through Hugging Face. This architecture is an improved version of BLIP [51] that approaches state-of-the-art performance on several challenging benchmarks [52]; we explored all versions of the backbone models except the t5-xxl model (due to GPU memory constraints) and found that the version that used Flan-T5 as a language model yielded the best accuracy. The chosen model version has a total of 4.1 billion parameters. For LLaVA, we test the 1.6 version available through Hugging Face. LLaVa is an open-source end-to-end trained large multimodal model based on the transformer architecture, which combines a vision encoder and the Nous Hermes 2 large language model for general-purpose visual and language understanding, achieving impressive chat capabilities [40]. The chosen model version has a total of 34 billion parameters. Qwen2.5-VL is the latest version of the Qwen model family [41]. It has been optimized for visual recognition, image reasoning, captioning, and answering

general questions about an image, outperforming many open-source and proprietary models on common industry benchmarks. We focus on the most powerful model version, which has a total of 72 billion parameters, but we also consider the smaller architecture with 7 billion parameters to measure performance gains with respect to model size. GPT-4V and Gemini Pro are regarded among the most powerful generalist AI systems to date, thanks to their unprecedented ability to understand and process an arbitrary mix of input images and texts. Technical details regarding the underlying architecture and inner working of these models (including engineered modules that might be used to solve specific tasks) have not been revealed; it has been speculated that both these models might have more than one trillion parameters.

For the numerosity production task, we consider three different image generation architectures that have proven capable of generating high-quality images following a textual description, also taking into account stylistic instructions, fine-grained details, and relational features (e.g., *A photo of an astronaut riding a horse in photorealistic style*). One is represented by the open-source Stable Diffusion (SD) model family, developed by Stability AI and freely available through Hugging Face. It is a latent diffusion model that combines an autoencoder with a diffusion model trained on the latent space of the autoencoder. We test both version 2.1 [44], which has approximately 500 million parameters, and the newest version 3.5 Large, which has approximately 8 billion parameters. We then consider the open-source FLUX model [46], which is also a diffusion-based architecture with a hybrid design that combines multimodal and parallel diffusion transformer blocks, allowing for a more effective processing of visual and textual data. FLUX has approximately 12 billion parameters, providing enhanced capacity for generating high-resolution, hyper-realistic images and accurately rendering complex visual scenes and text. Finally, we test two proprietary systems from the DALL-E model family using the API interface provided by OpenAI. DALL-E 2 [47] is an improved version of the original text-to-image DALL-E model, featuring a total of 3.5 billion parameters. DALL-E 3 [48] is the latest and most powerful version, which was trained using highly descriptive synthetic captions for the training images. Its number of parameters is currently unknown.

2.3 Counting-specific baselines

As baselines for the numerosity naming task we also evaluate two state-of-the-art architectures specifically tailored for counting tasks: the Point, Segment, and Count (PseCo) model [27], which is a detection-based counting model that utilizes point-level supervision and segmentation cues to improve object localization and enumeration, and the Training-Free Object Counting model [28], which can perform category-agnostic object counting without additional training, leveraging pre-trained feature extractors.

2.4 Evaluation metrics

2.4.1 Overall performance score. For each benchmark task, in addition to accuracy, we also compute the Normalized Absolute Error (NAE) score, which is a commonly used metric for evaluating counting abilities that addresses limitations of the most commonly used Mean Absolute Error (MAE). Indeed, MAE treats all errors equally, while NAE normalizes the absolute error by dividing it by the target value, making it sensitive to proportional rather than absolute differences. This normalization ensures fairness across scales, as larger targets inherently permit greater absolute errors without compromising accuracy. NAE also aligns with perceptual principles like Weber's law, reflecting the proportional nature of human numerosity estimation. NAE is defined as:

$$\text{NAE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|G_i - T_i|}{T_i}$$

where n is the total number of test samples, G_i is the generated numerosity for the i-th test sample and T_i is the target numerosity for the i-th test sample.

As a baseline, we report the NAE of a random model probed on the same number of test trials. This model generates responses randomly sampled from a uniform distribution in the range 1-20 (see related confusion matrix in Supplementary S1 Fig), obtaining a NAE of 2.38.

2.4.2 Counting level. We also assess the counting level of each model by applying standard criteria used in the literature on the development of counting skills [36]. To be considered an "n-knower" (i.e., "1-knower", "2-knower", "3-knower", "4-knower") the model has to: 1) correctly return n at least 67% of the time when tested for n; and 2) return n less than 50% of the time when the target number is different from n. The counting level is assessed on the average responses collected across all object categories.

2.4.3 Estimation performance. When explicit counting is precluded, in humans and other animal species numerosity estimation tasks yield a distribution of errors that varies systematically according to Weber's law [53]. In particular, while responses are error-free in the 1-4 subitizing range, for larger numerosities the standard error of the estimates increases proportionally to the mean, indicating scalar variability [54,55]. To investigate whether the AI responses follow a human-like estimation pattern, we measure the Pearson correlation between the confusion matrices produced by the models and that obtained from an ideal human observer (see related confusion matrix in Supplementary S1 Fig) that estimates numerosity in accordance with Weber's law, assuming a standard Weber fraction w of 0.15 [56] and error-free responses in the subitizing range.

2.4.4 Over- and under-estimation trends. To investigate the presence of systematic biases in visual enumeration, we analyzed the distribution of prediction errors (i.e., response - target), separately for each model and category. We applied the Wilcoxon signed rank test to assess whether prediction errors significantly deviated from zero (with alpha level set to 0.05). This non-parametric test, chosen for its robustness to non-normal distributions, involves ranking the absolute values of the errors, then comparing the sum of ranks for positive errors against the sum of ranks for negative errors. We then computed the effect size (r) using Cohen's formula:

$$r = \frac{|z|}{\sqrt{n}} \tag{1}$$

where z is the z-score derived from the Wilcoxon test statistic and n is the sample size. Following standard practice, we interpret $r \geq 0.3$ as evidence of a meaningful bias (medium-to-large effect size): a model is classified as overestimating if the sum of ranks for positive errors significantly exceeds that of negative errors and $r \geq 0.3$, and as underestimating if the opposite occurs and $r \geq 0.3$.

3 Results

Table 1 summarizes all results from our benchmark, ranking the models in terms of NAE. Despite the simplicity of the enumeration tasks, none of the models achieves perfect accuracy. Counting is particularly poor, with no model exceeding the level of 4 items. In humans, accurate performance up to 4 is also supported by subitizing [6], which in turn suggests that current multimodal foundation models cannot count at all. Some of the models mimic the

Table 1. Leader-board according to Normalized Absolute Error (NAE). The Corr w/ Human column reports the correlation with the confusion matrix produced by an ideal human observer. The last column reports the estimated number of model parameters (in Billions).

