

Getting Veo Running Through Google Cloud

A setup guide from ATP AI, Sales and Marketing. Written for business owners, not engineers. You bring Claude. Claude does the technical part. Checked against Google's own docs.

Veo is Google's video generation model. There are two ways in. One is Flow, a website you log into like any other. The other is the API, which needs a Google Cloud project, a billing account and code. Here is the honest part up front: most small businesses should use Flow and a subscription. The API buys you one thing, and that is video made while nobody is watching. This guide covers both doors and shows you exactly where the line is.

YOU WILL NOT NEED A TERMINAL

You already have Claude. That changes this whole guide. You will never type a command yourself. You do the browser parts, which are clicking, signing in and copying. Then you tell Claude what you want and Claude does the technical part. If you are in Claude Code, it runs the commands on your machine directly. If you are in Claude Desktop, it may hand you one line to paste and tell you exactly where to put it. If you are in Cowork, Claude works in its own sandbox, so ask it to give you the file or the line to run at home. That is the only difference, and I will not keep repeating it.

Before you start

- A Google account. A regular Gmail works. A Google Workspace account works too.
- Claude running. Claude Code, Claude Desktop or Cowork. Any of the three is fine.
- A credit card, if you go past the free Flow credits. Both routes bill to a card.
- A web browser and about 20 minutes for the Flow route.
- About 90 minutes for the API route if nothing goes sideways. Most of that is waiting, not working.

The steps

1 Pick your door before you spend a dollar

This step decides everything after it. There are two doors. Door one is Flow. It is Google's video studio at labs.google/fx/tools/flow. You sign in, type a prompt, get a video. It has a timeline editor, camera controls and a scene builder. You do not create a project or set up an API. With no subscription at all you get 50 Flow credits a day, free. Door two is the API. That is either the Gemini API or the Gemini Enterprise Agent Platform. Agent Platform is the new name for Vertex AI. Google renamed it in May 2026. The API is for making video from code. Think 200 product clips overnight, or a pipeline wired into your CRM. Here is the plain version. If a person looks at every clip before it ships, you want Flow. The API gives you one thing Flow does not have. That thing is automation. Ten videos a month by hand does not need it. What you do: open labs.google/fx/tools/flow in your browser and look at it.

ASK CLAUDE

I run [describe your business in one line]. I need about [number] videos a month, and [a person reviews each one / nobody reviews them, they post automatically]. Tell me straight which one I need: Google Flow with a subscription, or the Veo API. Do not be polite about it. Give me the reason in two sentences.

You know it worked when: You can say out loud which door you are taking, and why, without hedging.

WATCH OUT

People pick the API because it sounds more serious. It is not more capable. It is the same model underneath. You are trading a working editor for code.

2 Run the free Flow test before you spend anything

This step tells you whether Veo can even do your job. It costs nothing. What you do: go to labs.google/fx/tools/flow and sign in with your Google account. Type a prompt. Watch what comes back. You get 50 free credits a day with no subscription. The free credits work on Veo 3.1 Lite, Fast and Quality. Here is what each generation costs in credits, from Google's Flow help page: Veo 3.1 Lite: 10 credits, or 5 for Ultra subscribers Veo 3.1 Fast: 20 credits, or 10 for Ultra subscribers Veo 3.1 Quality: 100 credits for everyone Gemini Omni Flash: 15 to 30 credits, depending on clip length So your free 50 a day buys five Lite clips, or two Fast clips, and not one Quality clip. That is still enough to learn what the model can do. Do it first. One quirk. Your free daily refresh starts when you make your first generation. It is not tied to midnight. Unused credits do not carry over to tomorrow.

ASK CLAUDE

Write me five Veo prompts for my business. I sell [what you sell] to [who buys it]. Each prompt is one single shot, 8 seconds, and it must stand on its own. Do not put any on-screen text in the prompt. If anyone speaks, use a colon before the line instead of quote marks. Give me the five prompts as plain text I can copy one at a time.

You know it worked when: You have at least two finished clips you can play, and your credit counter in Flow has gone down.

WATCH OUT

The credit cost is per generation, not per request. One request in Flow can make two videos and charge you twice. Also, the moment you upgrade to a paid plan, any leftover free credits are gone.

3 Price the subscription against the API

This step stops you building something you did not need. Do the math before the setup, not after. What you do: open gemini.google/subscriptions and look at the plans. In Canadian dollars they read like this: Google AI Plus: CA\$6.99 a month, 200 Flow credits Google AI Pro: CA\$26.99 a month, 1,000 Flow credits Google AI Ultra: from CA\$139.99 a month for 10,000 credits, or CA\$279.99 a month for 25,000 Now compare. AI Pro at CA\$26.99 gives you 1,000 credits. That is 10 Veo 3.1 Quality clips, or 50 Fast clips, or 100 Lite clips. You also get the whole editing setup. The same 10 Quality clips through the API, at 8 seconds each, cost $8 \times \$0.40 \times 10$. That is \$32 USD. And you get no editor, no timeline and no scene builder. You get a block of data back. The API wins on one thing only, and that is volume nobody is watching. If that is not you, buy the subscription and go make videos.

