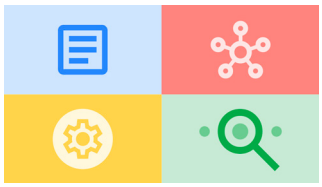


Imagen 4
Model Card

Model Cards are intended to provide developers with essential, summarized information on models, including overviews of known limitations and mitigation approaches. Model cards may be updated from time to time; for example, to include updated evaluations as the model is improved or revised.

Published: May 20, 2025

Model Information

Description: Imagen 4 is a latent diffusion model that generates high quality images from text prompts. Imagen 4 performs well in photorealistic composition settings and has improved spelling and typography, instruction following and richer colors, textures and details compared to previous Imagen models.

Inputs: The inputs consist of natural-language text strings (e.g. instructions for creating a synthetic image using a visual description) or image files.

Outputs: Outputs are generated high quality images in response to text and image inputs.

Architecture: Imagen 4 utilises latent diffusion, which is the de facto standard approach for modern image and video models, achieving high quality performance in generative media applications.

Model Data

Training Dataset: The Imagen 4 model was trained on a large dataset comprising images, text, and associated annotations.

Training Data Processing: The multi-stage safety and quality filtering process employs data cleaning and filtering methods in line with Google's commitment to advancing AI safely and responsibly. These methods include:

- **Safety and quality image filtering**: removal of unsafe, violent, or low-quality images.
- **Eliminating AI-generated images**: removal of AI-generated images prevents the model from learning artifacts or biases that may be found in AI-generated images.
- **Deduplicating images**: deduplication pipelines were utilized and similar images were down-weighted to minimize the risk of outputs overfitting training data.
- **Synthetic captions**: Synthetic captions were generated using Gemini models and allow the model to learn small details about the image.
- **Filtering captions:** filters were applied in order to minimise the presence of unsafe captions or personally identifiable information (PII).

Implementation and Sustainability

Hardware: Imagen 4 was trained using Google's Tensor Processing Units (TPUs). TPUs are specifically designed to handle the massive computations involved in training LLMs and can speed up training considerably compared to CPUs. TPUs often come with large amounts of high-bandwidth memory, allowing for the handling of large models and batch sizes during training, which can lead to better model quality. TPU Pods (large clusters of TPUs) also provide a scalable solution for handling the growing complexity of large foundation models. Training can be distributed across multiple TPU devices for faster and more efficient processing.

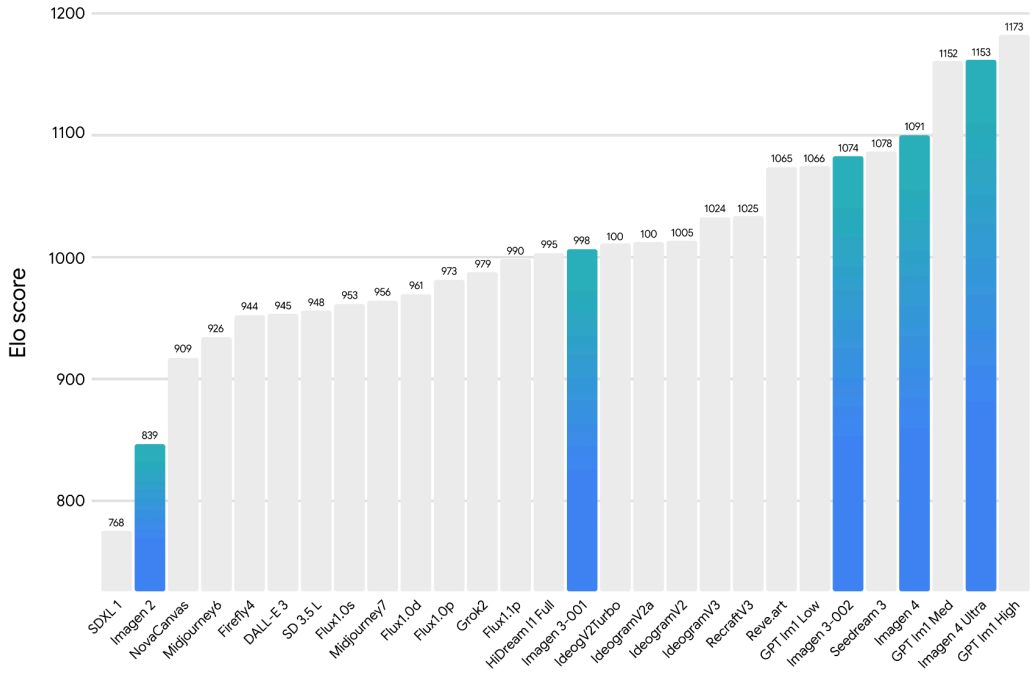
These advantages are aligned with Google's commitments to operate sustainably.