Model	Accuracy ↑	NAE ↓	Counting level	Corr w/ Human	Size (Billion)
Image-to-text models					
Qwen-72B	**0.89**	**0.01**	4-knower	0.89	72
Qwen-7B	0.82	0.02	4-knower	0.90	7
Gemini Pro	0.60	0.10	4-knower	0.93	600?
GPT-4V	0.74	0.13	4-knower	0.92	>1000?
LLaVa	0.37	0.13	2-knower	0.85	34
VILT	0.28	0.27	1-knower	0.67	0.1
BLIP2	0.29	0.33	1-knower	0.52	4
Text-to-image models					
FLUX	**0.44**	**0.25**	2-knower	0.89	12
SD3.5	0.41	0.26	3-knower	0.91	8
DALLE-2	0.34	0.37	2-knower	0.83	3.5
SD2.1	0.28	0.39	1-knower	0.77	0.5
DALLE-3	0.32	0.47	1-knower	0.84	?

https://doi.org/10.1371/journal.pone.0331566.t001

pattern of human visual estimation, as indicated by the strong correlations with the confusion matrices produced by the ideal human observer.

3.1 Image-to-text models

All models achieved the minimum number of eligible trials required, without the need of further prompting (total number of responses discarded for ViLT: 0; BLIP-2: 6; LLaVA: 0; Qwen: 0; GPT-4V: 0; Gemini: 20). For all models, the best performance was achieved with the generic "things" prompt, while for BLIP-2 and LLaVa the best performance was achieved with the category-specific prompts.

For ViLT and BLIP-2 the response accuracy was lower than 30%. The corresponding confusion matrices (CMs) reported in Fig 3 clearly show the presence of anchoring effects, leading these models to choose stereotyped responses (e.g., 4 or 6). The pattern of responses for these models also drastically varied between categories (minimum correlation between CMs for ViLT: 0.04, BLIP-2: 0.27), suggesting that they fail to abstract numerical information. Moreover, in sharp contrast with human adults, ViLT and BLIP-2 often returned wrong answers even for images with only one or two objects. According to standard criteria used in human developmental studies, these models can be considered "1-knowers", that is, they can only reliably enumerate single objects, as typical of children younger than three years of age [36]. The responses of LLaVA were slightly more accurate (37%) and consistent across object categories (minimum correlation between CMs: 0.52). This model achieves a "2-knower" counting level and also shows a more robust correlation with the confusion matrix produced by the ideal human observer.

Among the large multimodal models, Qwen achieved the highest accuracy (89%) with a NAE of 0.01, suggesting that it has the strongest enumeration capabilities. This model performs well across different object categories (minimum correlation between CMs: 0.85) and shows reliable enumeration up to four items, which makes it a "4-knower". Interestingly, the smaller version of Qwen with only 7B parameters still performs reasonably well, reaching an accuracy of 82% and demonstrating a higher correlation with the pattern of responses provided by the ideal human observer. The responses of proprietary foundation models, despite their larger size, are generally less accurate than Qwen (GPT-4V: 74%; Gemini: 60%), suggesting that these systems possess very rudimentary enumeration skills. Confusion matrices

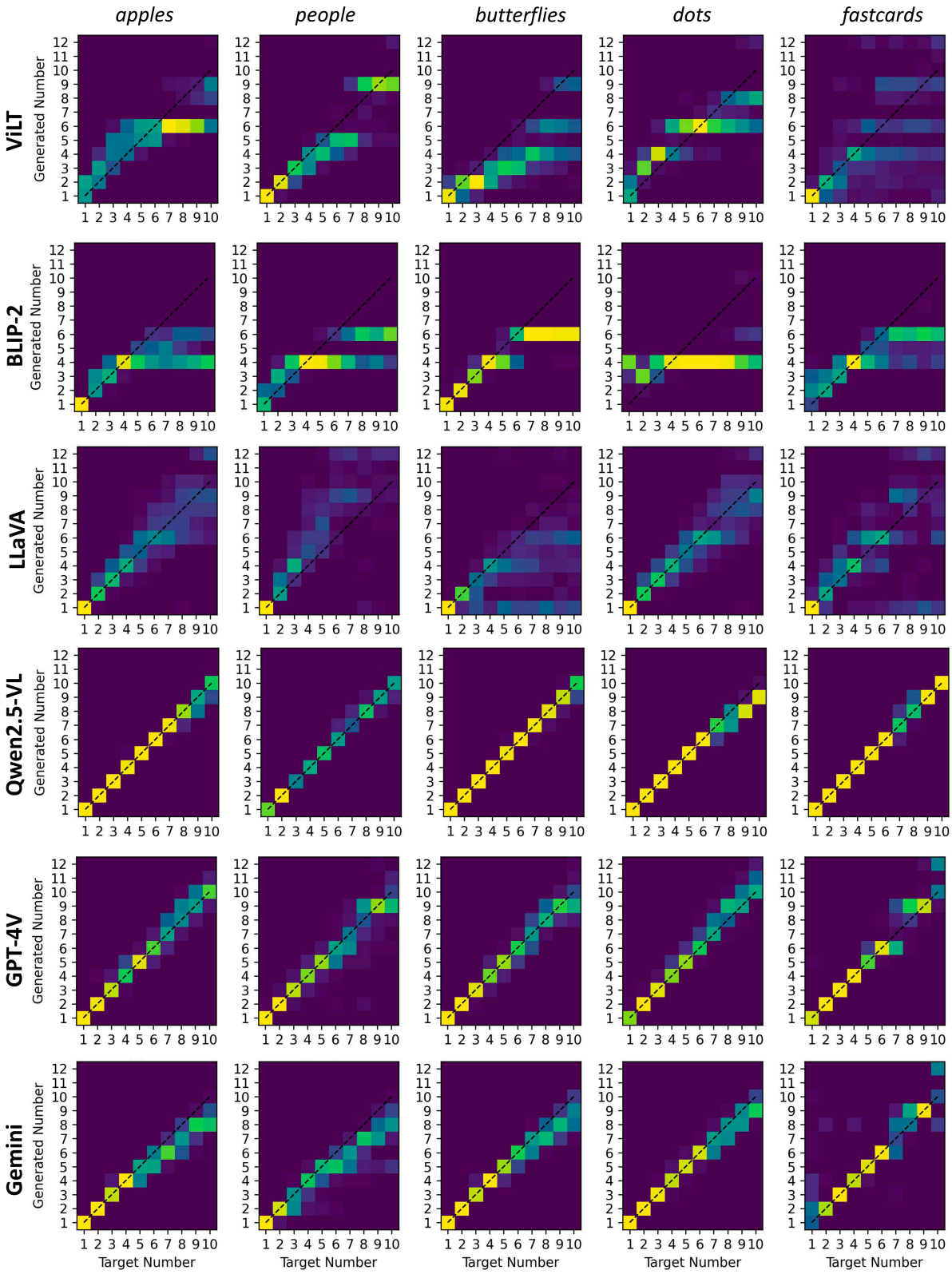


Fig 3. Confusion matrices for the numerosity naming task. Each panel shows the distribution of models' responses across different object categories: apples, people, butterflies, dots and fast cards. The x-axis represents the target number, while the y-axis represents the corresponding model responses. Response frequency is encoded using a perceptually uniform colormap (blue = 0%, yellow = 100%). Qwen2.5-VL stands for Qwen2.5-VL 72B.

https://doi.org/10.1371/journal.pone.0331566.g003

are consistent between categories (the minimum correlation is 0.87 for GPT-4V and 0.70 for Gemini). In some cases Gemini produces unexpected responses with images containing only one item, for example answering that *There are two things in the image: an apple and a white background* (these responses were discarded).

Considering the distribution of response errors, GPT-4V and Gemini, like Qwen, can be characterized as "4-knowers", that is, they exhibit reliable enumeration only up to four items. As noted before, this level of performance is consistent with subitizing (i.e., parallel individuation of visual items), but it also highlights the failure in mastering counting skills.

