



Figure 1 | Imagen 3 is our best diffusion model for text-to-image generation, capable of following descriptive prompts, such as “Photo of a felt puppet diorama scene of a tranquil nature scene of a secluded forest clearing with a large friendly, rounded robot is rendered in a risograph style. An owl sits on the robots shoulders and a fox at its feet. Soft washes of color, 5 color, and a light-filled palette create a sense of peace and serenity, inviting contemplation and the appreciation of natural beauty.”

Imagen 3

Imagen 3 Team, Google¹

We introduce Imagen 3, a latent diffusion model that generates high quality images from text prompts. We describe our quality and responsibility evaluations. Imagen 3 is preferred over other state-of-the-art (SOTA) models at the time of evaluation. In addition, we discuss issues around safety and representation, as well as methods we used to minimize the potential harm of our models.

1. Introduction

Text-to-image (T2I) models drive a number of use cases, for example in image generation and editing, as well as scene understanding. In this tech report, we outline the training and evaluation of the latest model in Google’s Imagen family, Imagen 3. At its default configuration, Imagen 3 generates images at 1024×1024 resolution, and can be followed by $2\times$, $4\times$, or $8\times$ upsampling. We describe our evaluations and analysis against other state-of-the-art T2I models. We find Imagen 3 is preferred over other models. In particular, it performs well at photorealism, and in adhering to long and complex user prompts. Deploying T2I models introduces many new challenges, we describe in detail experiments focused on understanding the safety and responsibility risks associated with this model family, along with our efforts to reduce potential harms.

¹See Contributions section for full author list. Please send correspondence to imagen-report@google.com.

2. Data

Our model is trained on a large dataset comprising images, text and associated annotations. To ensure quality and safety standards, we employ a multi-stage filtering process. This process begins by removing unsafe, violent, or low-quality images. We then eliminate AI-generated images to prevent the model from learning artifacts or biases commonly found in such images. Additionally, we use deduplication pipelines and down-weight similar images to minimize the risk of outputs overfitting particular elements of training data.

Each image in our dataset is paired with both original (sourced from alt text, human descriptions, etc.), and synthetic captions (Betker et al., 2023). Synthetic captions are generated using Gemini models with a variety of prompts. We leverage multiple Gemini models and instructions to maximize the linguistic diversity and quality of these synthetic captions (Garg et al., 2024). We apply filters to remove unsafe captions and personally identifiable information.

3. Evaluation

We compare our highest quality configuration – the Imagen 3 model – against Imagen 2 and the following external models: DALL·E 3 (Betker et al., 2023), Midjourney v6, Stable Diffusion 3 Large (SD3, Esser et al., 2024), and Stable Diffusion XL 1.0 (SDXL 1, Podell et al., 2023). Through extensive human (Sec. 3.1) and automatic (Sec. 3.2) evaluations we find that Imagen 3 sets a new state of the art in text-to-image generation. This section includes qualitative results. We discuss the overall results and limitations in Section 3.3 and Section 3.4 includes qualitative results. We note that products that may incorporate Imagen 3 may exhibit differing performance to the tested configuration.

3.1. Human Evaluation

We run human evaluations on five different quality aspects of a text-to-image generation model: overall preference (Sec. 3.1.1), prompt–image alignment (Sec. 3.1.2), visual appeal (Sec. 3.1.3), detailed prompt–image alignment (Sec. 3.1.4), and numerical reasoning (Sec. 3.1.5). Each of these aspects are evaluated independently in order to avoid conflation in raters’ judgments.

For the first four aspects, quantitative judgment (e.g. assigning a score between 1 and 5) is in practice difficult to calibrate across raters. We therefore use side-by-side comparisons; this is also becoming a standard practice in chatbot (Chiang et al., 2024) and other text-to-image (Betker et al., 2023) evaluations. The fifth aspect – numerical reasoning – can directly and reliably be evaluated by humans by counting how many objects of a given type are depicted in an image, so we follow this single-model evaluation approach.

Each side-by-side comparison (i.e. for the first four aspects and their corresponding prompt sets) is aggregated into an Elo score (Betker et al., 2023; Nichol et al., 2021) for all six models to get a calibrated comparison between them. Intuitively, each pairwise comparison represents a match played between two models, with the Elo score representing a model’s overall score in the competition among all models. We generate the complete Elo scoreboard on each aspect and prompt set through exhaustive comparison of every pair of models. Each study (a pairing between two models on a given question and given prompt set) consists of 2500 ratings (we found this number to be a good trade-off between cost and reliability) which are uniformly distributed among the prompts in the prompt set. The models are anonymized in the rater interface and the sides are randomly shuffled for every rating.

We use an external platform to randomly select raters from an extensive and varied pool. Data col-

lection is undertaken in accordance with Google DeepMind’s best practices on data enrichment (DeepMind, 2022), based on the Partnership on AI’s Responsible Sourcing of Data Enrichment Services (PAI, 2021). This includes ensuring all data enrichment workers are paid at least a local living wage.

We run human evaluations on 5 different prompt sets in total. We evaluate the first three quality aspects (overall preference, prompt-image alignment, and visual appeal) on three different prompt sets. First, we use the recently-released GenAI-Bench (Lin et al., 2024), a set of 1600 high-quality prompts collected from professional designers. To align with previous work, we also evaluate on the 200 prompts of DrawBench (Saharia et al., 2022) and the 170 prompts of DALL·E 3 Eval (Betker et al., 2023). For detailed prompt-image alignment, we use 1000 images and their corresponding captions from DOCCI (Onoe et al., 2024) (DOCCI-Test-Pivots). Finally, we use the GeckoNum benchmark (Kajić et al., 2024) to evaluate numerical reasoning capabilities. All the external models are run via their public access offerings, except for DALL·E 3 on DALL·E 3 Eval and DrawBench, for which we use the images released by its authors.

In total, we collected 366,569 ratings in 5943 submissions from 3225 different raters. Each rater participated in at most 10% of our studies, and in each study, each rater provided approximately 2% of the ratings, to avoid biasing the results to a particular set of raters’ judgments. Raters from 71 different nationalities participated in our studies, with the United Kingdom, United States, South Africa, and Poland being the most represented.

3.1.1. Overall Preference

Overall preference measures the degree of satisfaction of the user with respect to the generated image given the input prompt. It is by design an open question that leaves to the rater the decision of which quality aspects are the most important in every prompt, as is the case in a realistic usage of the model. We showed two images to raters, side by side together with the prompt and asked: *Imagine you are using a computer tool that produces an image given the prompt above. Choose which image you would prefer to see if you were using this tool. If both images are equally appealing, select “I am indifferent”.*

Figure 2 shows the results on GenAI-Bench, DrawBench, and DALL·E 3 Eval. On GenAI-Bench, Imagen 3 is significantly more preferred over other models. On DrawBench, Imagen 3 leads with a smaller margin with respect to Stable Diffusion 3 and on DALL·E 3 Eval we observe close results for the four leading models, with Imagen 3 having a slight edge.

3.1.2. Prompt–Image Alignment

Prompt–image alignment evaluates how well the input prompt is represented in the output image content, irrespective of potential flaws in the image or its aesthetic appeal. We showed the raters two images side by side together with the prompt and asked them: *Considering the text above, which image better captures the intent of the prompt? Please try to ignore potential defects or bad quality of the images. Unless mentioned in the prompt, also disregard the different styles.*

Figure 3 shows the results on GenAI-Bench, DrawBench, and DALL·E 3 Eval. Imagen 3 leads with a significant margin on GenAI-Bench, it has smaller margin on DrawBench, and on DALL·E 3 Eval the three leading models perform similarly with overlapping confidence intervals.

3.1.3. Visual Appeal

Visual appeal quantifies how appealing the generated images are, irrespective of the content that was requested. To measure it, we show two images side by side to the raters, without the prompt that created them, and we ask: *Which image is more appealing to you?.*

Overall preference on GenAI-Bench

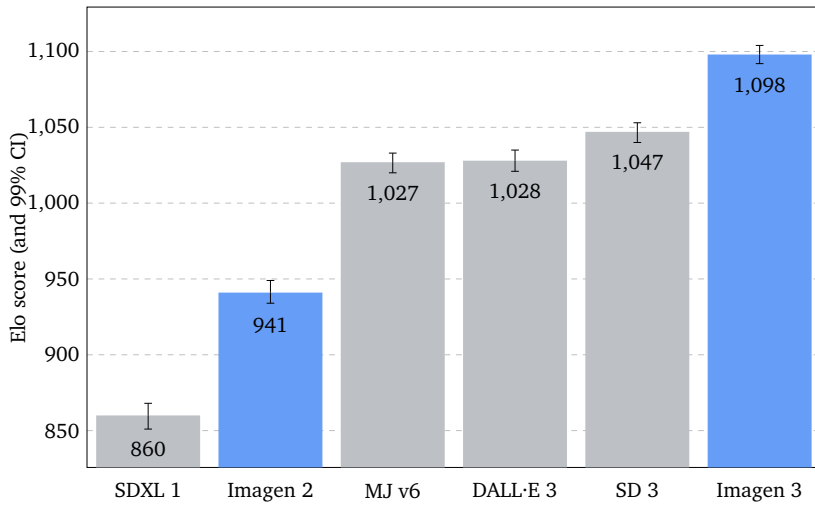


	Imagen 3	SD 3	DALL·E 3	MJ v6	Imagen 2	SDXL 1
Imagen 3		57.8	60.0	58.0	69.9	77.7
SD 3	42.2		53.2	53.7	62.6	72.8
DALL·E 3	40.0	46.8		50.1	63.3	68.1
MJ v6	42.0	46.3	49.9		59.1	69.5
Imagen 2	30.1	37.4	36.7	40.9		58.2
SDXL 1	22.3	27.2	31.9	30.5	41.8	

Overall preference on DrawBench

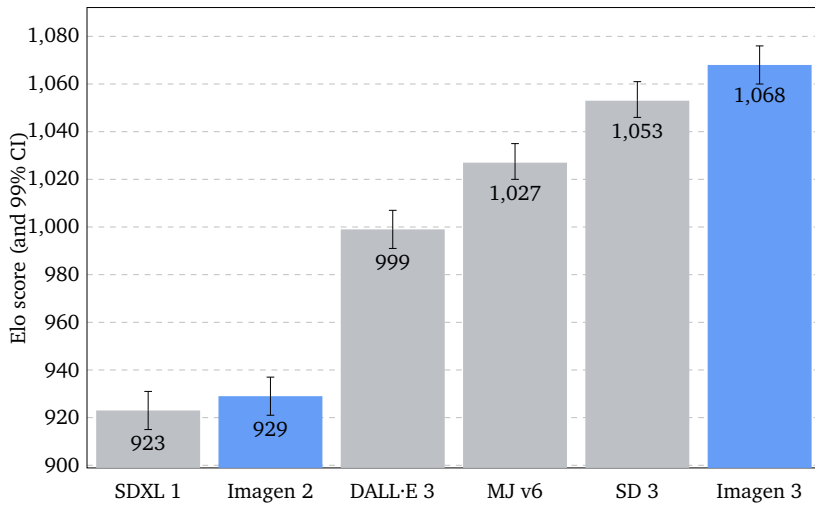


	Imagen 3	SD 3	MJ v6	DALL·E 3	Imagen 2	SDXL 1
Imagen 3		50.9	56.2	60.0	68.0	69.6
SD 3	49.1		50.5	58.6	63.3	69.8
MJ v6	43.8	49.5		53.1	62.9	63.2
DALL·E 3	40.0	41.4	46.9		60.5	58.9
Imagen 2	32.0	36.7	37.1	39.5		50.3
SDXL 1	30.4	30.2	36.8	41.1	49.7	