Software: Training was done using JAX, which allows researchers to take advantage of the latest generation of hardware, including TPUs, for faster and more efficient training of large models.

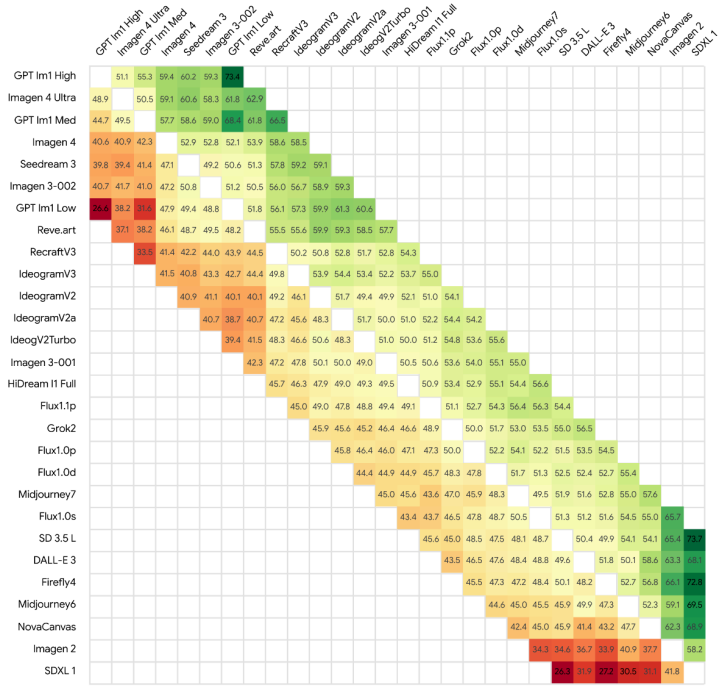
Evaluation

Approach: Human evaluations of three different quality aspects of text-to-image generation were conducted, including overall preference, prompt-image alignment and visual appeal. Automatic evaluation metrics were used to measure prompt-image alignment and image quality.

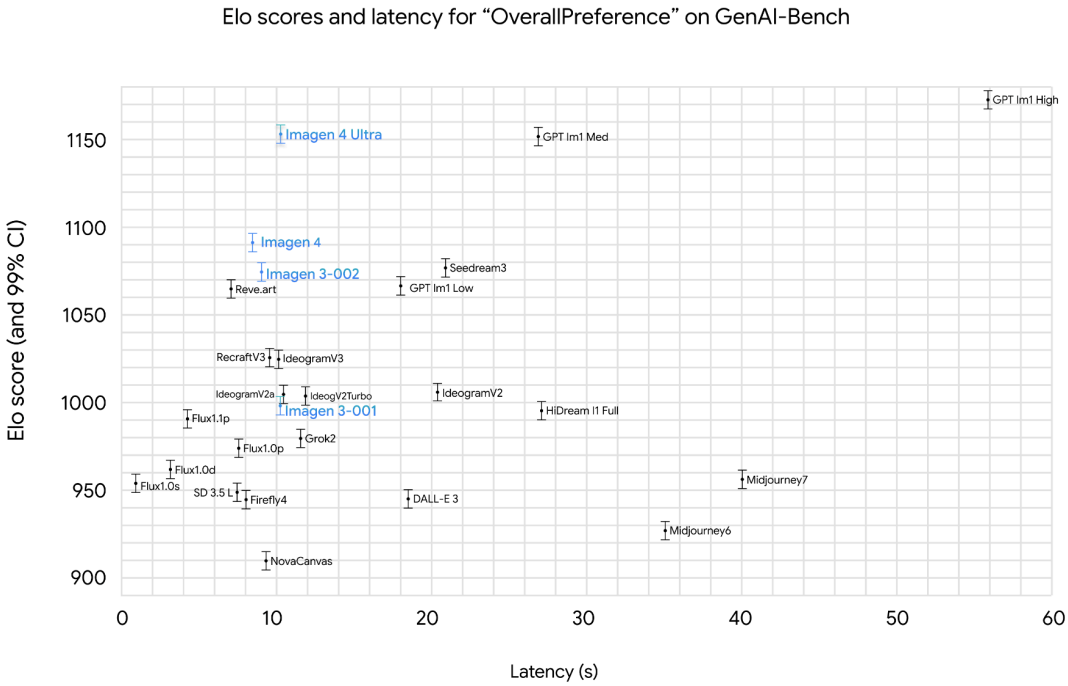
Results: Imagen 4 is Google's best text-to-image model yet. Imagen 4 scored high on human evaluations on GenAI-Bench, with one of the highest Elo scores for overall preference compared to other models. "Overall preference" is a measure of the fulfillment of the user's intent given the generated image.



