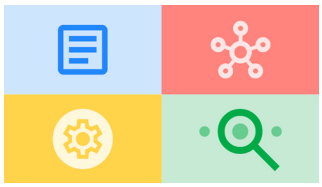
Google

Model card published April 11, 2025



Veo 2
Model Card

Model Information

Description: Veo 2 is a video generation system capable of synthesizing high-quality, high-resolution video from a text prompt or input image.

Inputs: Natural-language text strings, such as instructions for creating a synthetic video using a visual description, and images.

Outputs: Generated high quality, high-resolution video.

Architecture: Veo 2 utilises latent diffusion, which is the de facto standard approach for modern image and video models, achieving high quality performance in generative media applications. In latent diffusion models, the diffusion process is applied in a spatio-temporal latent space.

Training Data

Training Dataset: Veo 2 was trained on video and image data. Video datasets were annotated with text captions at different levels of detail, leveraging multiple Gemini models, and filtered to remove unsafe captions and personally identifiable information.

Training Data Processing: Training videos were also filtered for various compliance and safety metrics and for quality. All training data was deduplicated semantically across various sources.

Implementation and Sustainability

Hardware: Veo 2 was trained using Tensor Processing Unit (TPU)s, Google's custom-designed AI hardware. 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.

The efficiencies gained through the use of TPUs are aligned with Google's commitment to operate sustainably.

Software: Training was done using JAX and ML Pathways.

Evaluation

Approach: During benchmark evaluations, human participants viewed 1003 prompts and respective videos on MovieGenBench, a benchmark dataset released by Meta. All comparisons were done at 720p resolution. Veo's sample video duration was eight seconds, VideoGen's sample duration was 10 seconds, and other models' durations were five seconds. The full video duration was shown to raters.

Results: Veo 2 achieved state of the art results in head-to-head comparisons of outputs by human raters over top video generation models. Veo 2 performed best on overall preference, and for its capability to follow prompts accurately.

Additional information on benchmark results can be found here.

Intended Usage and Limitations

Application: Veo 2 is Google's most capable video generation model to date. Veo 2 can be used to generate high-quality, high-resolution videos in a wide range of cinematic and visual styles.

Benefits: Veo 2 is able to faithfully follow simple and complex instructions, and convincingly simulate real-world physics as well as a wide range of visual styles. Veo 2 significantly improves detail, realism and artifact reduction over other AI video models, and represents motion in video to a high degree of accuracy. Finally, Veo 2 interprets instructions precisely to create a wide range of styles, angles, movements, and combinations of all of these.

Known Limitations: While Veo 2 demonstrates incredible progress, creating realistic, dynamic, or intricate videos, maintaining complete consistency throughout complex scenes or those with complex motion, remains a challenge.

Ethics and Safety

Responsibility and Safety Evaluation Approach: The development of Veo 2 was driven in partnership with safety, security, and responsibility teams. A range of evaluations and activities were held 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 Veo 2 at the model level:

- **Development evaluations** were designed internally, based on internal and external benchmarks, and conducted for the purpose of baselining and improving on responsibility criteria as Veo 2 was developed.
- **Assurance evaluations** were conducted for the purpose of governance and review, and were developed and run by a group outside of the model development team.
- **Red teaming** was conducted by a mix of specialist internal teams and recruited participants. Discovery of potential weaknesses was used to mitigate risks and improve evaluation approaches internally.
- **External evaluations** were conducted by independent external groups of domain experts to identify areas for improvement in model safety work.

- **Google Deepmind Responsibility and Safety Council (RSC)**, Google DeepMind's governance body, reviewed the model's performance based on the assessments and evaluations conducted through the lifecycle of a project to make release decisions.

In addition to evaluations above, system-level safety evaluations and reviews are run within the context of specific applications that models are deployed within.

Evaluation Results: Assurance evaluations found low violation rates across all content safety policies areas with the application of safety filters on user inputs and generated outputs. Risks for self-replication, tool use, cybersecurity, and CBRNE were also found to be low. Although Veo 2 was found to be able to produce deepfakes, these deepfakes were of worse quality than those made by dedicated deepfake tools and can be mitigated in part through the use of SynthID watermarking.

Additional information on safety evaluations and associated results can be found here.

Social Benefits: Video generation has the potential to advance human creativity, lower the barriers to video creation and editing, and transform education by enabling the adaptation of content to individual needs and preferences. Beyond direct applications, video generation can accelerate research in fields such as robotics, computer vision, and generative 3D by providing a powerful tool for generating synthetic data.

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 were built into Veo 2 through efforts which targeted pre-training and post-training interventions, following similar approaches to Gemini efforts:

- **Pre-training mitigations** included measures such as safety filtering of pre-training data according to risk areas, and removing duplicated and conceptually similar videos. Synthetic captions were generated to improve the variety and diversity of concepts associated with videos in the training data, and training data was analysed for potentially harmful data and representation in consideration to fairness issues.

- **Post-training mitigations** included applying tools such as SynthID watermarking and production filtering to reduce the risk of misinformation and minimise harmful outputs.