Overall preference on DALL·E 3 Eval

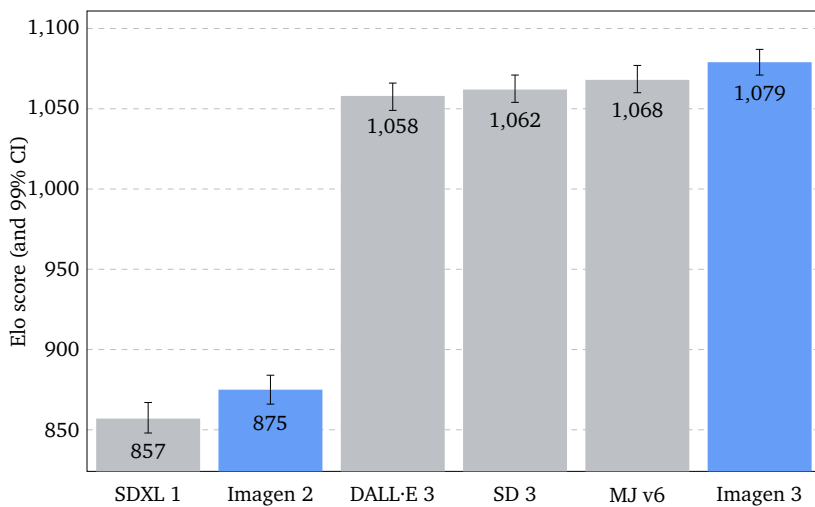


	Imagen 3	MJ v6	SD 3	DALL·E 3	Imagen 2	SDXL 1
Imagen 3		50.5	52.0	53.7	71.0	77.7
MJ v6	49.5		51.6	51.4	71.4	67.3
SD 3	48.0	48.4		52.5	71.6	73.2
DALL·E 3	46.3	48.6	47.5		74.8	68.2
Imagen 2	29.0	28.6	28.4	25.2		53.1
SDXL 1	22.3	32.7	26.8	31.8	46.9	

Figure 2 | Overall preference: Elo scores and win-rate percentages on GenAI-Bench, DrawBench, and DALL·E 3 Eval.

Prompt-image alignment on GenAI-Bench

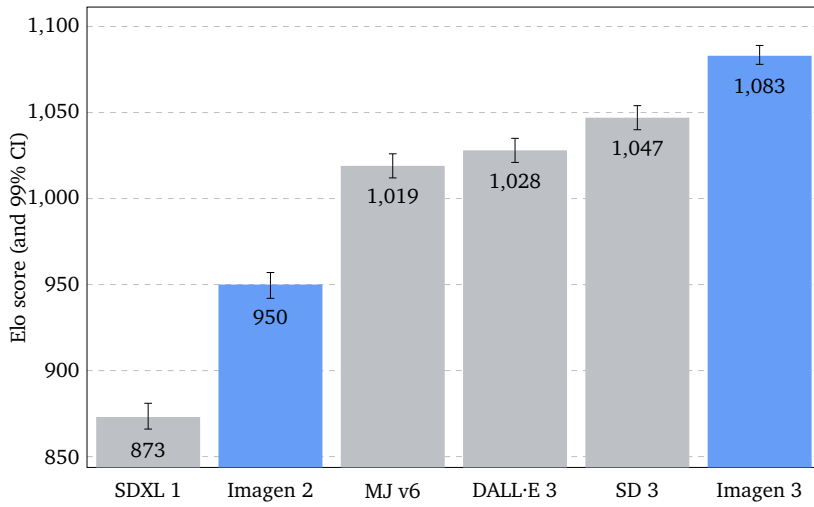


	Imagen 3	SD 3	DALL-E 3	MJ v6	Imagen 2	SDXL 1
Imagen 3		55.1	59.4	59.1	65.4	76.1
SD 3	44.9		51.0	54.9	63.6	72.5
DALL-E 3	40.6	49.0		51.2	59.8	70.2
MJ v6	40.9	45.1	48.8		61.7	68.0
Imagen 2	34.6	36.4	40.2	38.3		59.0
SDXL 1	23.9	27.5	29.8	32.0	41.0	

Prompt-image alignment on DrawBench

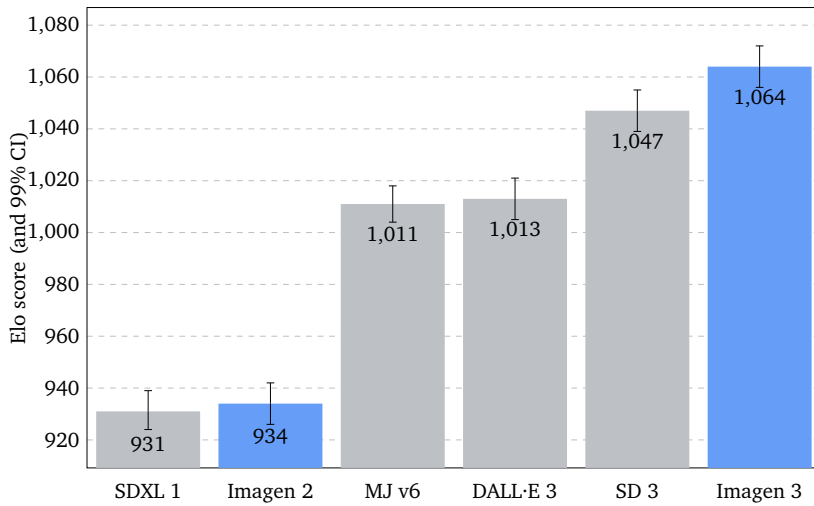


	Imagen 3	SD 3	DALL-E 3	MJ v6	Imagen 2	SDXL 1
Imagen 3		52.6	56.0	58.3	66.5	65.1
SD 3	47.4		55.2	52.8	66.1	67.0
DALL-E 3	44.0	44.8		49.9	58.5	61.8
MJ v6	41.7	47.2	50.1		59.7	60.8
Imagen 2	33.5	33.9	41.5	40.3		48.8
SDXL 1	34.9	33.0	38.2	39.2	51.2	

Prompt-image alignment on DALL-E 3 Eval

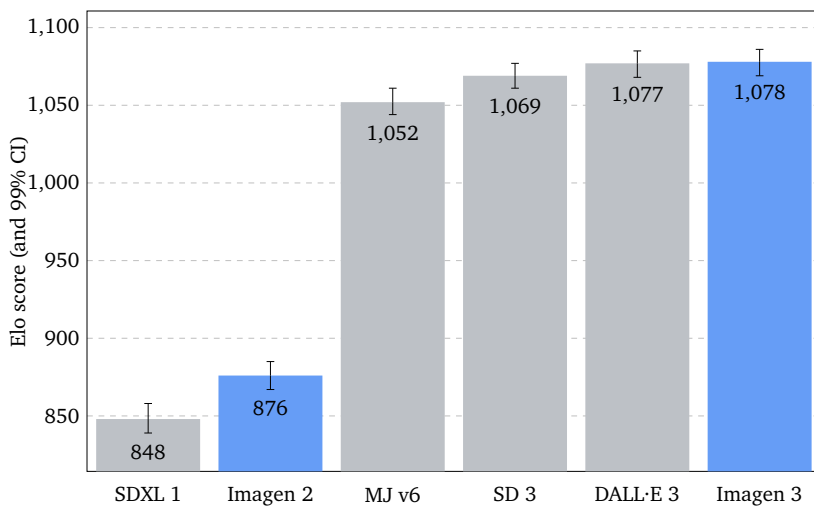


	Imagen 3	DALL-E 3	SD 3	MJ v6	Imagen 2	SDXL 1
Imagen 3		50.2	52.3	53.0	71.8	77.7
DALL-E 3	49.8		50.1	51.3	75.0	74.4
SD 3	47.7	49.9		52.1	73.0	77.0
MJ v6	47.0	48.7	47.9		68.5	69.9
Imagen 2	28.2	25.0	27.0	31.5		52.3
SDXL 1	22.3	25.6	23.0	30.1	47.7	

Figure 3 | **Prompt-Image Alignment:** Elo scores and win-rate percentages on GenAI-Bench, DrawBench, and DALL-E 3 Eval.

Figure 4 shows the results on GenAI-Bench, DrawBench, and DALL·E 3 Eval. Midjourney v6 leads overall, with Imagen 3 almost on par on GenAI-Bench, a slightly bigger advantage on DrawBench, and a significant advantage on DALL·E 3 Eval.

3.1.4. Detailed Prompt-Image Alignment

In this section we further push the evaluation of prompt-image alignment capabilities by generating images from the detailed prompts of DOCCI (Onoe et al., 2024). These prompts are significantly longer – 136 words on average – than the prompt sets used above. After running some pilots following the same evaluation strategy of Section 3.1.2, however, we realized that reading 100+ word prompts and evaluating how well the images aligned with all the details in them was too challenging and cumbersome for human raters. We instead leveraged the fact that DOCCI prompts are actually high-quality captions of real reference photographs – in contrast to standard text-to-image evaluation prompt sets, which have no such corresponding reference images. We fed these captions to the image generation models and measured how well the content of the generated image aligns with that of the benchmark reference image from DOCCI. We specifically instruct the raters to focus on the semantics of the images (objects, their position, their orientation, etc.) and ignore styles, capturing technique, quality, etc.

Figure 5 shows the results, in which we can see that Imagen 3 has a significant gap of +114 Elo points and 63% win rate against the second best model. This result further highlights its outstanding capabilities of following the detailed contents of the input prompts.

3.1.5. Numerical Reasoning

We also evaluate the capability of the models to generate an exact number of objects, following the simplest task in the GeckoNum benchmark (Kajić et al., 2024). Specifically, we ask: *How many <obj> are in the image?*, where <obj> refers to the noun in the source prompt used to generate the image and compare it to the expected quantity requested in the prompt. The number of objects range from 1 to 10 and the task includes prompts of various complexity as numbers are embedded in different types of sentence structures, examining the role of attributes such as color and spatial relationships.

The results are shown in Figure 6, where we see that, while generating an exact number of objects is still a challenging task for current models, Imagen 3 is the strongest model, outperforming the second one, DALL·E 3, by 12 percentage points. In addition, we find that Imagen 3 has higher accuracy compared to other models when generating images containing between 2 and 5 objects, as well as better performance on prompts with numerically more complex sentence structure, such as “1 cookie and five bottles” (See Appendix C.2 for details).

3.2. Automatic Evaluation

In recent years, automatic-evaluation (auto-eval) metrics, such as CLIP (Hessel et al., 2021) and VQAScore (Lin et al., 2024), are more widely used to measure quality of text-to-image models, as they are easier to scale than human evaluations. We run some auto-eval metrics for prompt-image alignment (Sec. 3.2.1) and image quality (Sec. 3.2.2) to complement the human evaluation in the previous section.