As shown in Table 2, image-to-text models have a general tendency to underestimate the numerosity across all object categories. Qwen-72b and GPT-4V are less biased compared to the other models, showing underestimation (Qwen) or overestimation (GPT) trends only for the Dots category.

Notably, the counting-specific architectures considered as baseline were not able to successfully count the target objects across all categories, suggesting that models trained with real images might struggle with generalizing their counting skills on synthetic visual stimuli (see S2 Fig in Supplementary Information).

3.2 Text-to-image models

Examples of generated images are shown in Fig 4. In comparing the outputs of different text-to-image models, we observed noteworthy stylistic trends across both open-source and proprietary systems. The earlier version of Stable Diffusion (SD2.1) demonstrates a broader artistic scope, spanning paintings, clip-art, and realistic renderings. The latest open-source models (SD3.5 and Flux) appear to have narrowed their stylistic variations, resulting in more coherent and realistic images. Proprietary models from the DALL-E family instead produce visual scenes with a characteristic synthetic style. DALL-E 2 mostly generates images with white backgrounds and well-defined objects, while DALL-E 3 injects more details, but still retaining a cartoon-like style (in a few cases DALL-E 2 generated images containing reflections: we discarded these trials to avoid ambiguous identification of objects by the automatic evaluation pipeline).

The mean response accuracy was fairly low for all models (ranging between 0.28 and 0.44), suggesting that the numerosity production task is far more challenging than the numerosity naming task. Confusion matrices are shown in Fig 5. Similarly to the naming task, the response patterns were not homogeneous across categories (minimum correlation between CMs for SD2.1: 0.43, SD3.5: 0.37, FLUX: 0.69, DALL-E 2: 0.63, DALL-E 3: 0.56). In a few cases SD3.5, FLUX and DALL-E 2 exhibit error-free responses, but that mostly happens for the generation of a single object (and not across all categories). All other models make errors

Table 2. Analysis of image-to-text models' estimation biases across object categories. "-" indicates no systematic bias, "Over." means the model's responses are systematically higher than the ground truth, and "Under." means they are systematically lower. Numbers in brackets represent the effect sizes (Cohen's *r*).

Model	Apples	People	Butterflies	Dots	Fastcards
ViLT	-	Under. (0.44)	Under. (0.68)	-	-
BLIP2	Under. (0.64)	Under. (0.61)	Under. (0.62)	Under. (0.46)	Under. (0.54)
llava34b	-	Over. (0.67)	Under. (0.68)	-	-
Qwen-72b	-	-	-	Under. (0.51)	-
Qwen-7b	Under. (0.46)	-	Under. (0.33)	Under. (0.40)	-
GPT-4V	-	-	-	Over. (0.48)	-
Gemini Pro	Under. (0.61)	Under. (0.67)	Under. (0.50)	Under. (0.42)	Over. (0.33)

https://doi.org/10.1371/journal.pone.0331566.t002

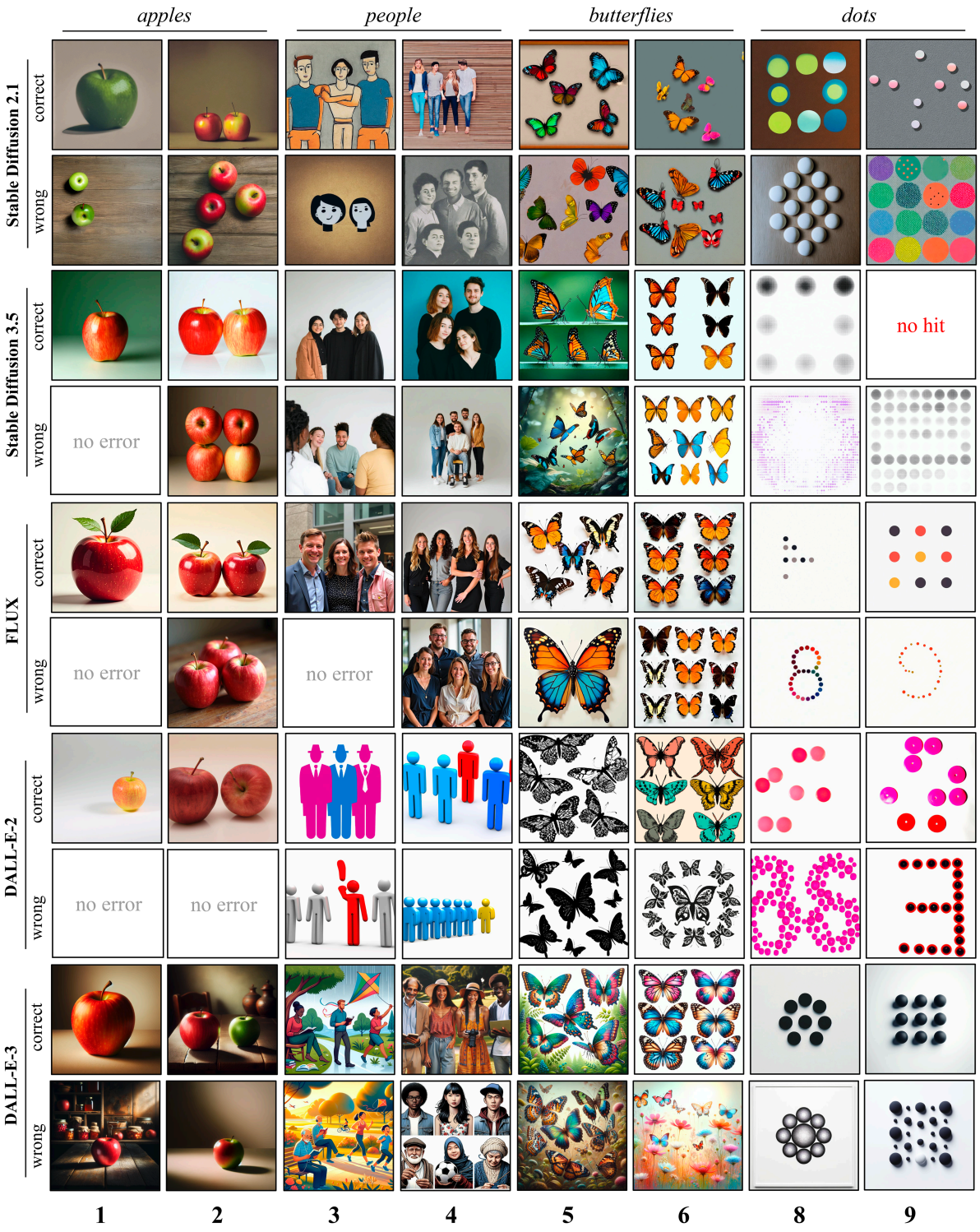


Fig 4. Examples of images generated by text-to-image models in the numerosity production task, showcasing both correct and wrong generations (the target number is indicated at the bottom). We report two images for each target category: apples, people, butterflies, and dots. For the dots category, in a few cases FLUX and DALL-E 2 generated images containing a wrong number of dots, which were nevertheless arranged according to the target digit shape (e.g., 8 in the figure). Surprisingly, SD3.5 was unable to generate a single correct response when the target numerosity was larger than 8.

