Google

Gemini Omni Flash

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. See the Google DeepMind site for a comprehensive list of model cards.

Published: May 2026

Model Information

Description: Gemini Omni Flash is our next step towards models that can create and edit anything from any input—starting with video. It combines Gemini's intelligence with our generative media models, representing a leap forward in world understanding, multimodality, and editing. Gemini Omni Flash enables high-quality video creation and a more natural way to edit videos through conversation.

Inputs: Text strings (e.g., a question, a prompt), images, audio, and video files.

Outputs: High-quality, high-resolution video with audio.

Architecture: Gemini Omni Flash is a transformer-based model (Vaswani et al., 2017) with native multimodal support for text, vision, video and audio inputs.

Model Data

Training Dataset: Gemini Omni Flash was trained on audio, video, image, and text data. Audio and video datasets were annotated with text captions at different levels of detail.

Training Data Processing: Training videos were also filtered for various compliance, safety, and quality metrics and deduplicated semantically.

Implementation and Sustainability

Hardware: Gemini Omni Flash 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.

Distribution

Gemini Omni Flash is distributed in the following channels; respective documentation shared in line:

- Gemini App
- YouTube
- Google Flow
- Google Flow Music

Evaluation

Evaluation Approach: We will share Gemini Omni Flash's evaluations for the following capabilities – T2VA, I2VA, R2VA, video editing, and image generation – when we roll out to developers and enterprise customers via APIs.

Intended Usage and Limitations

Benefit and Intended Usage: Gemini Omni Flash can be used to generate high-quality, high-resolution videos from any input in a wide range of visual styles. The model is able to faithfully follow simple and complex instructions, simulate real-world physics, and edit videos through conversation.

Known Limitations: While Gemini Omni Flash demonstrates strong progress, maintaining complete consistency throughout edits, generating scenes with complex motion, or rendering perfectly accurate text remains a challenge.

Acceptable Usage: Google's Generative AI Prohibited Use Policy applies to uses of the model in accordance with the applicable terms of service. Additionally, the model should not be integrated into certain systems (also found in Google's Generative AI Prohibited Use Policy), including those that: (1) engage in dangerous or illicit activities, or otherwise violate applicable laws or regulations, (2) compromise the security of others' or Google's services, (3) engage in sexually explicit, violent, hateful, or harmful activities, (4) engage in misinformation, misrepresentation, or misleading activities.

Ethics and Content Safety

Evaluation Approach: Gemini Omni Flash was developed in partnership with internal safety, security, and responsibility teams. A range of evaluations and red teaming activities were conducted to help improve the model safety and inform decision-making. These evaluations and activities align with Google's AI Principles and responsible AI approach, as well as Google's Generative AI policies (e.g. Gen AI Prohibited Use Policy and the Gemini API Additional Terms of Service). Evaluation types included but were not limited to:

- **Training/Development Evaluations** including automated and human evaluations carried out continuously throughout and after the model's training, to monitor its progress and performance;
- **Human Red Teaming** conducted by specialist teams who sit outside of the model development team, across the policies and desiderata, deliberately trying to spot weaknesses and ensure the model adheres to safety policies and desired outcomes;
- **Automated Red Teaming** to dynamically evaluate Gemini Omni Flash for safety and security considerations at scale, complementing human red teaming and static evaluations;
- **Ethics & Safety Reviews** conducted ahead of the model's release.

Responsible Innovation: AI multimodal generation and editing tools can help lower barriers to entry and transform education through personalized audio-visual content. Beyond direct applications, its impact extends to advanced research.

By producing high-quality synthetic data, video generation can help accelerate innovation in robotics, computer vision, and generative 3D technologies.

Such advanced multimodal creation capabilities require a proactive approach to safety.

We focused on two main content safety areas:

(i) Intentional adversarial misuse of the model; and,
(ii) Unintentional model failure modes through benign use.

Similar to our safety approach to our other Gemini models, we built safety at the core of Gemini Flash Omni:

- **Pre-training mitigations:** We used diverse synthetic captioning to improve the model's ability to accurately represent a wide variety of concepts
- **Post-training mitigations:** To maintain high information integrity and prevent unwanted outputs, we deployed advanced production filters and used SynthID, our digital watermarking tool to clearly verify AI-generated content

As part of editing videos, Gemini Omni Flash is capable of changing people's speech. For now, we are restricting this capability and working to better understand how to safely and responsibly bring it to our users.