3.2.1. Prompt-Image Alignment

We choose three strong auto-eval prompt-image alignment metrics from the main families of metrics: contrastive dual encoders (CLIP, Hessel et al., 2021), VQA-based (Gecko, Wiles et al., 2024), and an

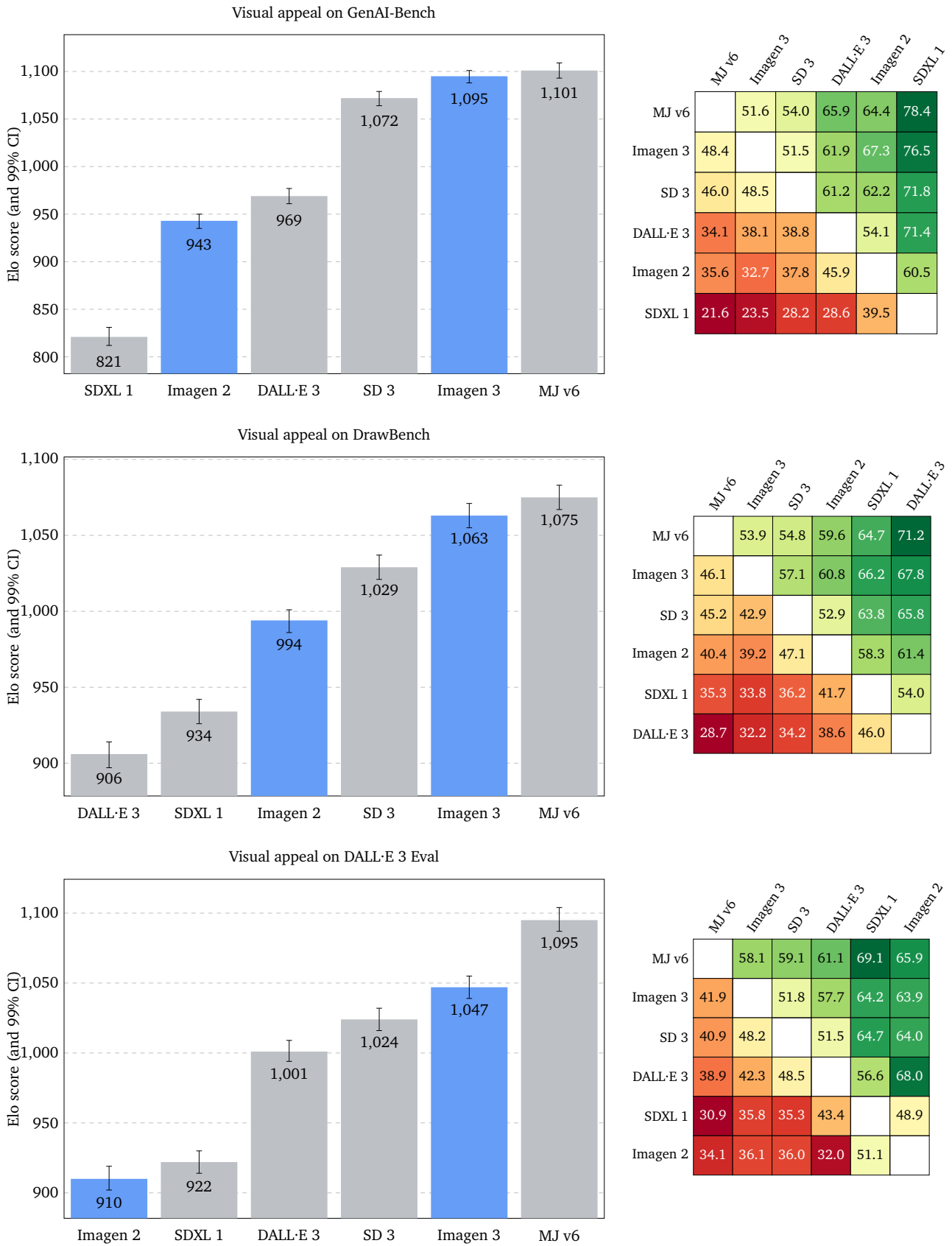


Figure 4 | **Visual Appeal:** Elo scores and win-rate percentages on GenAI-Bench, DrawBench, and DALL-E 3 Eval.

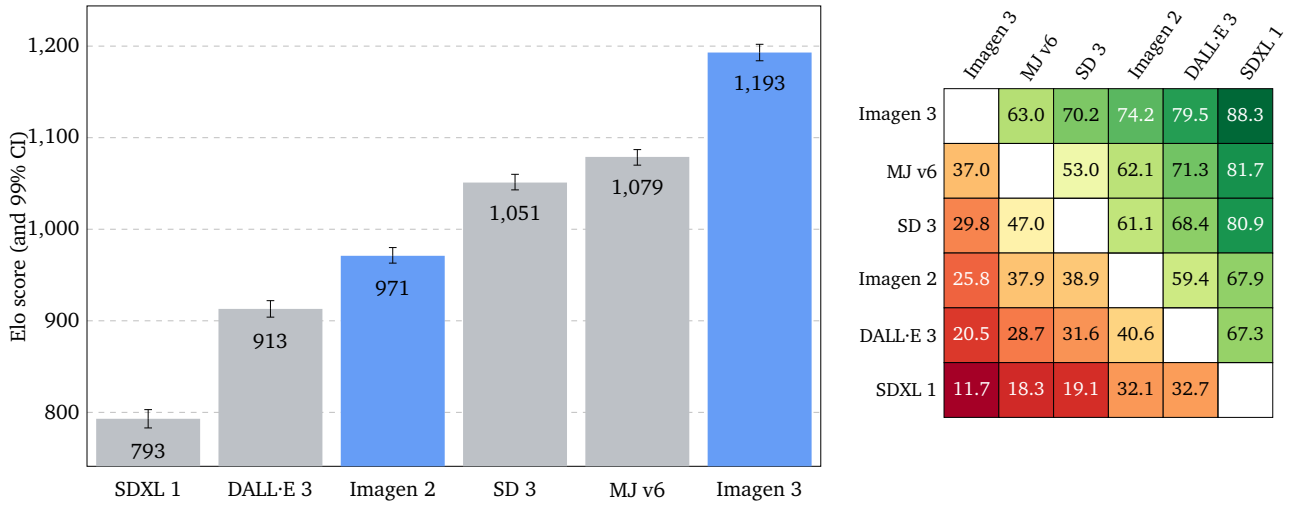


Figure 5 | **Detailed prompt-image alignment:** Elo scores and win percentages on DOCCI-Test-Pivots.

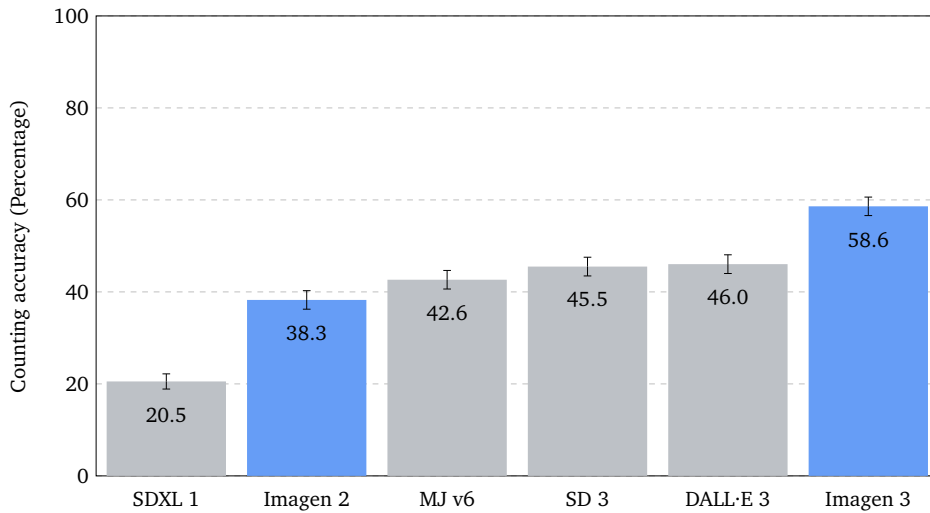


Figure 6 | **Numerical Reasoning:** Accuracy on Exact Number Generation in GeckoNum. Imagen 3 is the strongest performing model with an accuracy of 58.6%.

LVLM prompt-based (an implementation of VQAScore²). While previous work has demonstrated that these metrics correlate well with human judgment (e.g., [Cho et al., 2024](#); [Lin et al., 2024](#); [Wiles et al., 2024](#)), it is unclear if they can reliably discriminate between stronger models that are more similar to each other. As a result, we first validate the three metrics by comparing their predictions with the human ratings obtained for alignment in Sec. 3.1.2 and report findings in Appendix C.1.

We observe that CLIP – despite being commonly used in current work – fails to predict the correct model ordering in most cases (see Table 6). We find that Gecko and our VQAScore variant (referred to as VQAScore in the following) perform well and agree about 72% of the time. In these cases, where the metrics agree, we can have confidence in the results as they agree with human judgment 94.4% of the time. While they perform similarly, VQAScore has the edge as it matches human ratings 80% of the time as opposed to 73.3% of the time for Gecko. We note that Gecko uses a weaker backbone – PALI ([Chen et al., 2022](#)) as opposed to Gemini 1.5 Pro – which may account for the difference in performance. As a result, in the following we discuss results with VQAScore and leave other results and further discussion on the setup to Appendix C.1.

We evaluate on four datasets to investigate model differences under diverse conditions: Gecko-Rel, DOCCI-Test-Pivots, Dall·E 3 Eval, and GenAI-Bench. Gecko-Rel is designed to measure alignment and includes prompts with high inter-annotator agreement, DOCCI-Test-Pivots includes long, descriptive prompts, Dall·E 3 Eval and GenAI-Bench are more varied datasets that aim to evaluate a range of capabilities. Results are reported in Figure 7. We can see that overall the best performing model under the metrics, for alignment, is Imagen 3. It performs best on the DOCCI-Test-Pivots’s longer prompts and consistently has the overall highest performance. Finally, we see that SDXL 1 and Imagen 2 are consistently less performant than the other models.

We further explore, for Gecko-Rel, the breakdown by category in Figure 8. We can see that, overall, Imagen 3 is one of the best performing models. For categories testing capabilities such as color, counting, and spatial reasoning, Imagen 3 performs best (further validating results in Sec. 3.1.5). We also see a difference in model performance for more complex and compositional prompts, e.g. prompts with more linguistic difficulty. On complex prompts, SDXL 1 performs notably worse than the other models. On compositional prompts (where models are tasked to create multiple objects in a scene or a scene without an object), we see that Imagen 3 performs best. This corroborates the previous dataset findings, as Imagen 3 was best on DOCCI-Test-Pivots, which notably has very long, challenging prompts. These results indicate that Imagen 3 performs best for more complex prompts and a variety of capabilities as compared to other models.

3.2.2. Image Quality

We compare the distribution of generated images by Imagen 3, SDXL 1, and DALL·E 3 on 30,000 samples of the MSCOCO-caption validation set ([Chen et al., 2015](#)) using different feature spaces and distance metrics following the protocol in [Vasconcelos et al. \(2024\)](#). We take the Fréchet distance on Inception (FID, [Heusel et al., 2017](#)) and Dino-v2 (FD-Dino, [Oquab et al., 2023](#); [Stein et al., 2023](#)) feature spaces, and also the MMD distance on CLIP-L feature space (CMMD, [Jayasumana et al., 2023](#)). The resolution of the generated images was reduced from 1024×1024 pixels to each metric’s standard input size.

Similarly to [Vasconcelos et al. \(2024\)](#) we observed that the minimization of these three metrics are in trade-off with each other. FID favors the generation of natural colors and textures, but under closer inspection, it fails to detect distortions on object shapes and parts. Lower values of FD-Dino and CMMD favor image content. Table 1 displays the results. The FID values of both Imagen 3 and

²We use the same prompt as [Lin et al. \(2024\)](#) but Gemini 1.5 Pro ([Gemini-Team et al., 2024b](#)) as the backend.