ASK CLAUDE

Here is my real usage: about [number] clips a month, mostly [Lite / Fast / Quality]. Work out two numbers for me. One, the cheapest Google AI subscription that covers it, using Flow credit costs of 10 for Lite, 20 for Fast and 100 for Quality. Two, the same clips through the Veo API at \$0.10 per second for Fast, \$0.05 for Lite and \$0.40 for Standard, at 8 seconds each. Put both numbers side by side and tell me which is cheaper and by how much.

You know it worked when: You have two dollar figures in front of you and one is clearly smaller.

WATCH OUT

Prices show in your local currency, based on where Google thinks you are. The numbers above are what the page serves in Canada. Check your own before you put a number in a budget.

4 Create the Google Cloud project

A project is just a folder for your billing and your settings. Everything Veo does on the API route lives inside it. What you do: go to console.cloud.google.com and sign in. At the top of the screen there is a project selector dropdown. Click it. Click New Project. Give it a name you will still recognise in six months, like `veo-video-prod`. Click Create. Google gives one piece of advice here and it is good. If you are not sure you are keeping this, make a fresh project rather than using one you already have. Deleting the project later wipes everything inside it in one move. That is much cleaner than hunting down leftovers. If you are new to Google Cloud, new customers get \$300 in free credits.

ASK CLAUDE

```
I just created a Google Cloud project called veo-video-prod in the console. Check whether the Google Cloud command line tool is installed on this machine. If it is not, install it. If you cannot install it yourself, give me the single download link and tell me exactly what to click. Once it is working, list my projects, find that one, and tell me its project ID. Save the ID somewhere we can both use for the rest of this setup.
```

You know it worked when: The project name shows in the bar at the top of the Cloud console, and Claude has told you the project ID.

WATCH OUT

If you already run Google Cloud for something else, do not bolt Veo onto that project. Video generation bills fast. You want it on its own line in the billing report.

5 Choose the Gemini API or the Agent Platform

These are two different products and both run Veo. Their docs live in different places. Choose once and stick with it. The Gemini API, at ai.google.dev, signs in with a single API key. Setup takes minutes. Your finished videos sit on Google's server for two days, then go. This is the one a small business should use. The Gemini Enterprise Agent Platform is the old Vertex AI, renamed in May 2026. It signs in with Google Cloud credentials, not a key. It writes your video straight into a storage bucket you own, so nothing expires. It also uses the finished model IDs rather than preview ones. More setup, more control. The rename matters when you go searching. Half the guides online still say Vertex AI. Old cloud.google.com/vertex-ai links now redirect to docs.cloud.google.com/gemini-enterprise-agent-platform. The API address itself did not change, so old code still runs. Only the names and the docs moved. Steps 6 through 10 are the Gemini API route. Step 11 is the Agent Platform route. What you do: nothing in the browser yet. Just make the call.

ASK CLAUDE

Explain in plain words which one I should use: the Gemini API, or the Gemini Enterprise Agent Platform. What matters to me is [I want my files to stay put / I just want the fastest setup / I need the finished models, not preview]. Then remember my choice and use it for the rest of this setup. Tell me which model ID goes with my choice.

You know it worked when: You know which model ID you are using. The Gemini API uses `veo-3.1-generate-preview`. The Agent Platform uses `veo-3.1-generate-001`.

WATCH OUT

Do not mix them. The model IDs are different. Paste the wrong one and you get a model-not-found error that tells you nothing useful.

6 Turn on billing and create your API key

Veo has no free tier on the API. Google's pricing page lists Free Tier as Not available for every single Veo model. Billing has to be on before your first generation will run. What you do: go to aistudio.google.com/apikey. Sign in with the same Google account. Pick the project you made in step 4. Create an API key. Google AI Studio also has a Set up Billing option. The rate limits doc is clear: to move off the Free tier you have to set up billing in AI Studio first. One thing that catches people using old guides. Google is retiring the old key type. Every new key made in AI Studio is now an auth key, tied to a service account. Unrestricted standard keys are already being rejected. In September 2026 the Gemini API will reject all standard keys. If you have an old key sitting in a script somewhere, replace it now. A key is like a password. Never paste it into a chat, an email or a shared doc. Not even to Claude. You put it in the file yourself.

ASK CLAUDE

```
Set up a safe home on this machine for my Gemini API key. Create a .env
file in my working folder with one line, GEMINI_API_KEY, left empty. Make
sure that file can never be committed to git. Then tell me the exact file
path so I can paste the key in myself. Do not ask me to send you the key,
and do not write the key into any script.
```

You know it worked when: The .env file exists, you pasted your key into it yourself, and no key text appears anywhere in your chat.