Human evaluation on GenAI-Bench: Elo scores - a high overall preference benchmark for Imagen 4 vs other models.



Human evaluation on GenAI-Bench: win-rate percentages for overall preference between the pairs of models evaluated.



Human evaluation on GenAI-Bench: “Overall Preference” Elo scores plotted against latency for each model.

Latency:

- For the models with an API, we ran inference on the 1600 prompts of GenAI-Bench, recorded the 1600 observed latencies, and computed the median.
- For those models without an API, we estimated the latency on 10 generations via the respective UI and computed the median.
- For remaining models, we took the estimated latency published in https://artificialanalysis.ai/text-to-image.

In all cases, given that latencies were measured in a certain time interval and vary depending on load, location, and many other parameters; the latency observed can vary significantly.

Intended Usage and Limitations

Benefit and Intended Usage: Imagen 4 is Google's most capable image generation model to date. Imagen 4 can be used to generate high-quality, high-resolution images in a wide range of visual styles.

Known Limitations: While Imagen 4 and other current strong models achieve impressive performance, they still exhibit shortcomings in certain capabilities. In particular, tasks that require numerical reasoning, from generating an exact number of objects to reasoning about parts, are challenging for all models.

In addition, prompts that involve reasoning about scale (e.g. "the house is the same size as the cat"), compositional phrases (e.g. "one red hat and a black glass book") and actions ("a person throws a football") are the hardest across all models. This is followed by prompts that require spatial reasoning and complex language.

Responsibility and Safety

Responsibility and Safety Evaluation Approach: Imagen 4 was developed in partnership with safety and responsibility experts. A suite of evaluations was used across the end-to-end lifecycle of model development prior to release to improve models and inform decision-making. These evaluations and activities align with Google's AI Principles and responsible AI approach. The evaluations and reviews below were used for Imagen 4 at the model level, and further testing is anticipated as Imagen 4 is integrated into products:

- **Development**: Evaluations were conducted for policy violations such as violence, hate, explicit sexualization, and over-sexualization. Imagen 4 performed similar to or better than Imagen 3 across development safety evaluations.

- **Assurance:** Independently from the development team, evaluations were developed and conducted by specialized teams on content safety, including child safety, and representation.

- **Red teaming**: Red teaming was conducted by a mix of specialist internal teams and recruited internal participants throughout the model development process to inform development and assurance evaluation areas and to enable pre-launch mitigations.

- **Google Deepmind Responsibility and Safety Council (RSC):** Prior to model launches, Google DeepMind's Responsibility and Safety Council (RSC) reviews a model's performance based on the assessments and evaluations conducted throughout the lifecycle of a project to make release decisions. In addition to this process, system-level safety evaluations and reviews are conducted in the context of the specific applications in which models are deployed.

Risks: Two categories of content related risks were broadly identified:
(i) Intentional adversarial misuse of the model; and,
(ii) Unintentional model failure modes through benign use.

Mitigations: Safety and responsibility was built into Imagen 4 through pre-training and post-training mitigations following similar approaches to Gemini efforts. Pre-training mitigations included safety filtering, image deduplication, synthetic captioning, and data analysis. Post-training mitigations may include production filtering to minimize harmful outputs, and application of tools such as SynthID watermarking to reduce risks such as misinformation.