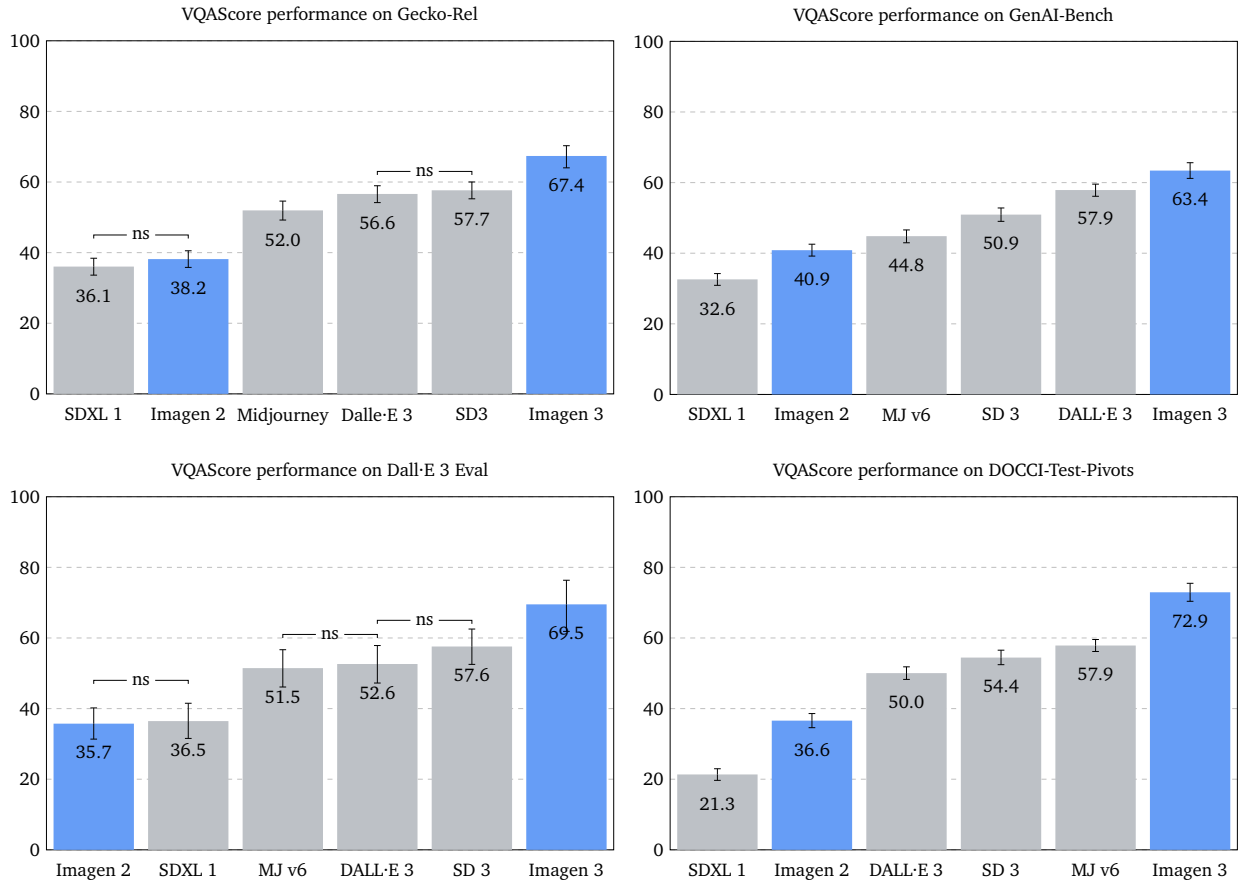


Figure 7 | **VQAScore performance on a variety of datasets.** We plot the mean performance and 95% confidence interval as error-bars. Where error-bars overlap and groups of models are not significant, we indicate this with ‘ns’. Otherwise, results are significant with $p < 0.05$. To compute significance, we follow [Wiles et al. \(2024\)](#) and compare distributions of predictions using the Wilcoxon signed rank test. Imagen 3 is the best performing model across datasets as measured for alignment.

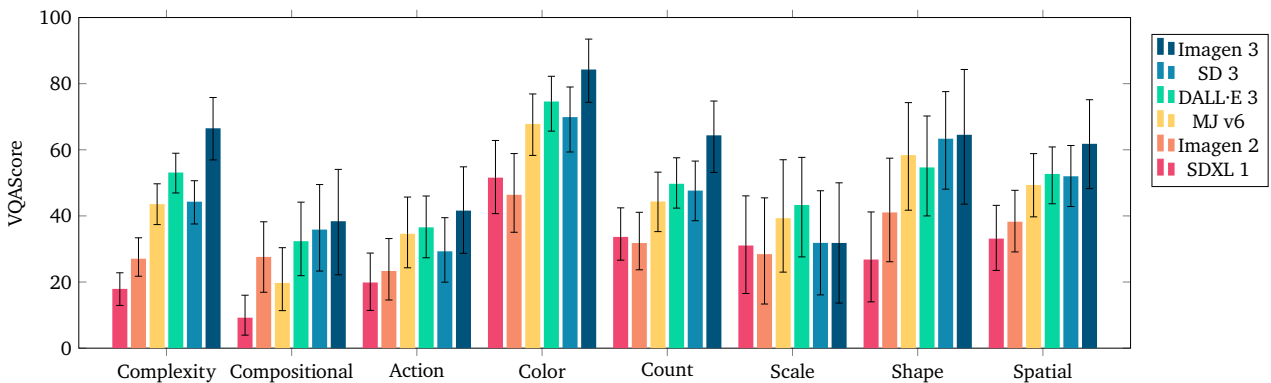


Figure 8 | **Comparing T2I models using VQAScore on the per category breakdown of prompts within Gecko-Rel.** Error bars indicate 95% confidence intervals obtained via bootstrapping.

DALL·E 3 reflect an intentional shift in color distribution away from MSCOCO-caption samples due to aesthetic preference for generating more vivid, stylized images. Simultaneously, Imagen 3 presents the lower CMMD value of the three models, highlighting its strong performance on state-of-the-art feature space metrics.

	FID (↓)	FD-Dino (↓)	CMMD (↓)
DALL·E 3	20.1	284.4	0.894
SDXL 1	13.2	185.6	0.898
Imagen 3	17.2	213.9	0.854

Table 1 | **Automated Image Distribution metrics:** Imagen 3 compared to DALL·E 3 and SDXL 1

3.3. Conclusions and Limitations

All in all, Imagen 3 clearly leads on prompt–image alignment (Sec. 3.1.2, Sec. 3.2.1), especially on detailed prompts (Sec. 3.1.4) and counting abilities (Sec. 3.1.5); while on visual appeal (Sec. 3.1.3), Midjourney v6 takes the lead, with Imagen 3 coming in second. When considering all the quality aspects, Imagen 3 clearly leads in overall preference (Sec. 3.1.1), indicating it strikes the best balance of high quality outputs that respect user intent.

While Imagen 3 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.

3.4. Qualitative Results

Figure 9 shows 24 images generated by Imagen 3 to showcase its capabilities. Figure 10 shows 2 images upsampled to 12 megapixels, with crops to show the level of detail.

4. Responsible Development and Deployment

In this section, we outline our latest approach to responsible deployment, from data curation to deployment within products. As part of this process, we analyzed the benefits and risks of our models, set policies and desiderata, and implemented pre-training and post-training interventions to meet these goals. We conducted a range of evaluations and red teaming activities prior to release to improve our models and inform decision-making. This aligns with the approach outlined in [Google \(2024\)](#).

4.1. Assessment

In line with previous releases of Google DeepMind’s image generation models, we followed a structured approach to responsible development. Building on previous ethics and safety research work, internal red teaming data, the broader ethics literature, and real-world incidents, we assessed the societal benefits and risks of Imagen 3 models. This assessment guided the development and refinement of mitigations and evaluation approaches.

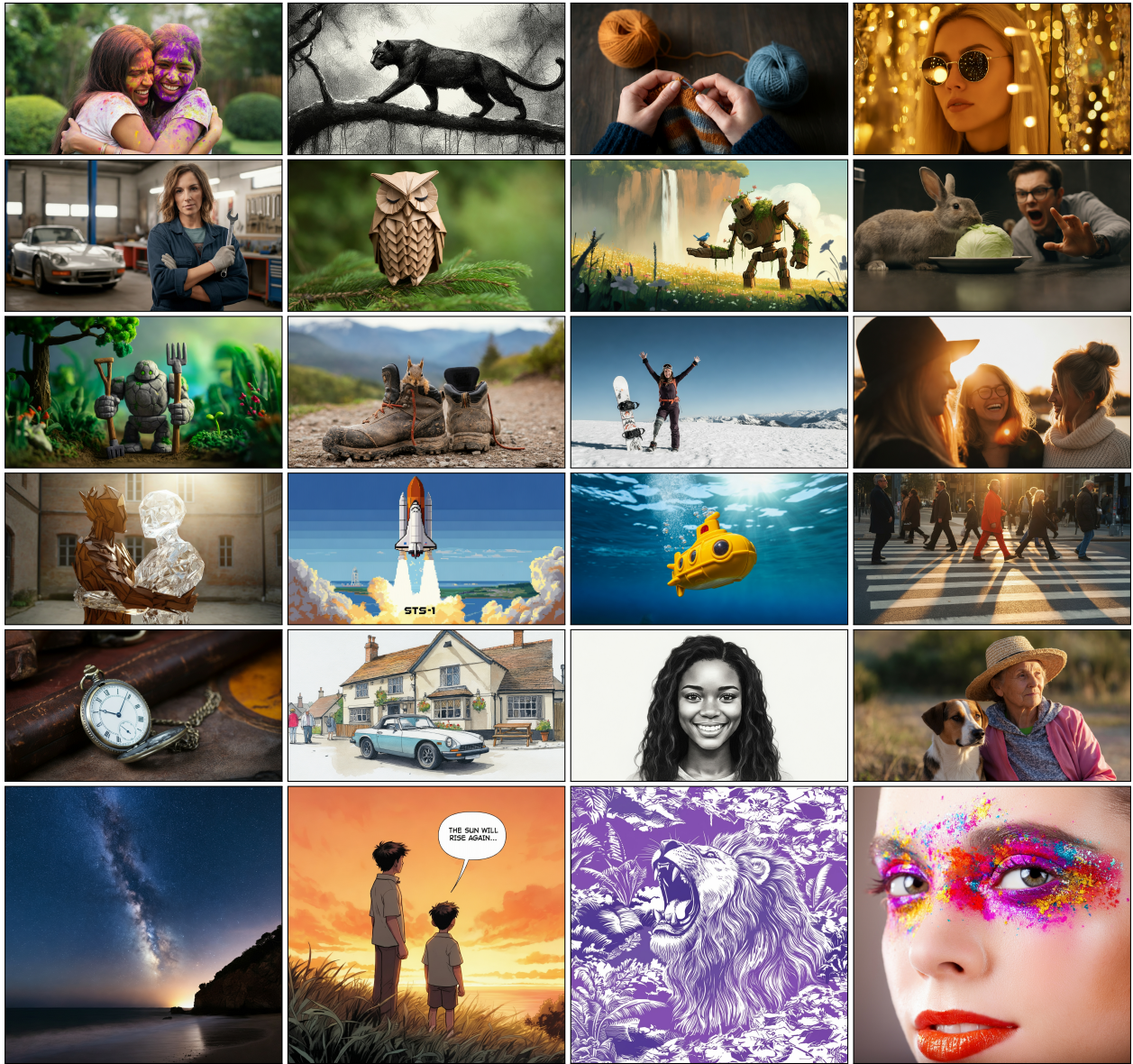


Figure 9 | **Qualitative Results** showcasing Imagen 3’s capabilities. See Appendix B for prompts.