https://doi.org/10.1371/journal.pone.0331566.g004

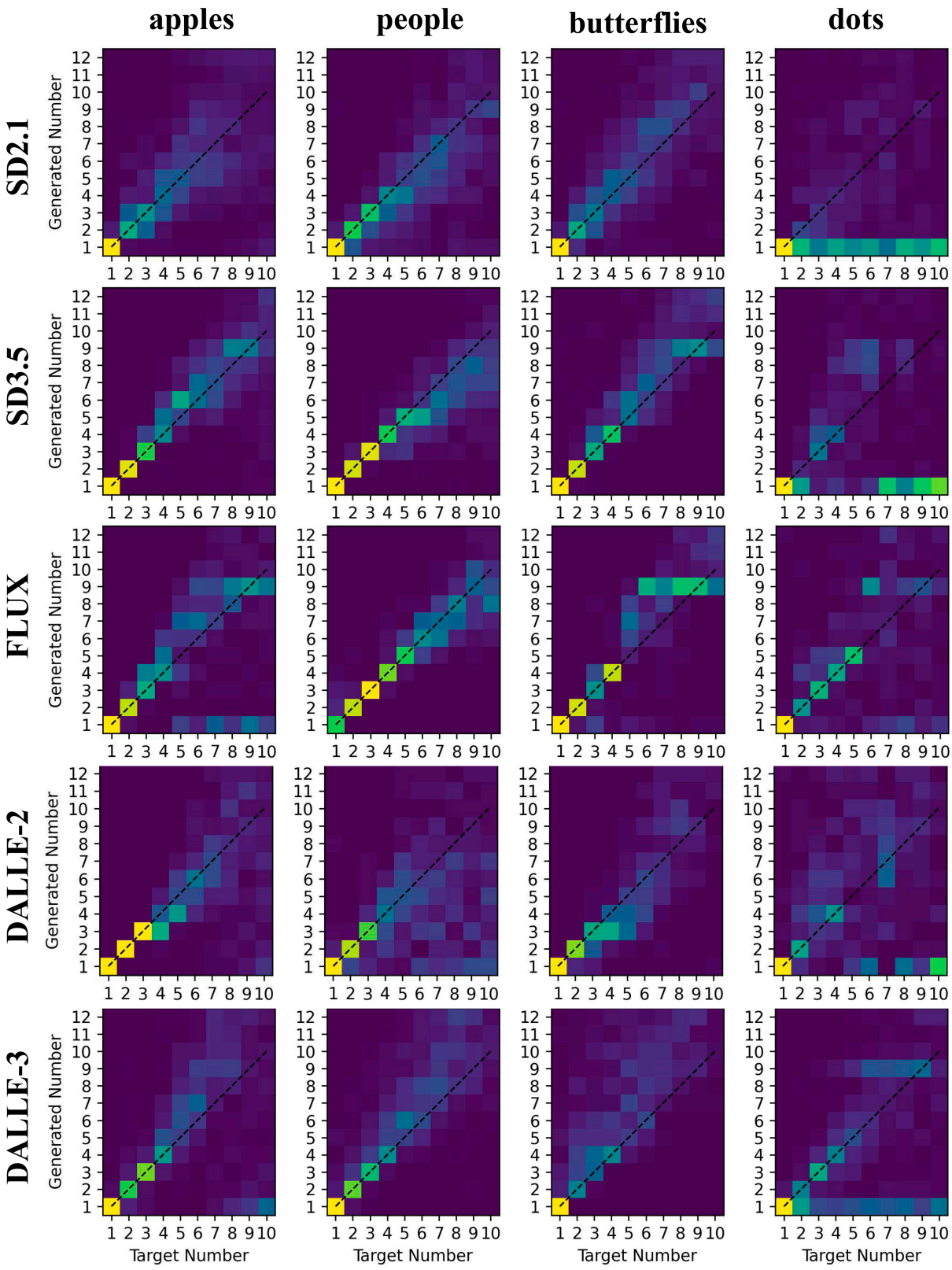


Fig 5. Confusion matrices for the numerosity production task. The x-axis represents the target number, while the y-axis represents the corresponding model responses. Response frequency is encoded using a perceptually uniform colormap (blue = 0%, yellow = 100%).

https://doi.org/10.1371/journal.pone.0331566.g005

even in this condition. The NAE ranges from 0.25 (FLUX) to 0.47 (DALL-E 3). According to criteria used in human developmental studies, the highest counting level is achieved by SD3.5, which can be classified as a "3-knower".

Table 3 shows the results of the analysis of over- vs. under-estimation trends for all text-to-image models; in this case we observe a tendency to overestimate.

3.3 Why is visual enumeration so challenging?

The poor performance of state-of-the-art AI models in a simple task like visual enumeration might seem surprising, given their impressive range of emergent abilities and considering that signatures of number sense have been observed in smaller-scale deep learning models (for discussion, see [18]).

One possible explanation lies in the statistical properties of the material used to train these foundation models. Indeed, it has been recently shown that the performance of multimodal models scales linearly as the concept frequency in pre-training data grows exponentially, which means that "zero-shot" performance in tasks involving underrepresented concepts will normally be poor [57]. To better characterize possible biases in the distribution of numerical information in popular training datasets, we conducted an in-depth analysis on the frequency of appearance of different numerosities in two datasets commonly used to train large-scale multimodal AI systems: Conceptual Captions 3 Million (CC3M) [58], which contains more than 3 million image-caption pairs, and LAION-400M [59], a large collection of 400 million English (image, text) pairs. Our investigation focused on examining the distribution of textual numerosities across all image captions present in the datasets, deploying natural language processing tools to identify numbers referring to countable objects in the image. In order to identify target numerosities, we first defined a valid numerosity as a textual segment containing a numerical token followed by either a noun or an optional adjective plus a noun (e.g., "1 apple" or "3 tall trees"). This criterion allowed to focus on explicit numerical references to countable objects while excluding textual references where numbers serve as model identifiers (e.g., "Porsche 911") or cardinal descriptors in product names (e.g., "35th anniversary"). To implement this filtering process, we employed the spaCy library to determine the part-of-speech (POS) tags for each word in the captions. After locating candidate structures, we systematically excluded measurements and units. The exclusion criteria encompassed a comprehensive range of metric prefixes, spanning from the microscopic to the massive: micro-, milli-, centi-, kilo-, mega-, giga-, nano-, tera-, and peta-. These prefixes were filtered out when combined with common measurement units. The excluded base units covered fundamental physical quantities: distance (meter/metre, mile, foot, inch, yard), volume (liter/litre), mass (gram, ounce, pound), power (watt), electrical potential (volt), current (amp), and energy (joule). Time-related measurements were similarly excluded, ranging from seconds to years, as were temperature scales (Fahrenheit, Celsius). By applying filtering criteria, we derived a more precise set of image-text pairs that genuinely reflected countable numerosities.

Table 3. Analysis of text-to-image models' estimation biases across object categories. "-" indicates no systematic bias, "Over." means the model's outputs are systematically higher than ground truth, and "Under." means they were systematically lower. Numbers in brackets represent the effect sizes (Cohen's *r*).

