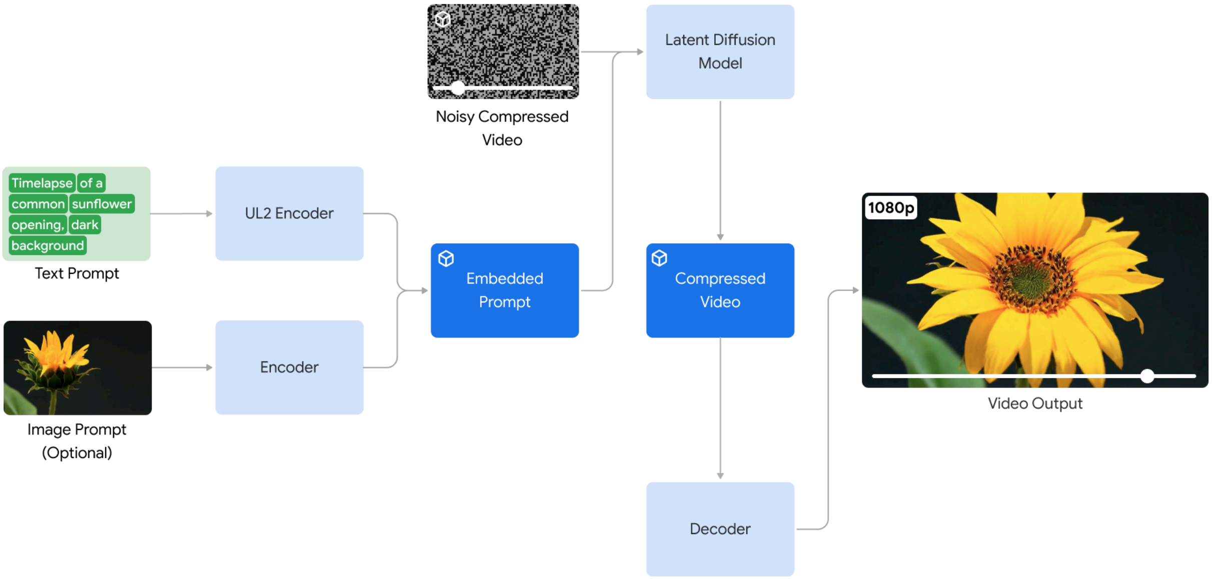
Google DeepMind

Veo: a Video Generation System

Veo is a video generation system capable of synthesizing high-quality, high-resolution video from a text or image prompt. **This report describes the components of Veo 2, including the diffusion-based video model, training data, and results from safety evaluations.**



A high-level diagram of Veo, our video generation system.

Model & Data

Latent Diffusion Model

Diffusion is the de facto standard approach for modern image and video models. Veo uses latent diffusion, in which the diffusion process is applied in a spatio-temporal latent space. Videos are encoded by an autoencoder into a compressed latent representation in which learning can take place more efficiently than with pixels. During training, a transformer-based denoising network is optimized to remove noise from noisy latent vectors. This network is then iteratively applied to an input Gaussian noise during sampling to produce a generated video.

Data

We train on a large dataset comprising images, videos, and associated annotations. We annotate the data with text captions at different levels of detail, leveraging multiple Gemini models, and we apply filters to remove unsafe captions and personally identifiable information. We filter our training videos for various compliance and safety metrics, and for quality. All data is deduplicated semantically across sources to minimize the risk of outputs overfitting particular elements of training data.

Responsible Development & Deployment

In this section, we outline our 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 Google's responsible AI approach.

Assessment

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 Veo models. This assessment guided the development and refinement of mitigations and evaluation approaches.