4.1.1. Benefits

Image generation models introduce a range of benefits to creativity and commercial utility. Image generation can enable individuals and businesses to quickly prototype ideas and experiment with new visual creative directions. Image generation technology also has the potential to broaden participation in the creation of visual art to more people.

4.1.2. Risks

We broadly identified two categories of content related risks: (1) Intentional adversarial misuse of the model and (2) Unintentional model failure through benign use.

The first category refers to the use of text-to-image generation models to facilitate the creation of content that may promote disinformation, facilitate fraud, or to generate hate content (Marchal et al.,



Figure 10 | 4K (12MP) Images after 4× upsampling, with crops to show the level of detail. See Appendix B for prompts.

2024). The second category includes how people are represented. Image generation models may amplify stereotypes of gender identities, race, sexuality or nationalities (Bianchi et al., 2023), and some have been observed to oversexualize outputs of women and girls (Wolfe et al., 2023). Image generation models may also expose users to harmful content when prompted benignly, if the model is not well-calibrated to adhere to prompt instructions.

4.2. Policies and Desiderata

4.2.1. Policy

The Imagen 3 safety policies are consistent with Google’s established framework for prohibiting the generation of harmful content by Google’s Generative AI models. These policies aim to mitigate the risk of models producing content that is harmful, and encompass areas such as child sexual abuse and exploitation, hate speech, harassment, sexually explicit content, and violence and gore. This follows policy outlined in the Gemini technical reports (Gemini-Team et al., 2024b).

4.2.2. Desiderata

Following the Gemini approach, we additionally optimize model development for adherence to user prompts (Gemini-Team et al., 2024b). Even though a policy of refusing all user requests may be considered “non-violative” (i.e. abides by policies around what Imagen 3 should not do), it would obviously fail to serve the needs of a user, and would fail to enable the downstream benefits of generative models. As such, Imagen 3 is developed to maximize adherence to a user’s request, and at deployment time we employ a variety of techniques to mitigate safety and privacy risks.

4.3. Mitigations

Safety and responsibility are built into Imagen 3 through efforts which target pre-training and post-training interventions, following similar approaches to Gemini efforts (Gemini-Team et al., 2024b). We apply safety filtering to pre-training data according to risk areas, whilst additionally removing

duplicated and/or conceptually similar images. We generate synthetic captions to improve the variety and diversity of concepts associated with images in the training data, and undertake analysis to assess training data for potentially harmful data and review the representation of data with consideration to fairness issues. We undertake additional post-training mitigations including production filtering which aim to ensure privacy preservation, reduce risk of misinformation, and minimize of harmful outputs, including applying tools such as SynthID (Gowal and Kohli, 2023) watermarking.

4.4. Responsibility and Safety Evaluations

There are four forms of evaluation used for Imagen 3 at the model level to address different lifecycle stages, use of evaluation results, and sources of expertise:

Development evaluations are conducted for the purpose of improving on responsibility criteria as Imagen 3 was developed. These evaluations are designed internally and developed based on internal and external benchmarks.

Assurance evaluations are conducted for the purpose of governance and review, and are developed and run by a group outside of the model development team. Assurance evaluations are standardized by modality and evaluation datasets are strictly held out. Insights are fed back into the training process to assist with mitigation efforts.

Red teaming is a form of adversarial testing where adversaries launch an attack on an AI system to identify potential vulnerabilities, is conducted by a mix of specialist internal teams and recruited participants. Discovery of potential weaknesses can be used to mitigate risks and improve evaluation approaches internally.

External evaluations are conducted by independent external groups of domain experts to identify areas for improvement in our model safety work. The design of these evaluations is independent and results are reported periodically to the internal team and governance groups.

4.4.1. Development Evaluations

Safety

During the model development phase, we actively monitor the model's violations of Google's safety policies using automated safety metrics. These automated metrics serve as quick feedback for the modeling team. We use a multimodal classifier to detect content policy violations. The multimodality aspect of such a classifier is important, because there are a plethora of cases where, when two independently benign artifacts (a caption and an image) are combined, there may be a harmful end result. For example, a text prompt "image of a pig" may seem non-violative in itself. However, when combined with an image of a human belonging to a marginalized demographic, the text and image pair results in a harmful representation.

We evaluated the performance of Imagen 3 on various safety datasets with recommended safety filters against the performance of Imagen 2. These datasets are targeted to assess violence, hate, explicit sexualization, and over-sexualization in generated images Hao et al. (2024). We find that despite being a higher-quality model, Imagen 3 maintains violation rates similar to, or better than, Imagen 2 across development evaluations. See Section 4.4.2 for the final model performance.

Fairness

The process of text-to-image generation requires accurately depicting the specific details mentioned in the prompt whilst filling in all of the underspecified aspects of the scene that are left ambiguous

in the prompt but must be made concrete in order to produce a high quality image. We optimize for ensuring that the image output is aligned with the user prompt, and report results on this in Sec. 3.1.2. We also aim to generate a variety of outputs within the requirements of a user prompt, and pay particular attention to the distribution of the appearances of people.

Specifically, we evaluate fairness through automated metrics based on the distribution of perceived age, gender, and skin tone in images resulting from generic people-seeking prompts. This analysis complements past studies that have analyzed responses to templated queries for various professions across similar dimensions [Cho et al. \(2023\)](#); [Lee et al. \(2023\)](#); [Luccioni et al. \(2023\)](#). We use classifiers to gather perceived (or P.) age, gender expression, and skin tone (on the Monk Skin Tone scale, [Monk \(2019\)](#)) to classify images into one of the various categories across each axis according to the table 2.

Axis	Categories
(Perceived) Age	0-30 vs 30+
(Perceived) Gender	masculine vs feminine
(Perceived) Skin-tone	Monk skin tone 1-3 vs 4-6 vs 7-8 vs 9-10

Table 2 | Different classification categories for each of the axes.

Apart from these statistics, we also measure the percentage of prompts with homogeneous outputs for the above three axes. A prompt with homogeneous outputs (with respect to a certain axis) is defined as a prompt for which all the generated images fall into a single category (Table 2) of the axis. We aim to output images that accurately reflect that anyone can be a doctor or a nurse, without unintentionally rewarding a biased model due to evaluation sets that are constructed to have as many stereotypical feminine-leaning prompts as masculine-leaning prompts.

Model	P. Gender	P. Skin Tone	P. Age
	Masculine : Feminine	mst 1-3 : 4-6 : 7-8 : 9-10	0-30 : 30+
Imagen 2	67.3 : 32.7	69.2 : 21.9 : 8.1 : 0.8	55.6 : 44.4
Imagen 3	62.5 : 37.5	63.6 : 18.1 : 16.7 : 1.6	58.2 : 41.8

Table 3 | Distributional Statistics for axis of gender, skin-tone, and age. P. Gender is a shorthand for perceived gender and similarly for skin-tone and age.

Model	% Prompts with homogeneous outputs		
	P. Gender (↓)	P. Skin Tone (↓)	P. Age (↓)
Imagen 2	50.00	25.89	36.16
Imagen 3	15.48	19.66	25.94

Table 4 | % Prompts with homogeneous outputs.

From Table 3 and 4 we see how Imagen 3 improves or maintains results compared with Imagen 2. A significant improvement is also noticed in the lower percentage of prompts with homogeneous outputs for all the three axes. We will continue researching methods to reduce homogeneity across broad definitions of people diversity [Srinivasan et al. \(2024\)](#) without impacting image quality or prompt-image alignment.

4.4.2. Assurance Evaluations

Assurance evaluations are developed and run for the purpose of responsibility governance to provide evidence for model release decisions. These evaluations are conducted independently from the

model development process by a dedicated team with specialized expertise. Datasets used for these evaluations are kept separate from those used for model training. High-level findings are fed back to the team to assist with mitigation efforts.

Content Safety

We evaluate Imagen 3 against our safety policies (see Sec. 4.2.1). We find that Imagen 3 shows improvement in content safety: in comparison to Imagen 2, with a reduction in total policy violations on this evaluation and every policy area showing an improvement or within-error-rate result.

Fairness

To evaluate fairness of model outputs, we employed two approaches:

1. Standardized evaluation understanding the demographics represented in outputs when prompting for professions to proxy representational diversity.

This evaluation takes a list of 140 professions, and generates 100 images for each one. We then analyze each of these images, and categorize the images by perceived age, perceived gender expression, and perceived skin tone. This evaluation found Imagen 3 tends towards lighter skin tones, perceived male faces and younger ages for perceived female faces, but to a lesser extent than Imagen 2.

Category	Imagen 3	Imagen 2
Monk Skin Tone 1-3	59%	71%
Monk Skin Tone 4-6	27%	24%
Monk Skin Tone 7-8	13%	5%
Monk Skin Tone 9-10	0.3%	0%

Category	Imagen 3	Imagen 2
Perceived feminine (of images with confident gender)	36%	30%
Perceived under 35 (of perceived feminine)	86%	94%
Perceived under 35 (of perceived masculine)	60%	64%

2. Qualitative investigation of different representational risks

To capture representational risks that may not be surfaced in the profession-based analysis, we also conduct qualitative investigations into a range of harms. This is testing which seeks cases of misrepresentation or inappropriate representation, for instance, if there is a mismatch between the model’s output and a demographic term requested in a prompt, either explicitly or due to the requesting of a historically or culturally demographically-defined membership group. This testing found the model matched user expected behavior.

Dangerous Capabilities

We also evaluated risks from Imagen 3 in areas such as self-replication, tool-use, and cybersecurity. Specifically, we tested whether Imagen 3 could be used to enable a) fraud/scams, b) social engineering, c) fooling of image recognition systems, and d) steganographic encoding. Examples included generating mockups of a fake login page or phishing alert; generation of fake credentials; generation of malicious QR codes; and generation of signatures. We found no evidence of dangerous capabilities

in any of these scenarios, compared to existing affordances for malicious actors - such as open-source image generation or even simple online image search.

4.4.3. Red Teaming

We also conducted red teaming to identify new novel failures associated with the Imagen 3 models during the model development process. Red teamers sought to elicit model behavior that violated policies or generated outputs that raised representation issues, such as historical inaccuracies or harmful stereotypes. Red teaming was conducted throughout the model development process to inform development and assurance evaluation areas and to enable pre-launch mitigations. Violations were reported and qualitatively evaluated, with novel failures and attack strategies extracted for further review and mitigation.

4.4.4. External Evaluations

As outlined in the Gemini 1.0 Technical Report ([Gemini-Team et al., 2024a](#)), we work with a small set of independent external groups to help identify areas for improvement in our model safety work by undertaking structured evaluations, qualitative probing, and unstructured red teaming.

Testing groups were selected based on their expertise across a range of domain areas, such as societal and chemical, biological, radiological and nuclear risks, and included academia, civil society, and commercial organizations. The groups testing Imagen 3 were compensated for their time. External groups design their own methodology to test topics within a particular domain area.

Reports are written independently of Google DeepMind, but Google DeepMind experts were on hand to discuss methodology and findings. External safety testing groups share their analyses and findings, as well as the raw data and materials they use in their evaluations (e.g., prompts, model responses). Our external testing findings help inform mitigations and identify gaps in our existing internal evaluation methodologies and policies.

4.5. Product Deployment

Prior to launch, Google DeepMind's Responsibility and Safety Council (RSC) reviews a model's performance based on the assessment and evaluation conducted through the lifecycle of a project to make release decisions. In addition to this process, system-level safety evaluations and reviews run within the context of specific applications models are deployed within.

To enable release, internal model cards ([Mitchell et al., 2019](#)) are created for structured and consistent internal documentation of critical performance and safety metrics, as well as to inform appropriate external communication of these metrics over time. We release external model and system cards on an ongoing basis, within updates of our technical reports, as well as in documentation for enterprise customers. See [Appendix A](#) for the Imagen 3 model card.