Model	Apples	People	Butterflies	Dots
SD2.1	Over. (0.3)	-	Over. (0.45)	-
SD3.5	-	Under. (0.44)	Over. (0.57)	-
FLUX	-	-	Over. (0.45)	-
DALLE-2	-	Under. (0.34)	-	-
DALLE-3	-	Over. (0.64)	Over. (0.76)	-

https://doi.org/10.1371/journal.pone.0331566.t003

We analyzed the numerical information in image captions merging together heterogeneous object types, that is, all numerosities mentioned in a caption were summed into a single value, regardless of the object categories. For example, the caption "An image of 3 bananas and 2 apples" is counted as a total numerosity of 5, treating the quantities as a single aggregate and thus capturing the overall number of items described in the caption. Our analysis revealed that in both datasets numerosities are distributed according to a power law, which implies that larger numerosities are strongly underrepresented compared to small numerosities (see Fig 6). Interestingly, the frequency of appearance of decades (i.e., 10, 20, 30, 40, etc.) decreases

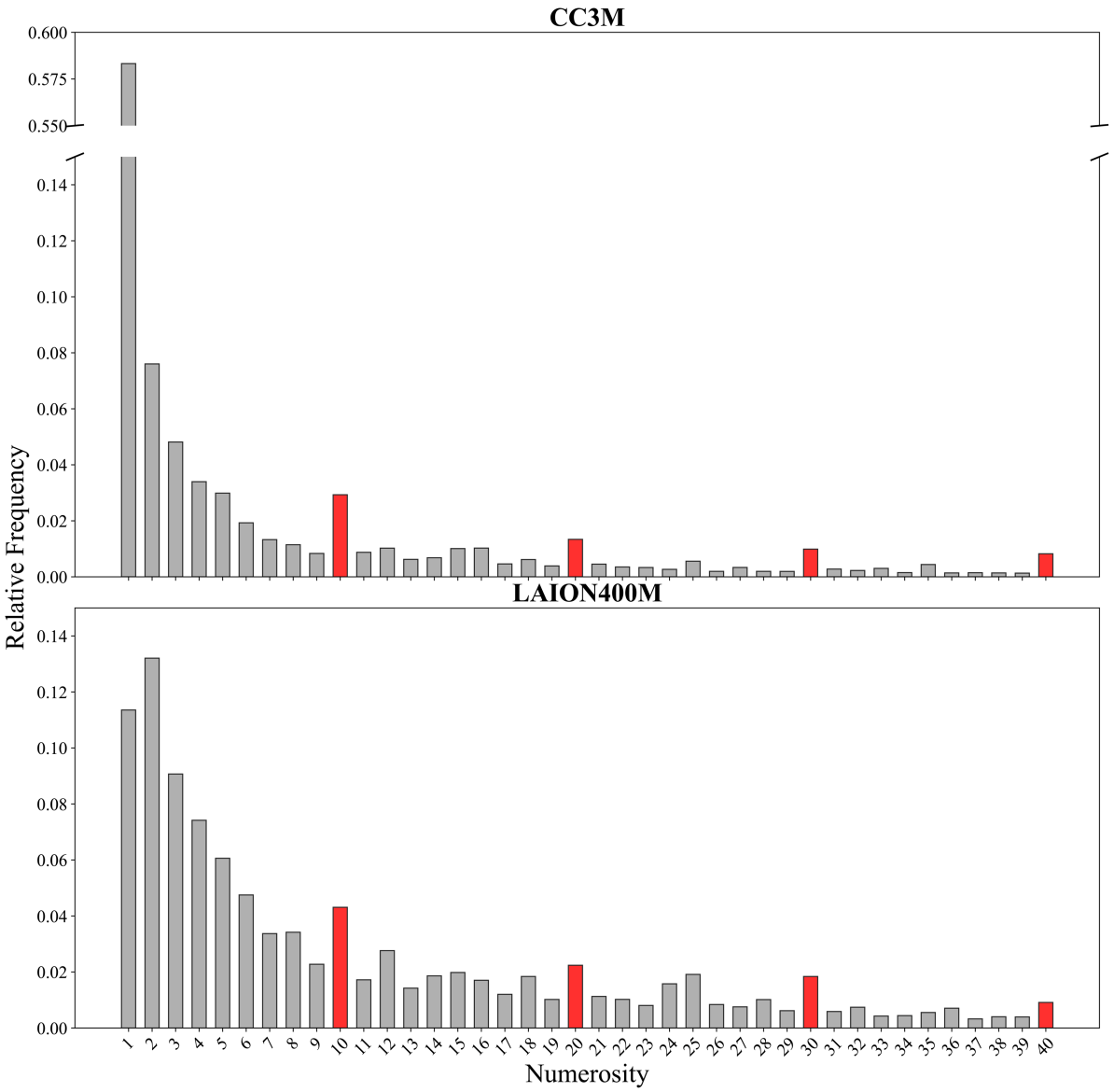


Fig 6. Power-law distribution of textual numerosities related to countable objects in the CC3M dataset (upper panel) and LAION-400M dataset (lower panel). Decade numbers are highlighted in red. The y-axis represents the relative frequency of appearance. A broken y-axis is used to accommodate the large difference in peak frequencies between the two datasets. The layout is proportioned such that the lower-frequency regions of both datasets share the same vertical height, making them visually comparable despite the scale difference.

https://doi.org/10.1371/journal.pone.0331566.g006

less sharply, indicating a preferential bias for these regular numerosities in the training material.

Such biased distribution of numerosities might contribute to the weak enumeration abilities we observed in all models, especially for images containing a large number of items. Nevertheless, it has been previously shown that signatures of number sense can emerge in smaller-scale models trained with synthetic (black and white) images containing a variable number of geometric shapes, even when numerosities are sampled according to the power-law Zipfian distribution observed in natural images [60]. This suggests that the poor number sense of large-scale AI systems might also (at least partially) stem from the more complex visual properties of real images, whose richness in finer-grained details could prevent the emergence of more explicit numerosity representations.

Finally, an additional issue could be the presence of noise in the linguistic captions, since the numerical information provided in the descriptive text sometimes might be completely unrelated to the numerosity depicted in the image. We manually checked a few image-text pairs containing explicit numerical information in their captions to assess whether the reported numerosities accurately matched the corresponding visual content, and indeed we found that sometimes the numerical text is misaligned with the image content (see examples in Supplementary S3 Fig). These noisy (image, text) pairs make it challenging to achieve an accurate semantic alignment between visual and textual representations, amplifying problems related to the modality gap observed in multimodal AI systems [61].

4 Discussion

The present work demonstrates that modern AI systems cannot yet reliably enumerate the number of objects in a visual scene, both in image-to-text and text-to-image tasks. Such a striking deficit is often observed even for sets containing only a few items, suggesting a counting level that is at best comparable to that of preschool children who do not fully master the counting principles [36,37]. Exact enumeration up to n=4 is consistent with subitizing, which in humans is supported by fast parallel individuation related to object tracking [62] and is independent from counting skills. This observation fits well with the finding that the best performing models generate responses to larger numbers that broadly follow the pattern of human numerical estimation, with scalar variability of the response distribution. Overall, these findings demonstrate that multimodal foundation models do not master counting skills, though the best models exhibit sparks of human-like number sense.

