

Introducing agentic video understanding with Gemini

1.9.2026 - Rohan Doshi | Google Italia

Our new agentic feature for video analysis cuts token consumption by up to 88%, reduces costs by up to 66%, and boosts quality by up to 7%.

Today, we're launching agentic video understanding across our latest models: Gemini 3.7 Flash, 3.6 Flash and 3.5 Flash-Lite. This new capability improves accuracy while dramatically reducing token usage and costs for video analysis. Similar to agentic vision, which combines code execution with Gemini models' native image understanding, agentic video understanding uses Gemini's native video tools to improve performance and unlock new capabilities for video processing like sub-second moment retrieval, more accurate anomaly detection, precise counting and more.

The feature is available today for video uploads and YouTube videos via the Gemini API in Google AI Studio and the Gemini Enterprise Agent Platform.

Benchmarks

Unlike current 'static' processing, where the model ingests the video at a fixed frames-per-second rate (default 1 FPS, adjustable via API), agentic video understanding pairs the model's core reasoning with native video tools to dynamically search, scan, and inspect target video segments across visual frames, audio, and transcripts. Across standard video analysis benchmarks, Gemini models with agentic video understanding reduce analysis costs by up to 66% and token consumption by up to 88%, while improving accuracy by up to 7%.

These efficiency gains are especially pronounced on long-form video (from 10-minute how-to guides to 90-minute lectures and multi-hour recordings), where static processing forces developers to choose between high token costs or techniques that drop critical details.

Activating agentic video understanding drops token consumption by up to 88% and boosts accuracy by up to 7% with Gemini 3.7 Flash.

While these gains span all three supported models, Gemini 3.7 Flash with agentic understanding offers the best possible quality overall and the best combination of quality and cost efficiency, putting it at the accuracy-to-cost pareto frontier among tested models for video understanding.

Using agentic video understanding places Gemini 3.7 Flash at the accuracy-to-cost pareto frontier for video analysis.

How it works

Instead of static processing where the model ingests media streams at a fixed frame rate, agentic video understanding enables Gemini to take an active, goal-directed role in determining *what* to watch, at *what* speed, and through *which* modality (frames, audio, or transcript), fetching only the moments and signals needed. While developers could previously do this manually, with agentic video understanding, Gemini can accomplish it through an agentic loop, invoking an internal tool to load the relevant part of the video file, significantly reducing development overheads.

Capabilities and use cases

Agentic video understanding transforms how developers can process long-form video content across a variety of demanding applications.

- Sub-second moment retrieval: Pinpoint split-second state changes and tight cut boundaries that are easily missed at 1 FPS, making precise automated video editing possible.
- Long-form needle-in-a-haystack search: Answer complex queries across multi-hour videos without consuming millions of tokens.
- Anomaly detection: Resample interesting time windows at higher FPS to inspect rapid motion and subtle visual artifacts.
- Counting action & object: Accurately track repeated physical movements and distinct objects over time.

Token-efficient long-form video analysis

See how Gemini 3.7 Flash performs with and without agentic video understanding on LongVideoBench, a long-form video understanding benchmark. Notice the large token reductions and accuracy improvements.

Accurate fast action analysis with dynamic FPS

With agentic video understanding, 3.7 Flash is able to accurately count a fast-paced movement by scanning and rewatching the video at different frames per second, as needed.

Token-efficient needle-in-a-haystack search

Using agentic video understanding, Gemini 3.7 is able to accurately answer complex questions based on the content of the video while consuming a significantly lower number of tokens compared to static analysis.

Real-world results

Many of our early access partners saw strong performance while testing with agentic video understanding. Here's what they have to say:

Getting started

Agentic video understanding is available via the Gemini API in Google AI Studio and Gemini Enterprise Agent Platform, launching across Gemini 3.7 Flash, 3.6 Flash, and 3.5 Flash-Lite. It uses standard Gemini API token pricing with no additional feature fee.

To enable it, simply set processing to "agentic" in the API configuration. Read our developer guide to get more insights into the feature and how to get started.

We are also bringing the efficiency and quality improvements of agentic video understanding to billions of users across Google products. The feature will roll out to all users in the Gemini app across Flash and Flash-Lite models soon. And in the coming months, agentic video understanding will also power YouTube's 'Ask YouTube' feature on the video watch page, leveraging Gemini to deliver higher-quality answers grounded in the visuals.

Acknowledgement for their contribution to this work: Sergi Caelles, Filip Pavetić, Ahmet Iscen,

Suhas Yogin, and the Agentic Vision team.

<https://blog.google/innovation-and-ai/models-and-research/gemini-models/introducing-agentic-video-in-gemini>