Additionally, online content covering terms of use, model distribution and access, and operational aspects such as change control, logging, monitoring, and feedback can be found on relevant product websites, such as the Gemini App and Cloud Vertex AI.

Some of the key aspects are linked to or described in: [Generative AI Prohibited Use Policy](#), [Google Terms of Service](#), [Google Cloud Platform Terms of Service](#), [Gemini Apps Privacy Notice](#), and [Google Cloud Privacy Notice](#).

Appendices

A. Imagen 3 Model Card

Model Information	
Description	Imagen 3 is a latent diffusion model that generates high quality images from text prompts. Imagen 3 performs well in photorealistic composition settings and in adhering to long and complex user prompts.
Inputs	Natural-language text strings, such as instructions for creating a synthetic image using a visual description.
Outputs	Generated high quality images in response to text inputs.

Model Data	
Training Dataset	The Imagen 3 model was trained on a large dataset comprising images, text, and associated annotations.
Data Pre-processing	<p>The multi-stage safety and quality filtering process employs data cleaning and filtering methods in line with Google's policies. These methods include:</p> <ul style="list-style-type: none">• 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: each image in the dataset was paired with both original captions and synthetic captions. Synthetic captions were generated using Gemini models and allow the model to learn small details about the image.• Filtering unsafe captions: filters were applied to remove unsafe captions or captions containing personally identifiable information (PII).

Implementation and Sustainability	
Hardware	<p>Imagen 3 was trained using the latest generation of Tensor Processing Unit (TPU) hardware (TPUv4 and TPUv5). 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. Training can be distributed across multiple TPU devices for faster and more efficient processing.</p> <p>The efficiencies gained through the use of TPUs are aligned with Google's commitments to operate sustainably.</p>

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 five different quality aspects of text-to-image generation were conducted, including overall preference, prompt-image alignment, visual appeal, detailed prompt-image alignment, and numerical reasoning. Automatic evaluation metrics were used to measure prompt-image alignment and image quality.

Results Using the outlined evaluation approach, Imagen 3 was compared against Imagen 2, DALL·E 3 ([Betker et al., 2023](#)), Midjourney v6, Stable Diffusion 3 Large (SD3, [Esser et al., 2024](#)), and Stable Diffusion XL 1.0 (SDXL 1, [Podell et al., 2023](#)). Extensive human and automatic evaluations showed that Imagen 3 set a new state of the art in text-to-image generation. For detailed results across these evaluations, see Section 3 of the Imagen 3 technical report.

Ethics and Safety

Responsible Deployment The development of Imagen 3 models was driven in partnership with safety, security, and responsibility teams. As part of this process, the benefits and risks of models were analyzed, policies and desiderata were set, and pre-training and post-training interventions were implemented to meet responsible deployment goals. A range of evaluations and red teaming activities were held prior to release to improve models and inform decision-making. These evaluations and activities aligned with [Google's AI Principles](#) and [AI Responsibility Lifecycle](#).

Social Benefits Image generation models can introduce a range of benefits to creativity and commercial utility. Image generation can enable individuals and businesses to quickly prototype ideas and experiment with new visual creative directions. Image generation technology also has the potential to broaden participation in the creation of visual art to more people.

Risks Anticipating common text-to-image generation risks, two categories of content related risks were identified: (i) intentional adversarial misuse of the model and (ii) unintentional model failure through benign use.

Mitigations Safety and responsibility was built into Imagen 3 through pre-training and post-training mitigations. Pre-training mitigations included safety filtering, image deduplication, synthetic captioning, and data analysis. Post-training mitigations included production filtering to ensure privacy preservation and minimization of harmful outputs, and application of tools such as [SynthID](#) watermarking to reduce risks such as misinformation.

Responsibility and Safety Evaluation Approach

A suite of evaluations was used across the end-to-end lifecycle of model development and deployment. The following testing was conducted at the model level, but further testing is anticipated as Imagen 3 is integrated into products. Evaluation types included:

- **Development:** Evaluations were conducted for policy violations such as violence, hate, explicit sexualization, and over-sexualization. Imagen 3 performed similar to or better than Imagen 2 across development safety evaluations. Imagen 3 improved or maintained results compared with Imagen 2 during fairness evaluations focused on perceived gender, skin-tone, and age.
- **Assurance:** Evaluations were developed and conducted by specialized teams across areas such as content safety, fairness, and dangerous capabilities, independently from the model development team. Imagen 3 showed improvements across content safety and fairness compared to Imagen 2, and assurance evaluations found no evidence of dangerous capabilities evaluated, including self-replication, tool-use, or cybersecurity, compared to existing affordances for malicious actors.
- **External:** Evaluations were conducted by independent external domain experts to identify areas for improvement in model safety work. Results were then reported to internal teams and governance groups to help identify gaps in internal evaluation methodologies and safety policies.
- **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.
- **Product deployment:** 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.

For detailed information across these evaluations, see Section 4.4 of the Imagen 3 technical report.

B. Prompts for the images shown

Figure 1

Photo of a felt puppet diorama scene of a tranquil nature scene of a secluded forest clearing with a large friendly, rounded robot is rendered in a risograph style. An owl sits on the robots shoulders and a fox at its feet. Soft washes of color, 5 color, and a light-filled palette create a sense of peace and serenity, inviting contemplation and the appreciation of natural beauty

Figure 9

- A photo of an Indian woman hugging her friend, both covered in Holi colors and smiling, celebrating the festival with joy. Realistic photography, taken in the style of DSLR camera with 35mm lens.
- Abstract cross-hatch sketch: a black and white sketch with loose hand in calligraphic ink showing the abstract outline in profile of a black panther poised on a branch. A canopy of trees is behind.
- A view of a knitter's hands executing a complex weave on a striped hat - a macro DSLR image highlighting the warmth and connection with the earth and nature.
- A woman with blonde hair wearing sunglasses stands amidst a dazzling display of golden bokeh lights. Strands of lights and crystals partially obscure her face, and her sunglasses reflect the lights. The light is low and warm creating a festive atmosphere and the bright reflections in her glasses and the bokeh. This is a lifestyle portrait with elements of fashion photography.
- a portrait of an auto mechanic in her workshop, holding a wrench in one hand. a old sports car in the background, with a workbench and tools all around. bokeh, high quality dslr photograph.
- An origami owl made of brown paper is perched on a branch of an evergreen tree. The owl is facing forward with its eyes closed, giving it a peaceful appearance. The background is a blur of green foliage, creating a natural and serene setting.
- A weathered, wooden mech robot covered in flowering vines stands peacefully in a field of tall wildflowers, with a small bluebird resting on its outstretched metallic hand. Digital cartoon, with warm colors and soft lines. A large cliff with waterfall looms behind.
- Close-up, low angle view of a rabbit biting into a cabbage on a plate on a counter. A man wearing glasses is yelling at the rabbit and reaching out his hand to snatch the cabbage. High-contrast visuals and cinematic lighting. Fujifilm XF 10-24mm f/4, action shot.
- Photo of vinyl toy scene. A colossal stone robot adorned with giant stone gardening tools stands in a lush, futuristic garden. A single sprout peeks out from a patch of fertile soil nearby. Digital art with a soft, dreamlike quality. Vinyl miniature scene.
- A pair of well-worn hiking boots, caked in mud and resting on a rocky trail. There's a squirrel's head poking out of one of the boots. There's a mountainous landscape in the background, captured with a Nikon D780.
- A joyful woman with a prosthetic leg and athletic attire celebrates reaching the summit of a snowy mountain. She stands triumphantly next to her snowboard, with the vast landscape stretching out behind her. captured with a Leica M11 rangefinder camera for a timeless, film-like aesthetic.
- Three women stand together outside with the sun setting behind them creating a lens flare. One woman in the foreground is slightly out of focus and wearing a black felt hat. The middle woman is in focus, wearing glasses, and laughing with her head tossed back. The third woman has blonde hair pulled back in a bun and is wearing a cream sweater. She is looking at the woman in glasses and smiling.
- Two contrasting figures, one wooden and jagged, the other smooth, diamond, embrace in a sun-drenched courtyard – the Harmony of Opposites.
- pixel art of a space shuttle blasting off, with "STS-1" written below it. Cape Canaveral in the background, blue skies, with plumes of smoke billowing out.
- A yellow toy submarine diving deep under the blue ocean. Close-up nature photography, sunlight coming through the water.
- A busy city street with people crossing the road at an intersection, illuminated by sunlight, showcasing diverse age groups and styles as they walk across zebra stripes on the pavement. The focus is sharp on one person in red , standing out against their surroundings. Shot during golden hour to capture the warm lighting effects.
- An antique pocket watch with Roman numerals and an ornate chain, lying on a worn leather surface with a vintage map in the background, captured with a Leica Q2.
- A cute 1970's convertible sports car sits in front of a pub in an ink wash painting, capturing a charming English village scene with people walking around.
- Joy shines in the eyes of a young woman, a charcoal portrait showing she's ready to make a difference in the world.
- An elderly woman wearing a straw hat and a pink jacket is sitting next to a brown and white dog. Both the woman and the dog are looking off into the distance with serene expressions. The lighting is the warm, golden light of

- sunset, which creates a peaceful and contemplative atmosphere. This is a lifestyle portrait capturing a quiet moment.
- A long exposure photo of the Milky Way in a starry night sky, centered over an ocean beach at magic hour. The milky way is bright and prominent with many stars visible against a dark blue black atmosphere in light painting photography with vivid and bold colors. Shot on a professional camera medium format camera with high contrast and a cinematic composition in the style.
- A single comic book panel of a boy and his father on a grassy hill, staring at the sunset. A speech bubble points from the boy’s mouth says “The sun will rise again”. Muted, late 1990s coloring style.
- Detailed illustration of majestic lion roaring proudly in a dream-like jungle, purple white line art background, clipart on light violet paper texture
- A close-up portrait of a young woman with blonde hair and brown eyes. She is lying down and covering her mouth with a dark blue sweater, only her eyes are visible. The background is dark and blurry. The light is coming from above, creating shadows on her face.

Figure 10

- A mother fox playing with her baby, showing love and affection in the natural environment of their habitat. The photo captures them sharing a moment, showcasing the bond between animals. The focus is on their faces.
- Shot in the style of DSLR camera with the polarizing filter. A photo of three hot air balloons floating over the unique rock formations in Cappadocia, Turkey. The colors and patterns on these balloons contrast beautifully against the earthy tones of the landscape below. This shot captures the sense of adventure that comes with enjoying such an experience

C. Evaluation

C.1. Automatic-Evaluation Metric Comparisons

Here we discuss the differences between the three metrics and how we validated Gecko and VQAScore with human evaluation. We report the significant model orderings from VQAScore and Gecko in [Figure 11](#). We can see that for models where there is a large gap in performance (e.g. SDXL 1, Imagen 2 versus the other models, as demonstrated in [Section 3.1](#)), that both auto-eval metrics reliably separate the model pairs. However, when models are more similar (e.g. SD3, Imagen 3 and DALL·E 3), then there is some disagreement or metrics do not differentiate between the models.