The ability to represent and manipulate visual numerosity should be regarded as a foundational skill for multimodal AI systems because it would ground the subsequent learning of more complex numerical and arithmetic concepts. Numerosity in humans is a primary visual feature (just as orientation or color) [63] and it is encoded by neuronal populations in multiple cortical regions in the primate brain [64–67]. Computational modeling studies have shown that sensitivity to numerosity can indeed emerge in small-scale deep learning models trained to generate synthetic images of object sets [19,20,68]: diffusion models, such as Stable Diffusion, FLUX and DALL-E, are trained with a similar objective on huge and heterogeneous image datasets that most likely contain substantial variability in numerosity. However, our analyses have shown that the empirical distribution of numerosities in the (image, text) pairs commonly used to train these systems follows a power law, therefore it might be possible that oversampling of small numerosities in the training corpora of foundation models has detrimental effects on their emergent representational space [57]. Another issue might lie in the mapping between perceptual numerosity representations, encoded in image embeddings, and number symbols (number words or Arabic digits) encoded in text

embeddings. In children, establishing such a bidirectional mapping is a sophisticated developmental process, which takes many years and requires explicit instruction [69]. The noisy nature of the (image, text) pairs used to train multimodal models might prevent the creation of a systematic mapping between different input modalities, with detrimental effects on the emergence of abstract numerosity representations.

It is also interesting to observe that even the most advanced proprietary models do not exhibit perfect accuracy on such simple tasks, suggesting that numerosity estimation has not been explicitly built-in in these systems. Nevertheless, one cannot exclude the possibility that some counting mechanisms were at least partially engineered as extra processing layers during prompt elaboration. Furthermore, it is well-known that the most recent AI systems can exploit the self-generation of code snippets to fulfill a user request, as in the case of mathematical problem solving [70], therefore models like GPT or Gemini could in principle also exploit external tools (e.g., based on object detection and a symbolic counting algorithm) to carry out these visual enumeration tasks. Quite surprisingly, however, in fact we noticed that compared to our own previous investigations [32] the performance of some proprietary models (DALL-E) has slightly degraded, suggesting that rather than pushing for improving number sense, the newest API provided by OpenAI might have actually reduced the compute budget available to the user. These issues highlight the dangers of using proprietary models in academic research [71], and calls for a broader adoption of open-source alternatives. In the case of our visual enumeration benchmark, it is encouraging to see that the best performance in both numerosity naming and numerosity production tasks was achieved by open-source models (Qwen and FLUX, respectively), demonstrating that future scientific investigations could be carried out by relying on transparent and well-documented neural architectures.

5 Conclusions

This work shows that systematic visual counting skills do not spontaneously emerge even in the most advanced foundation models, suggesting that significant progress in the design and training of multimodal architectures is still required to create systems that can reliably process visual quantities [72]. Efforts to improve the enumeration capabilities of multimodal models have been recently flourishing, highlighting the relevance of this type of benchmark for a comprehensive assessment of their emergent abilities [32,73]. For example, recent work suggests that fine-tuning the basic CLIP image-to-text model with a counting-contrastive loss can improve its ability to count the number of objects in images up to ten [74]. Another recently proposed approach is to enhance numerical reasoning by improving the numerical captions of the images used in training corpora [75]. For text-to-image models, others have proposed to implement counting guidance by using gradients of a counting network during the generative diffusion process, which however seems effective only for objects with a relatively simple shape [76]. We believe that our benchmark constitutes an important step forward to investigate these issues, possibly extending the upper limit of tested numerosities well beyond the 1-10 range as the performance of future AI models will improve.

It should also be noted that all models tested in the present work generate their output in a single step, thereby mimicking the kind of processing supported by an approximate estimation system operating in parallel [77]. However, in order to determine the exact number of items in a visual scene humans have learned to deploy iterative counting algorithms, which allow establishing a one-to-one mapping between visual items and number symbols [10]. This nevertheless requires generating outputs in a sequential manner, in analogy to "chain-of-thought" methods that have indeed proven useful to improve AI reasoning [78]. Whether

a similar approach could be used to tackle the enumeration tasks presented in this work is still an open question , which might be addressed by considering modern methods that can incrementally guide image synthesis and editing using multiple modalities [79].

In conclusion, we believe that a better visual grounding of numeracy development in AI models could be the key to enabling these systems to acquire and more reliably master mathematical knowledge [80] or even geometrical principles [81] without resorting to highly specialized hybrid architectures [82], which could open new possibilities for the use of AI in symbolic reasoning and knowledge discovery [83].

Supporting Information

S1 Fig. Confusion matrices for baseline models. On the left is the random choice model, while on the right is the ideal human observer, which has perfect accuracy in the subitizing range (1-4) and approximately estimates larger numbers according to Weber's law.
(PDF)

S2 Fig. Confusion matrices for counting-specific models for the numerosity naming task. Each panel shows the distribution of models' responses across different object categories: apples, people, butterflies, dots and fast cards. The x-axis represents the target number, while the y-axis represents the corresponding model responses. Response frequency is encoded using a perceptually uniform colormap (blue = 0%, yellow = 100%).
(PDF)

S3 Fig. Examples of numerical misalignment between images and textual captions in the LAION-400M training corpora. The numerical information in the text is highlighted in red.
(PDF)

S4 Fig. NAE of the image validation set as a function of Grounding DINO's threshold. The optimal threshold was found at 0.40.
(PDF)

Author contributions

Conceptualization: Alberto Testolin, Marco Zorzi.

Data curation: Kuinan Hou.

Funding acquisition: Alberto Testolin, Marco Zorzi.

Investigation: Alberto Testolin, Kuinan Hou.

Methodology: Alberto Testolin, Marco Zorzi.

Visualization: Alberto Testolin, Kuinan Hou.

Writing – original draft: Alberto Testolin.

Writing – review & editing: Kuinan Hou, Marco Zorzi.

References

1. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89–94. https://doi.org/10.1038/s41586-019-1799-6 PMID: 31894144
2. Gilardi F, Alizadeh M, Kubli M. ChatGPT outperforms crowd workers for text-annotation tasks. Proc Natl Acad Sci U S A. 2023;120(30):e2305016120. https://doi.org/10.1073/pnas.2305016120 PMID: 37463210