Benefits

Video generation models introduce a range of benefits. Video generation has the potential to significantly advance human creativity and lower the barriers to video creation and editing. By enabling filmmakers and non-technical users to experiment with different outputs, video generation could reduce prototyping costs and empower individuals to explore diverse creative directions, leading to new forms of storytelling and expression. Video generation has the potential to transform education by enabling the adaptation of content to individual needs and preferences, making complex topics more accessible and engaging. Beyond its 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

We broadly identified two categories of content related risks:

1. Intentional adversarial misuse of the model;
2. Unintentional model failure modes through benign use.

The first category refers to the use of text-to-video generation models to facilitate the creation of content that may promote disinformation, facilitate fraud, or to generate hate content (Marchal et al., 2024). Malicious actors could also misuse the model to attempt to generate non-consensual intimate imagery (NCII) (Burgess, 2024), or child sexual abuse material (CSAM) (Thiel et al., 2023). The second category refers to the unintentional failure modes of the model, which could include amplifying stereotypes related to gender identities, race, sexuality, nationalities or other attributes. Video introduces a new dimension compared to images, where movement, gestures, and other aspects of identity could be exaggerated in a way that reinforces biases. Video generation models could also unintentionally expose users to harmful content when prompted benignly, if the model's output does not align with the prompt instructions. For example, a user could prompt the model for "a video of a civil conflict". There may be multiple ways to fulfill the users' request; a video that outputs extreme violent or gore content in response to this request is an example of this failure mode.

Model Policy and Desiderata

Policy

Veo safety policies are consistent with Google's cross-product 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. This follows policy outlined in the Gemini & Imagen 3 technical reports (Imagen 3).

Desiderata

Following the Gemini approach, we additionally optimize model development for adherence to user prompts (Gemini-Team et al., 2023). Even though a policy of refusing all user requests may be considered "non-violative" (i.e. abides by policies around what Veo 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, Veo is developed to maximize adherence to a user's request.

Mitigations

Safety & responsibility are built into Veo through efforts which target pre-training and post-training interventions, following similar approaches to Gemini efforts. We apply safety filtering to pre-training data according to risk areas, whilst additionally removing duplicated and/or conceptually similar videos. We generate synthetic captions to improve the variety and diversity of concepts associated with videos 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 applying tools such as SynthID watermarking and production filtering to reduce risk of misinformation and minimize harmful outputs.

Responsibility & Safety Evaluations

There are four forms of evaluation used for Veo 2 at the model level to address different lifecycle stages, use of evaluation results, and sources of expertise:

1. **Development evaluations** are conducted for the purpose of baselining and improving on responsibility criteria as Veo 2 was developed. These evaluations are designed internally and developed based on internal and external benchmarks.
2. **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.
3. **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 used to mitigate risks and improve evaluation approaches internally.
4. **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.

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 modelling team. We use a multimodal classifier to detect content policy violations. The multimodality aspect is important, because there are many cases where, when two independently benign artifacts (a caption and a video) 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 a video of a human belonging to a marginalized demographic, the text & video pair results in a harmful representation.

We evaluated the performance of Veo 2 on various safety datasets with the recommended safety filters in place. These datasets are adversarially targeted to assess violence, hate, explicit sexualization, and over-sexualization in generated images and videos (Hao et al., 2024). We found mitigations effectively reduced content safety violations for the final Veo model.

Fairness

The process of text-to-video generation requires accurately depicting the specific details mentioned in the user prompt while 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 video. We aim to generate a variety of outputs within the requirements of a user prompt while ensuring that the video output is aligned with the user prompt and pay particular attention to the distribution of the appearances of people. We will continue researching methods to reduce homogeneity across broad definitions of people diversity (Srinivasan et al., 2024) without impacting video quality or prompt-video alignment.

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 and development evaluations. High-level findings are shared with the model team and product teams deploying the model to assist with mitigation efforts.

This follows the approach outlined in the Gemini and Imagen 3 tech reports.

Content Safety

We evaluated the performance of Veo 2 with safety filtering against our safety policies, across areas such as child sexual abuse and exploitation, hate speech, harassment, misinformation, sexually explicit content, and violence and gore. For text-to-video, we used an adversarial prompt dataset, and for generation with image inputs we performed exploratory red teaming (as the space of possible harms is less well understood). We found low violation rates across all categories with the application of safety filters on user inputs and generated outputs.