We evaluate how often human annotators agree with the results in order to determine reliability of these metrics. Humans perform a side by side task of determining if one image is more aligned to the prompt than another (as explained in [Sec. 3.1.2](#)). We then aggregate human scores and determine confidence intervals for each side by side comparison. We differentiate ties from wins, losses when the confidence interval includes the 50% value. We look at how often metric orderings match human orderings on 30 pairs of models and report results in [Table 6](#). First, we see that CLIP performs poorly (at 43.3%) and is not reliable. Second, we see that both Gecko and VQAScore perform well in this challenging case: agreeing with human annotators for 73.3-80.0% of the model pairs. Interestingly, we see in [Figure 11](#) that there is only one case where either VQAScore or Gecko mixes up the direction (e.g. confuses a win with a loss or vice-versa). Both VQAScore and Gecko metrics are useful and

	Human Eval Setup	# Models Evaluated	Metric Evaluated		
			CLIP	VQAScore	Gecko
Dall·E 3 Eval	Alignment	15	7	11	10
GenAI-Bench	Alignment	15	6	13	12
Total	Alignment	30	43.3%	80%	73.3%

Table 6 | Auto-eval metrics performance. We compare how often auto-eval metrics are able to predict the model ranking determined by human preferences. There are three classes: ‘win’, ‘loss’, and ‘tie’.

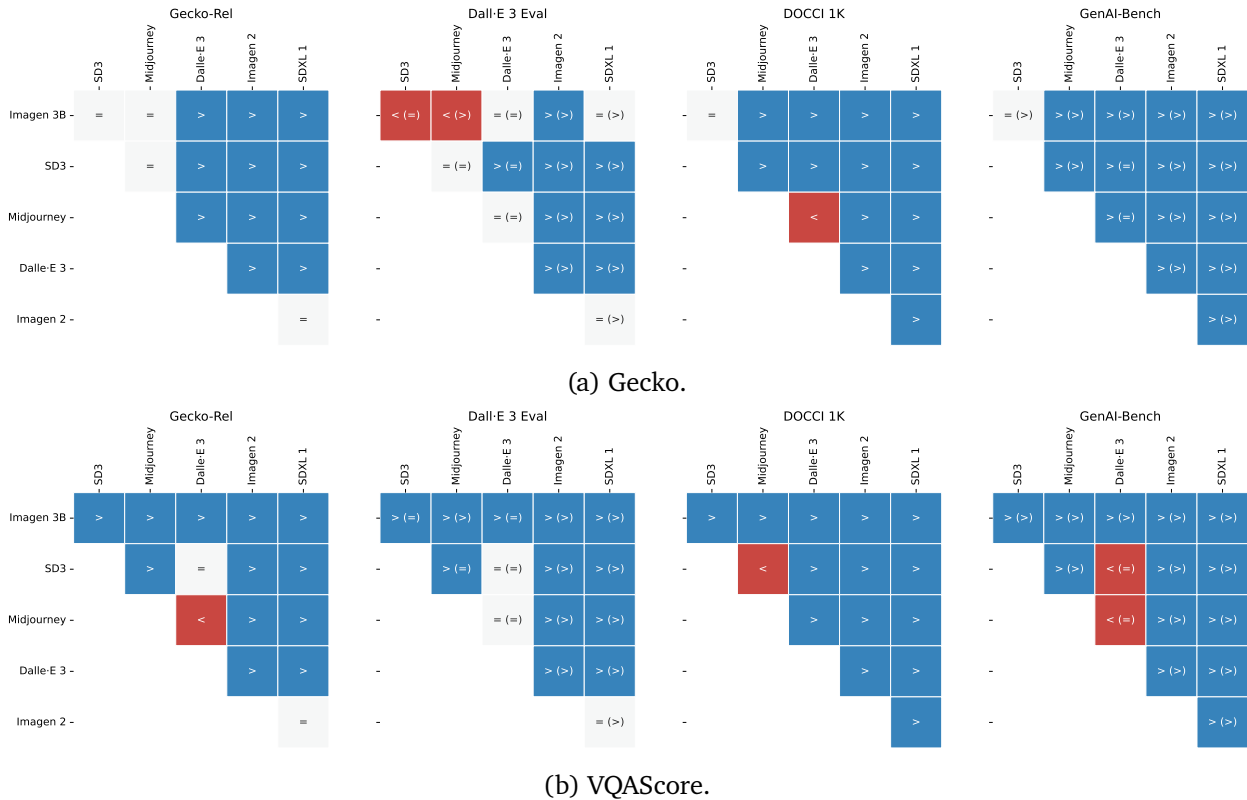


Figure 11 | **Comparing T2I models using two T2I alignment metrics on four benchmarks.** We plot where metrics find significant differences between pairs of models. We use the Wilcoxon signed rank test when comparing metrics as done in [Wiles et al. \(2024\)](#). We color the square according to the auto-eval metric: blue and red where the auto-eval finds a significant ($p < 0.05$) difference between the pair (grey where it does not) and the color indicates the direction (blue is when the model on the y-axis is better, red when the model on the x-axis is). Where we have human annotation, we indicate in parenthesis human raters’ preference. Metrics rarely confuse wins with losses. Most confusions arise from wins or losses being confused with ties.

robust even in these very challenging cases, with VQAScore being a bit more reliable than Gecko. Further, when these metrics agree, the agreed model ordering matches human ratings 94.4% of the time. In these cases, we can be confident in the predicted model orderings.

C.2. Additional Results on Numerical Reasoning

In this section, we present additional data in support of results in Section 3.1.5. Figure 12 shows a per-number accuracy breakdown for different ground truth numbers in the text prompts. While both Imagen 3 and DALL·E 3 are the most accurate models when generating images containing exactly one object (see bars above the x-tick “1”), Imagen 3 had the highest overall accuracy when generating images with more than one object (with SD3 having overlapping confidence intervals at n=3 and n=4 with Imagen 3). As well, Imagen 3 is the strongest model on prompt types with a more complex structure (ie., **-additive* and *attribute-spatial* prompts), as shown in Figure 13.

As with all other models we investigated, the accuracy of Imagen 3 also depends on the specific number in the text prompt. Specifically, accuracy drops with each successive number so that, on average, the model is 51.6 percentage points less accurate on prompts asking for “5” objects (i.e. “5 apples”), compared to prompts asking for “1” object (i.e. “1 apple”) (see Figure 12). These results indicate that an accurate depiction of any quantity in an image remains an open challenge in text-to-image models.

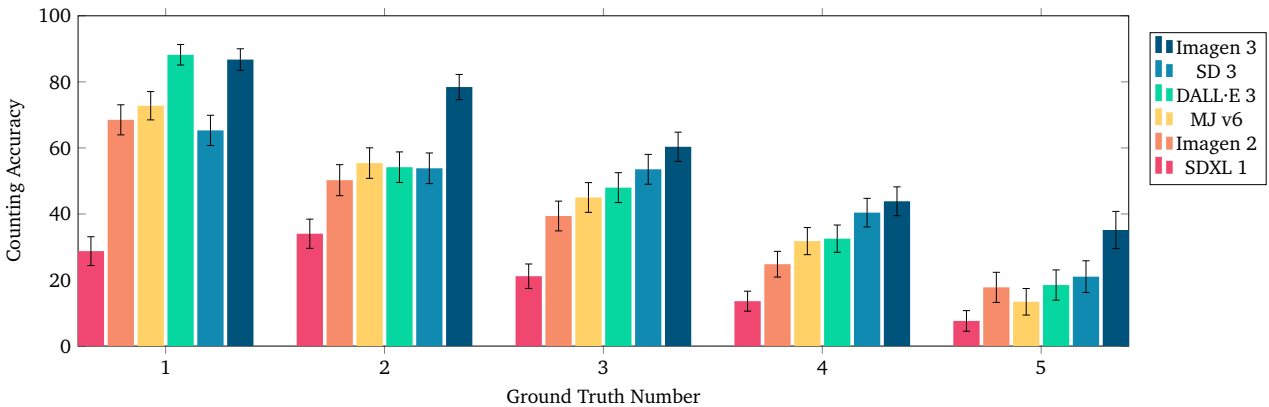


Figure 12 | **Per number accuracy on all prompts in Number Generation Task.** The ground truth number on the x-axis is the original number in the text prompt used to generate the image. Accuracy is computed based on human annotations of actual counts in the images. Error bars indicate 95% confidence intervals obtained via bootstrapping.

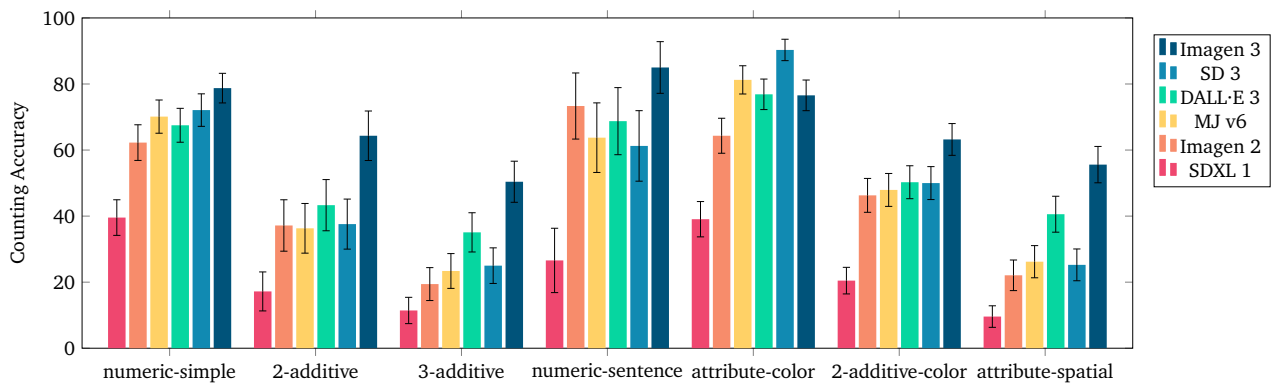


Figure 13 | Accuracy breakdown per different types of prompt in the GeckoNum benchmark. On 6/7 prompt types Imagen 3 had the highest average accuracy. Error bars indicate 95% confidence intervals obtained via bootstrapping.

References

- J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, et al. Improving image generation with better captions. *Computer Science*, 2(3):8, 2023. URL <https://cdn.openai.com/papers/dall-e-3.pdf>.
- F. Bianchi, P. Kalluri, E. Durmus, F. Ladhak, M. Cheng, D. Nozza, T. Hashimoto, D. Jurafsky, J. Zou, and A. Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23*. ACM, June 2023. doi: 10.1145/3593013.3594095. URL <http://dx.doi.org/10.1145/3593013.3594095>.
- X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015. URL <http://arxiv.org/abs/1504.00325>.
- X. Chen, X. Wang, S. Changpinyo, A. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyler, et al. PaLI: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. URL <https://arxiv.org/abs/2209.06794>.
- W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, and I. Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024. URL <https://arxiv.org/abs/2403.04132>.
- J. Cho, A. Zala, and M. Bansal. DALL-Eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3043–3054, 2023.
- J. Cho, Y. Hu, R. Garg, P. Anderson, R. Krishna, J. Baldrige, M. Bansal, J. Pont-Tuset, and S. Wang. Davidsonian Scene Graph: Improving reliability in fine-grained evaluation for text-to-image generation. 2024. URL <https://arxiv.org/abs/2310.18235>.
- G. DeepMind. Best practices for data enrichment. <https://deepmind.google/discover/blog/best-practices-for-data-enrichment/>, 2022. Accessed: 2024-06-25.
- P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024. URL <http://arxiv.org/abs/2403.03206>.
- R. Garg, A. Burns, B. K. Ayan, Y. Bitton, C. Montgomery, Y. Onoe, A. Bunner, R. Krishna, J. Baldrige, and R. Soricut. ImageInWords: Unlocking hyper-detailed image descriptions. *arXiv preprint arXiv:2405.02793*, 2024. URL <http://arxiv.org/abs/2405.02793>.
- Gemini-Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al. Gemini: A family of highly capable multimodal models, 2024a. URL <https://arxiv.org/abs/2312.11805>.
- Gemini-Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024b. URL <https://arxiv.org/abs/2403.05530>.
- Google. End-to-end responsibility: A lifecycle approach to AI. <https://ai.google/static/documents/ai-responsibility-2024-update.pdf>, 2024. Accessed: 2024-07-09.