3. Dehaene S. The number sense: How the mind creates mathematics. OUP USA. 2011.
4. Izard V, Sann C, Spelke ES, Streri A. Newborn infants perceive abstract numbers. Proc Natl Acad Sci U S A. 2009;106(25):10382–5. https://doi.org/10.1073/pnas.0812142106 PMID: 19520833
5. Sella F, Berteletti I, Lucangeli D, Zorzi M. Spontaneous non-verbal counting in toddlers. Dev Sci. 2016;19(2):329–37. https://doi.org/10.1111/desc.12299 PMID: 25754974
6. Feigenson L, Dehaene S, Spelke E. Core systems of number. Trends Cogn Sci. 2004;8(7):307–14. https://doi.org/10.1016/j.tics.2004.05.002 PMID: 15242690
7. Cicchini GM, Anobile G, Burr DC. Spontaneous perception of numerosity in humans. Nat Commun. 2016;7:12536. https://doi.org/10.1038/ncomms12536 PMID: 27555562
8. Gallistel CR, Gelman R. Preverbal and verbal counting and computation. Cognition. 1992;44(1–2):43–74. https://doi.org/10.1016/0010-0277(92)90050-r PMID: 1511586
9. Halberda J, Mazzocco MMM, Feigenson L. Individual differences in non-verbal number acuity correlate with maths achievement. Nature. 2008;455(7213):665–8. https://doi.org/10.1038/nature07246 PMID: 18776888
10. Carey S, Barner D. Ontogenetic origins of human integer representations. Trends Cogn Sci. 2019;23(10):823–35. https://doi.org/10.1016/j.tics.2019.07.004 PMID: 31439418
11. Dolfi S, Decarli G, Lunardon M, De Filippo De Grazia M, Gerola S, Lanfranchi S, et al. Weaker number sense accounts for impaired numerosity perception in dyscalculia: Behavioral and computational evidence. Dev Sci. 2024;27(6):e13538. https://doi.org/10.1111/desc.13538 PMID: 38949566
12. Arteta C, Lempitsky V, Zisserman A. Counting in the wild. In: European conference on computer vision (ECCV), Amsterdam, The Netherlands; 2016. 483–98.
13. Khan MA, Menouar H, Hamila R. Revisiting crowd counting: State-of-the-art, trends, and future perspectives. Image Vision Comput. 2023;129:104597. https://doi.org/10.1016/j.imavis.2022.104597
14. Gao J, Zhao L, Li X. NWPU-MOC: A benchmark for fine-grained multicategory object counting in aerial images. IEEE Trans Geosci Remote Sensing. 2024;62:1–14. https://doi.org/10.1109/tgrs.2024.3356492
15. Trott A, Xiong C, Socher R. Interpretable counting for visual question answering. In: International conference on learning representations; 2018.
16. Zhang Y, Hare J, Prügel-Bennett A. Learning to count objects in natural images for visual question answering. In: International conference on learning representations; 2018.
17. Chattopadhyay P, Vedantam R, Selvaraju RR, Batra D, Parikh D. Counting everyday objects in everyday scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 1135–44.
18. Zorzi M, Testolin A. An emergentist perspective on the origin of number sense. Philos Trans R Soc Lond B Biol Sci. 2017;373(1740):20170043. https://doi.org/10.1098/rstb.2017.0043 PMID: 29292348
19. Stoianov I, Zorzi M. Emergence of a "visual number sense" in hierarchical generative models. Nat Neurosci. 2012;15(2):194–6. https://doi.org/10.1038/nn.2996 PMID: 22231428
20. Testolin A, Dolfi S, Rochus M, Zorzi M. Visual sense of number vs. sense of magnitude in humans and machines. Sci Rep. 2020;10(1):10045. https://doi.org/10.1038/s41598-020-66838-5 PMID: 32572067
21. Nasr K, Viswanathan P, Nieder A. Number detectors spontaneously emerge in a deep neural network designed for visual object recognition. Science advances. 2019;5(5):eaav7903.
22. Mistry PK, Strock A, Liu R, Young G, Menon V. Learning-induced reorganization of number neurons and emergence of numerical representations in a biologically inspired neural network. Nat Commun. 2023;14(1):3843. https://doi.org/10.1038/s41467-023-39548-5 PMID: 37386013
23. Li C, Gan Z, Yang Z, Yang J, Li L, Wang L, et al. Multimodal foundation models: From specialists to general-purpose assistants. FNT Comput Graph Vision. 2024;16(1–2):1–214. https://doi.org/10.1561/0600000110
24. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S. Emergent abilities of large language models. Trans Mach Learn Res. 2022.
25. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S. On the opportunities and risks of foundation models. arXiv preprint; 2021. https://doi.org/arXiv:210807258
26. Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint; 2023. https://doi.org/arXiv:230312712
27. 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; 2024. p. 17067–76.

28. Shi Z, Sun Y, Zhang M. Training-free object counting with prompts. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision; 2024. p. 323–31.
29. Romeo Z, Testolin A. Artificial intelligence can emulate human normative judgments on emotional visual scenes. R Soc Open Sci. 2025;12(7):250128. https://doi.org/10.1098/rsos.250128 PMID: 40740716
30. Testolin A. Can neural networks do arithmetic? A survey on the elementary numerical skills of state-of-the-art deep learning models. Appl Sci. 2024;14(2):744. https://doi.org/10.3390/app14020744
31. Rane S, Ku A, Baldridge J, Tenney I, Griffiths T, Kim B. Can generative multimodal models count to ten? In: Proceedings of the annual meeting of the cognitive science society; 2024.
32. Testolin A, Hou K, Zorzi M. Visual enumeration is challenging for largescale generative AI; 2024. https://doi.org/arXiv:240203328
33. Binz M, Schulz E. Using cognitive psychology to understand GPT-3. Proc Natl Acad Sci U S A. 2023;120(6):e2218523120. https://doi.org/10.1073/pnas.2218523120 PMID: 36730192
34. Revkin SK, Piazza M, Izard V, Cohen L, Dehaene S. Does subitizing reflect numerical estimation?. Psychol Sci. 2008;19(6):607–14. https://doi.org/10.1111/j.1467-9280.2008.02130.x PMID: 18578852
35. Whalen J, Gallistel CR, Gelman R. Nonverbal counting in humans: The psychophysics of number representation. Psychol Sci. 1999;10(2):130–7. https://doi.org/10.1111/1467-9280.00120
36. Le Corre M, Carey S. One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. Cognition. 2007;105(2):395–438. https://doi.org/10.1016/j.cognition.2006.10.005 PMID: 17208214
37. Lee MD, Sarnecka BW. Number-knower levels in young children: Insights from Bayesian modeling. Cognition. 2011;120(3):391–402. https://doi.org/10.1016/j.cognition.2010.10.003 PMID: 21109239
38. Kim W, Son B, Kim I. ViLT: Vision-and-language transformer without convolution or region supervision. In: International conference on machine learning; 2021. p. 5583–94.
39. Li J, Li D, Savarese S, Hoi S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning; 2023.
40. Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. Adv Neural Inform Process Syst. 2024;36.
41. Bai S, Chen K, Liu X, Wang J, Ge W, Song S. Qwen2. 5-vl technical report. 2025. https://doi.org/arXiv:250213923
42. Yang Z, Li L, Lin K, Wang J, Lin CC, Liu Z, et al. The dawn of lmms: Preliminary explorations with gpt-4v (ision). arXiv preprint. 2023;9(1):1. https://doi.org/10.48550/arXiv.230917421
43. Team G, Anil R, Borgeaud S, Wu Y, Alayrac JB, Yu J. Gemini: A family of highly capable multimodal models. arXiv preprint; 2023. https://doi.org/arXiv:2312.11805
44. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In: IEEE/CVF conference on computer vision and pattern recognition; 2022.
45. Podell D, English Z, Lacey K, Blattmann A, Dockhorn T, Müller J. Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint; 2023. https://doi.org/10.48550/arXiv.2307.01952
46. Labs BF. Flux. https://github.com/black-forest-labs/flux. 2023.
47. Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M. Hierarchical text-conditional image generation with CLIP latents. arXiv preprint; 2022. p. 3. https://doi.org/arXiv:220406125
48. Betker J, Goh G, Jing L, Brooks T, Wang J, Li L. Improving image generation with better captions. OpenAI report; 2023.
49. Hou K, Zorzi M, Testolin A. Estimating the distribution of numerosity and non-numerical visual magnitudes in natural scenes using computer vision. Psychol Res. 2024;89(1):31. https://doi.org/10.1007/s00426-024-02064-2 PMID: 39625570
50. Liu S, Zeng Z, Ren T, Li F, Zhang H, Yang J. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision; 2025. p. 38–55.
51. Li J, Li D, Xiong C, Hoi S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning; 2022. p. 12888–900.
52. Liang PP, Goindani A, Chafekar T, Mathur L, Yu H, Salakhutdinov R. Hemm: Holistic evaluation of multimodal foundation models. arXiv preprint; 2024. https://doi.org/arXiv:240703418
53. Shepard RN, Kilpatric DW, Cunningham JP. The internal representation of numbers. Cogn Psychol. 1975;7(1):82–138. https://doi.org/10.1016/0010-0285(75)90006-7

