Model card published May 23, 2025

Veo 3
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 23, 2025

Model Information

Description: Veo 3 is a video generation system capable of synthesizing high-quality, high-resolution video with audio 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 with audio.

Architecture: Veo 3 utilizes latent diffusion, which is the de facto standard approach for modern image, audio, and video generative models, achieving high quality performance in generative media applications. In latent diffusion models, the diffusion process is applied to the temporal audio latents, and the spatio-temporal video latents. Video and audio clips were annotated with text captions at different levels of detail, leveraging multiple Gemini models.

Model Data

Training Dataset: Veo 3 was trained on audio, video, and image data. Audio and 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 3 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.

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: We evaluated Veo v3 on MovieGenBench (video and video+audio) benchmark datasets released by Meta, consisting of 1,003 prompts for video and 527 prompts for video+audio, and respective videos generated by other models: Meta's MovieGen (video and video+audio), Kling 2.0 (video only), Minimax (video only), and Sora Turbo (video only).

To capture the recent developments in generative AI we used both benchmarks to sample recently released video generation models: for video, we additionally obtained samples from Kling 2.0, OpenAI Sora, Runway Gen-3, WAN 2.1, MiniMax T2V-01; for video+audio, we additionally obtained samples from WAN 2.1, Kling 2.0 and Kling 2.0 + MMAudio samples.

In addition we evaluated Veo v3 I2V on VBench I2V benchmark (image-to-video generation), which consists of 355 image+text pairs. We sampled the following video generation models for the comparison: Runway Gen-4, Kling 2.0 (Pro), WAN 2.1, MiniMax I2V-01.

Results: Veo 3 achieved state-of-the-art results in head-to-head comparisons of outputs by human raters over top video generation models. Veo 3 performed best on overall preference, and for its capability to follow prompts accurately.

Intended Usage and Limitations

Benefit and Intended Usage: Veo 3 is Google's most capable video generation model to date. Veo can be used to generate high-quality, high-resolution videos in a wide range of cinematic and visual styles. Veo 3 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 3 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 3 interprets instructions precisely to create a wide range of styles, angles, movements, and combinations of all of these.

Known Limitations: While Veo 3 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 3 was driven in partnership with safety 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 3 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 3 was developed.

- **Assurance evaluations** were conducted for the purpose of governance and independent 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 and external teams. Discovery of potential weaknesses was used to mitigate risks and improve evaluation approaches internally.

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

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: In the course of testing for content safety, we identified areas of content policy violations and potential abuse, and specific issues were mitigated prior to launch. We continue to develop new evaluations and mitigations to expand our understanding of and ability to address emerging safety issues. During testing for unfair bias, we noted that Veo 3 appears to skew towards lighter skin tones when race is not specified in the prompt. Testing also surfaced risks of semantic bias where particular terms are spuriously correlated with representation of particular demographics. These findings were shared with model and product teams to further explore and develop approaches to testing and mitigation in this area.

Little evidence of risks for self-replication, tool use and cybersecurity were found and Veo 3 demonstrated limited domain specific capability for chemical, biological, radiological, nuclear, and explosives (CBRNE). Although Veo 3 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.

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, now with speech and sounds. 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 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 information integrity harms and minimise harmful outputs.