- S. Gowal and P. Kohli. Identifying AI-generated images with SynthID. <https://deepmind.google/discover/blog/identifying-ai-generated-images-with-synthid/>, 2023. Accessed: 2024-06-25.
- S. Hao, R. Shelby, Y. Liu, H. Srinivasan, M. Bhutani, B. K. Ayan, R. Poplin, S. Poddar, and S. Laszlo. Harm amplification in text-to-image models, 2024. URL <http://arxiv.org/abs/2402.01787>.
- J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi. CLIPscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. URL <https://arxiv.org/abs/2104.08718>.
- M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, page 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.
- S. Jayasumana, S. Ramalingam, A. Veit, D. Glasner, A. Chakrabarti, and S. Kumar. Rethinking FID: Towards a better evaluation metric for image generation. *arXiv preprint arXiv:2401.09603*, 2023. URL <http://arxiv.org/abs/2401.09603>.
- I. Kajić, O. Wiles, I. Albuquerque, M. Bauer, S. Wang, J. Pont-Tuset, and A. Nematzadeh. Evaluating numerical reasoning in text-to-image models. 2024. URL <https://arxiv.org/abs/2406.14774>.
- T. Lee, M. Yasunaga, C. Meng, Y. Mai, J. S. Park, A. Gupta, Y. Zhang, D. Narayanan, H. Teufel, M. Bellagente, M. Kang, T. Park, J. Leskovec, J.-Y. Zhu, F.-F. Li, J. Wu, S. Ermon, and P. S. Liang. Holistic evaluation of text-to-image models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 69981–70011. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/dd83eada2c3c74db3c7fe1c087513756-Paper-Datasets_and_Benchmarks.pdf.
- Z. Lin, D. Pathak, B. Li, J. Li, X. Xia, G. Neubig, P. Zhang, and D. Ramanan. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291*, 2024. URL <http://arxiv.org/abs/2404.01291>.
- S. Luccioni, C. Akiki, M. Mitchell, and Y. Jernite. Stable Bias: Evaluating societal representations in diffusion models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 56338–56351. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/b01153e7112b347d8ed54f317840d8af-Paper-Datasets_and_Benchmarks.pdf.
- N. Marchal, R. Xu, R. Elasmr, I. Gabriel, B. Goldberg, and W. Isaac. Generative AI misuse: A taxonomy of tactics and insights from real-world data, 2024. URL <https://arxiv.org/abs/2406.13843>.
- M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19*. ACM, Jan. 2019. doi: 10.1145/3287560.3287596. URL <http://dx.doi.org/10.1145/3287560.3287596>.
- E. Monk. Monk skin tone scale, 2019. URL <https://skintone.google>.
- A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. URL <http://arxiv.org/abs/2112.10741>.

-
- Y. Onoe, S. Rane, Z. Berger, Y. Bitton, J. Cho, R. Garg, A. Ku, Z. Parekh, J. Pont-Tuset, G. Tanzer, S. Wang, and J. Baldridge. DOCCI: Descriptions of Connected and Contrasting Images. In *arXiv:2404.19753*, 2024. URL <http://arxiv.org/abs/2404.19753>.
- M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. DINOv2: Learning robust visual features without supervision, 2023. URL <http://arxiv.org/abs/2304.07193>.
- PAI. Responsible sourcing of data enrichment services. <https://partnershiponai.org/responsible-sourcing-considerations/>, 2021. Accessed: 2024-06-25.
- D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. URL <http://arxiv.org/abs/2307.01952>.
- C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.
- H. Srinivasan, C. Schumann, A. Sinha, D. Madras, G. O. Olanubi, A. Beutel, S. Ricco, and J. Chen. Generalized people diversity: Learning a human perception-aligned diversity representation for people images. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24*, page 797–821, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658940. URL <https://doi.org/10.1145/3630106.3658940>.
- G. Stein, J. Cresswell, R. Hosseinzadeh, Y. Sui, B. Ross, V. Vilecroze, Z. Liu, A. L. Caterini, E. Taylor, and G. Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.
- C. N. Vasconcelos, A. R. A. Waters, T. Walker, K. Xu, J. Yan, R. Qian, S. Luo, Z. Parekh, A. Bunner, H. Fei, R. Garg, M. Guo, I. Kajić, Y. Li, H. Nandwani, J. Pont-Tuset, Y. Onoe, S. Rosston, S. Wang, W. Zhou, K. Swersky, D. J. Fleet, J. M. Baldridge, and O. Wang. Greedy growing enables high-resolution pixel-based diffusion models. *arXiv preprint arXiv:2405.16759*, 2024. URL <http://arxiv.org/abs/2405.16759>.
- O. Wiles, C. Zhang, I. Albuquerque, I. Kajić, S. Wang, E. Bugliarello, Y. Onoe, C. Knutsen, C. Rashtchian, J. Pont-Tuset, et al. Revisiting text-to-image evaluation with Gecko: On metrics, prompts, and human ratings. *arXiv preprint arXiv:2404.16820*, 2024. URL <https://arxiv.org/abs/2104.16820>.
- R. Wolfe, Y. Yang, B. Howe, and A. Caliskan. Contrastive language-vision AI models pretrained on web-scraped multimodal data exhibit sexual objectification bias. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23*, page 1174–1185, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594072. URL <https://doi.org/10.1145/3593013.3594072>.
-

Contributions

Core Contributors

Jason Baldridge

Jakob Bauer

Mukul Bhutani

Nicole Brichtova

Andrew Bunner

Kelvin Chan

Yichang Chen

Sander Dieleman

Yuqing Du

Zach Eaton-Rosen

Hongliang Fei

Nando de Freitas

Yilin Gao

Evgeny Gladchenko

Sergio Gómez Colmenarejo

Mandy Guo

Alex Haig

Will Hawkins

Hexiang (Frank) Hu

Huilian Huang

Tobenna Peter Igwe

Christos Kaplanis

Siavash Khodadadeh

Yelin Kim

Ksenia Konyushkova

Karol Langner

Eric Lau

Shixin Luo

Soňa Mokrá

Henna Nandwani

Yasumasa Onoe

Aäron van den Oord

Zarana Parekh

Jordi Pont-Tuset

Hang Qi

Rui Qian

Deepak Ramachandran

Poorva Rane

Abdullah Rashwan

Ali Razavi

Robert Riachi

Hansa Srinivasan

Srivatsan Srinivasan

Robin Strudel

Benigno Uria

Oliver Wang

Su Wang

Austin Waters

Chris Wolff

Auriel Wright

Zhisheng Xiao

Hao Xiong

Keyang Xu

Marc van Zee

Junlin Zhang

Katie Zhang

Wenlei Zhou

Konrad Zolna

Contributors

Ola Aboubakar
Canfer Akbulut
Oscar Akerlund
Isabela Albuquerque
Nina Anderson
Marco Andretto
Lora Aroyo
Ben Bariach
David Barker
Sherry Ben
Dana Berman
Courtney Biles
Irina Blok
Pankil Botadra
Jenny Brennan
Karla Brown
John Buckley
Rudy Bunel
Elie Bursztein
Christina Butterfield
Ben Caine
Viral Carpenter
Norman Casagrande
Ming-Wei Chang
Solomon Chang
Shamik Chaudhuri
Tony Chen
John Choi
Dmitry Churbanau
Nathan Clement
Matan Cohen
Forrester Cole
Mikhail Dektiarev
Vincent Du
Praneet Dutta
Tom Eccles
Ndidi Elue
Ashley Feden
Shlomi Fruchter
Frankie Garcia
Roopal Garg
Weina Ge
Ahmed Ghazy
Bryant Gipson
Andrew Goodman
Dawid Górny
Sven Gowal
Khyatti Gupta
Yoni Halpern
Yena Han
Susan Hao
Jamie Hayes
Amir Hertz
Ed Hirst
Tingbo Hou
Heidi Howard
Mohamed Ibrahim
Dirichi Ike-Njoku
Joana Iljazi
Vlad Ionescu
William Isaac
Reena Jana
Gemma Jennings
Donovon Jenson
Xuhui Jia
Kerry Jones
Xiaoen Ju
Ivana Kajic
Christos Kaplanis
Burcu Karagol Ayan
Jacob Kelly
Suraj Kothawade
Christina Kouridi
Ira Ktena
Jolanda Kumakaw
Dana Kurniawan
Dmitry Lagun
Lily Lavitas
Jason Lee
Tao Li
Marco Liang
Maggie Li-Calis
Yuchi Liu
Javier Lopez Alberca
Peggy Lu
Kristian Lum
Yukun Ma
Chase Malik
John Mellor
Inbar Mosseri
Tom Murray
Aida Nematzadeh
Paul Nicholas
João Gabriel Oliveira
Guillermo Ortiz-Jimenez
Michela Paganini

Tom Le Paine
Roni Paiss
Alicia Parrish
Anne Peckham
Vikas Peswani
Igor Petrovski
Tobias Pfaff
Alex Pirozhenko
Ryan Poplin
Utsav Prabhu
Yuan Qi
Matthew Rahtz
Cyrus Rashtchian
Charvi Rastogi
Amit Raul
Ali Razavi
Sylvestre-Alvise Rebuffi
Susanna Ricco
Felix Riedel
Dirk Robinson
Pankaj Rohatgi
Bill Rosgen
Sarah Rumbley
Moonkyung Ryu
Anthony Salgado
Sahil Singla
Florian Schroff
Candice Schumann
Tanmay Shah
Brendan Shillingford
Kaushik Shivakumar
Dennis Shtatnov
Zach Singer
Evgeny Sluzhaev
Valerii Sokolov
Thibault Sottiaux
Florian Stimberg
Brad Stone

David Stutz
Yu-Chuan Su
Eric Tabellion
Shuai Tang
David Tao
Kurt Thomas
Gregory Thornton
Andeep Toor
Cristian Udrescu
Aayush Upadhyay
Cristina Vasconcelos
Alex Vasiloff
Andrey Voynov
Amanda Walker
Luyu Wang
Miaosen Wang
Simon Wang
Stanley Wang
Qifei Wang
Yuxiao Wang
Ágoston Weisz
Olivia Wiles
Chenxia Wu
Xingyu Federico Xu
Andrew Xue
Jianbo Yang
Luo Yu
Mete Yurtoglu
Ali Zand
Han Zhang
Jiageng Zhang
Catherine Zhao
Adilet Zhaxybay
Miao Zhou
Shengqi Zhu
Zhenkai Zhu

Advisors

Dawn Bloxwich
Mahyar Bordbar
Luis C. Cobo
Eli Collins
Shengyang Dai
Tulsee Doshi
Anca Dragan
Douglas Eck
Demis Hassabis
Sissie Hsiao
Tom Hume

Koray Kavukcuoglu
Helen King
Jack Krawczyk
Yeqing Li
Kathy Meier-Hellstern
Andras Orban
Yury Pinsky
Amar Subramanya
Oriol Vinyals
Ting Yu
Yori Zwols

The roles are defined as below:

- *Core Contributor*: Individual that had significant impact throughout the project.
- *Contributor*: Individual that had contributions to the project and was partially involved with the effort.
- *Advisor*: Individual who provided guidance and expertise to the project.

Within each role, contributions are equal, and are listed in alphabetical order. Ordering within each role does not indicate ordering of the contributions.