54. Dehaene S. The neural basis of the Weber-Fechner law: A logarithmic mental number line. Trends Cogn Sci. 2003;7(4):145–7. https://doi.org/10.1016/s1364-6613(03)00055-x PMID: 12691758

55. Gallistel C, Gelman I I. Non-verbal numerical cognition: From reals to integers. Trends Cogn Sci. 2000;4(2):59–65. https://doi.org/10.1016/s1364-6613(99)01424-2 PMID: 10652523

56. Piazza M, Facoetti A, Trussardi AN, Berteletti I, Conte S, Lucangeli D, et al. Developmental trajectory of number acuity reveals a severe impairment in developmental dyscalculia. Cognition. 2010;116(1):33–41. https://doi.org/10.1016/j.cognition.2010.03.012 PMID: 20381023

57. Udandarao V, Prabhu A, Ghosh A, Sharma Y, Torr P, Bibi A. No "zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. In: The thirty-eighth annual conference on neural information processing systems; 2024.

58. Sharma P, Ding N, Goodman S, Soricut R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: In: Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long papers); 2018. p. 2556–65.

59. Schuhmann C, Vencu R, Beaumont R, Kaczmarczyk R, Mullis C, Katta A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint; 2021. https://doi.org/arXiv:211102114

60. Testolin A, Zou WY, McClelland JL. Numerosity discrimination in deep neural networks: Initial competence, developmental refinement and experience statistics. Dev Sci. 2020;23(5):e12940. https://doi.org/10.1111/desc.12940 PMID: 31977137

61. Liang VW, Zhang Y, Kwon Y, Yeung S, Zou JY. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. Adv Neural Inform Process Syst. 2022;35:17612–25.

62. Fu W, Dolfi S, Decarli G, Spironelli C, Zorzi M. Electrophysiological signatures of numerosity encoding in a delayed match-to-sample task. Front Hum Neurosci. 2022;15:750582. https://doi.org/10.3389/fnhum.2021.750582 PMID: 35058763

63. Burr D, Ross J. A visual sense of number. Curr Biol. 2008;18(6):425–8. https://doi.org/10.1016/j.cub.2008.02.052 PMID: 18342507

64. Harvey BM, Klein BP, Petridou N, Dumoulin SO. Topographic representation of numerosity in the human parietal cortex. Science. 2013;341(6150):1123–6. https://doi.org/10.1126/science.1239052 PMID: 24009396

65. Castaldi E, Piazza M, Dehaene S, Vignaud A, Eger E. Attentional amplification of neural codes for number independent of other quantities along the dorsal visual stream. Elife. 2019;8:e45160. https://doi.org/10.7554/eLife.45160 PMID: 31339490

66. Paul JM, van Ackooij M, Ten Cate TC, Harvey BM. Numerosity tuning in human association cortices and local image contrast representations in early visual cortex. Nat Commun. 2022;13(1):1340. https://doi.org/10.1038/s41467-022-29030-z PMID: 35292648

67. Nieder A. The neuronal code for number. Nat Rev Neurosci. 2016;17(6):366–82. https://doi.org/10.1038/nrn.2016.40 PMID: 27150407

68. Boccato T, Testolin A, Zorzi M. Learning numerosity representations with transformers: Number generation tasks and out-of-distribution generalization. Entropy (Basel). 2021;23(7):857. https://doi.org/10.3390/e23070857 PMID: 34356398

69. Mundy E, Gilmore CK. Children's mapping between symbolic and nonsymbolic representations of number. J Exp Child Psychol. 2009;103(4):490–502. https://doi.org/10.1016/j.jecp.2009.02.003 PMID: 19327782

70. Drori I, Zhang S, Shuttleworth R, Tang L, Lu A, Ke E, et al. A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level. Proc Natl Acad Sci U S A. 2022;119(32):e2123433119. https://doi.org/10.1073/pnas.2123433119 PMID: 35917350

71. Palmer A, Smith NA, Spirling A. Using proprietary language models in academic research requires explicit justification. Nat Comput Sci. 2024;4(1):2–3. https://doi.org/10.1038/s43588-023-00585-1 PMID: 38177494

72. Testolin A. The challenge of modeling the acquisition of mathematical concepts. Front Hum Neurosci. 2020;14:100. https://doi.org/10.3389/fnhum.2020.00100 PMID: 32265678

73. Kajic I, Wiles O, Albuquerque I, Bauer M, Wang S, Pont-Tuset J. Evaluating numerical reasoning in text-to-image models. arXiv preprint; 2024. https://doi.org/arXiv:240614774

74. Paiss R, Ephrat A, Tov O, Zada S, Mosseri I, Irani M. Teaching clip to count to ten. In: Proceedings of the IEEE/CVF international conference on computer vision; 2023. p. 3170–80.

75. Jeong Y, Choi Y. NuCap: A numerically aware captioning framework for improved numerical reasoning. Appl Sci. 2025;15(10):5608. https://doi.org/10.3390/app15105608

76. Kang W, Galim K, Koo HI. Counting guidance for high fidelity text-to-image synthesis. arXiv preprint arXiv:230617567; 2023. https://arxiv.org/abs/2306.17567

77. Nieder A, Miller EK. Analog numerical representations in rhesus monkeys: Evidence for parallel processing. J Cogn Neurosci. 2004;16(5):889–901. https://doi.org/10.1162/089892904970807 PMID: 15200715

78. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inform Process Syst. 2022;35:24824–37.

79. Zhan F, Yu Y, Wu R, Zhang J, Lu S, Liu L, et al. Multimodal image synthesis and editing: The generative AI era. IEEE Trans Pattern Anal Mach Intell. 2023;45(12):15098–119. https://doi.org/10.1109/TPAMI.2023.3305243 PMID: 37624713

80. Mirzadeh I, Alizadeh K, Shahrokhi H, Tuzel O, Bengio S, Farajtabar M. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. arXiv preprint; 2024. https://doi.org/arXiv:241005229

81. Rudman W, Golovanevsky M, Bar A, Palit V, LeCun Y, Eickhoff C. Forgotten polygons: Multimodal large language models are shape-blind. arXiv preprint; 2025. https://doi.org/10.48550/arXiv.250215969

82. Romera-Paredes B, Barekatain M, Novikov A, Balog M, Kumar MP, Dupont E, et al. Mathematical discoveries from program search with large language models. Nature. 2024;625(7995):468–75. https://doi.org/10.1038/s41586-023-06924-6 PMID: 38096900

83. Wang H, Fu T, Du Y, Gao W, Huang K, Liu Z, et al. Scientific discovery in the age of artificial intelligence. Nature. 2023;620(7972):47–60. https://doi.org/10.1038/s41586-023-06221-2 PMID: 37532811