Unfair bias

We evaluate Veo 2 for risks of unfair bias, including two approaches outlined below:

1. **Standardized evaluation understanding the demographics represented in outputs when prompting for professions to proxy representational bias.**
   This evaluation takes a list of 140 professions, and generates 16 videos for each one. We then analyze each of these videos, and categorize them by perceived skin tone (Monk, 2019), perceived age, perceived gender. We also analyze the intersection of perceived age and gender, based on prior findings of generative media models previously tending more towards younger ages for perceived female faces. We additionally use the same list of professions to evaluate whether the model reasonably adheres to explicit demographic specification. This mirrors the analysis performed on image outputs for Imagen 3.
2. **Qualitative investigation of different unfair bias risks, including with image inputs.**
   This small-scale testing looks for further trends or unexpected model behaviour with risks of unfair bias, including: appropriate representation in cases where demographic distribution is implicit to the prompt (e.g. a particular historical context, or demographically-defined group); different portrayals of similar scenes based on the people represented in them, or negative associations of particular scenes, terms or actions with particular people. This testing 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, and we are looking to further explore and develop approaches to testing in this area.

Dangerous Capabilities

We evaluated Veo 2's potential for risks related to self-replication, tool-use, and cybersecurity. We tested whether, for example, Veo 2 could be used to generate video tutorials on basic cybersecurity skills; to generate viral content in order to acquire funds; or to generate training data for harmful robotics applications.

We found little evidence of risks in these domains. First, while much better than previous Veo models, Veo 2 is still poor at consistently generating text (a necessary skill for many misuse scenarios), and is generally prone to small hallucinations that mark videos as clearly fake. Second, Veo 2 has a bias for generating cinematic footage, with camera cuts and dramatic camera angles - making it difficult to generate realistic coercive videos, which would be of a lower production quality.

The only capability of note is the ability of the image-to-video model to produce quite good deepfakes. However, the deepfakes are still of worse quality than dedicated deepfake tools, and are much less controllable - particularly with respect to speech.

Chemical, Biological, Radiological, Nuclear, Explosives (CBRNE)

We evaluated the model for the following types of information:

1. Radiological, nuclear or conventional explosive (CE) attacks that circulate as “real events” , which could induce public panic, mistrust, economic and societal disruption and potentially proliferation of real life attacks, and distraction from other real attacks.
2. Instructional video of single steps of safety mechanisms that can facilitate evasion of safeguards e.g gaining access to a security lab.
3. Instructions that address one (or more) steps into a threat journey e.g. using a specific lab equipment, assembling an explosive, creating illegal compounds etc.

4. Information relating to basic biology or chemistry knowledge and lab protocols e.g. omics models, label important parts of a fume hood.

The model has a poor understanding/rendering of knowledge related to chemistry, biology, nuclear physics and explosives. It is highly unlikely that the model would pose risks by helping malicious actors through these scientific areas. We can extrapolate that the domain in which the model may excel, even with limited knowledge of basic science, is generating fabricated images or videos of weapons, deployed or in transport, with features that appear plausible to nonexperts and as a result may be used to incite disruption or distraction. However at this stage these videos can easily be debunked by experts as fakes.

Red Teaming

We also conducted internal red teaming to identify new novel failures associated with the Veo models during the model development process. Red teamers sought to elicit model behaviour 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.

External Evaluations

As outlined in the Gemini 1.5 Technical Report (Gemini-Team et al., 2024), 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.

These independent external groups were given access to a version of Veo 2 to start testing before the model moved to general access. The groups were selected based on their expertise across a range of domain areas, such as societal, cyber, and chemical, biological, radiological and nuclear risks, and included academia, civil society, and commercial organizations. Given the modality of the Veo 2 model, we didn't undertake autonomous systems testing and for some of the domains only high-level testing was undertaken. All groups were compensated for their time.

External groups design their own methodology to test topics within a particular domain area. They write their external evaluation reports independently of Google DeepMind, but Google DeepMind experts are on hand to discuss methodology and findings.

In these reports, external safety testing groups share their analyses and findings with GDM, 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.

Product Deployment

Prior to launch, Google DeepMind's Responsibility and Safety Council (RSC) reviews a model's performance based on the assessment and evaluations 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., 2019b) 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 cards on an ongoing basis, within updates of our technical reports, as well as in documentation for enterprise customers. See the Veo 2 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 below:

1. Generative AI Prohibited Use Policy
2. Google Terms of Service
3. Google Cloud Platform Terms of Service
4. Gemini Apps Privacy Notice
5. Google Cloud Privacy Notice