WATCH OUT

Turning on billing does not give you big limits right away. A new paid project starts on Tier 1. That is capped at \$250 and has a spend limit of \$10 per rolling 10 minutes. At Veo 3.1 Standard prices that is about three 8-second clips per 10 minutes before you start getting errors.

7 Set a budget alert, and learn what it does not do

This is a smoke alarm, not a sprinkler. Set it anyway. What you do: in the Cloud console go to Billing, then Budgets and alerts, then Create budget. Choose Alerts only. Name it. Set the scope to your Veo project. Set an amount you would be genuinely annoyed to go over. Fifty dollars is a sensible start. Now read this twice, because it is the most expensive misunderstanding in Google Cloud. Google's own documentation says an alerts-only budget does not automatically cap usage or spending. A budget alert sends you an email. It stops nothing. If a script has a bug and loops, the alert tells you about it after the money is gone. Your real protections are the tier spend limit from step 6, and a cap written into your own script. That second one is Claude's job.

ASK CLAUDE

```
From now on, every Veo script you write for me needs a hard spend guard built in. Cap the number of clips per run at a number I approve first. Print the running dollar total after every single clip. Stop the script the moment the total passes $50. Show me the cap and the price per clip before you run anything, and wait for me to say go.
```

You know it worked when: Your budget shows in the Budgets and alerts list, and Claude has shown you a clip cap and a dollar cap before running anything.

WATCH OUT

An alert is not a cap. A real hard stop means wiring budget notices through Pub/Sub to a function that switches billing off. That is a real build, not a checkbox.

8 Test in Veo Studio before any code exists

This step separates two very different problems, and it saves hours. What you do: Google AI Studio has a demo app called Veo Studio. It runs against your paid API key with no code at all. Open it and generate five or ten clips. Get a feel for how the prompts behave. Confirm your key and your billing work from end to end. Here is why it matters. If your video comes back looking wrong, that is a prompt problem. If nothing comes back at all, that is a billing or sign-in problem. Veo Studio tells you which one you have before there is any code to blame. Everything you make here bills at the normal per-second rate. It is not free just because you did not write the code.

ASK CLAUDE

I am testing in Veo Studio. Here is what happened: [paste the error message, or describe what came back]. Tell me whether this is a prompting problem or a billing and sign-in problem. Then give me the next thing to check, one step at a time.

You know it worked when: At least one clip comes back and plays. Your billing page shows a charge for it.

WATCH OUT

Veo Studio needs a paid Gemini API key. If step 6 is not finished it will not run, and the error will not necessarily tell you that is why.

9 Have Claude run your first real generation

This is the step people quit on. You are not going to quit on it, because you are not doing it. Claude is. What you do: nothing but read and approve. No typing commands. So you know what is happening, here is the plain version. Claude installs a small piece of Google code called google-genai. It writes a short script. It sends your prompt to Veo. Then it waits, because video is not instant. The API takes your request and hands back a job ticket. The script keeps checking that ticket until the video is finished. Google puts the wait at 11 seconds at the fastest and up to 6 minutes at busy times. Ask for the Fast model on your first run. Your first working call costs \$0.80 instead of \$3.20. Move up to veo-3.1-generate-preview once the plumbing works. If you ever need the raw address, it is POST <https://generativelanguage.googleapis.com/v1beta/models/veo-3.1-generate-preview:predictLongRunning>. You will not type it. Claude will.

ASK CLAUDE

```
Install the google-genai Python client on this machine. Then write a script called veo_test.py that makes one video with the Veo Fast model, model ID veo-3.1-fast-generate-preview, using this prompt: A cinematic shot of a majestic lion in the savannah. Read my key from the GEMINI_API_KEY environment variable and never write the key into the script. This is a long-running job, so check the operation every 10 seconds until it reports done, and do not time out before 6 minutes. Download the finished video into this folder the moment it is ready. Run it, then show me the file. If you cannot run commands yourself, give me the one line to paste and tell me exactly where to paste it.
```

You know it worked when: There is an .mp4 file on your computer, about 8 seconds long, with sound. Your billing shows roughly \$0.80.

WATCH OUT

You are charged for the generation, not for the clip you keep. Every clip you throw away was still billed. Google documents one exception: if an audio processing issue stops the video generating, you are not charged. You only pay if the video is successfully generated.

10 Download the video within two days or lose it

On the Gemini API route, your finished videos sit on Google's server for 2 days and are then removed. There is no grace period and no recovery. What you do: check your own folder after every run and confirm the file is actually there. Do not trust that it is waiting for you online. The fix is simple. The download belongs in the same script that does the generating. Never plan to come back for it later. Later is Wednesday and the file is gone. One wrinkle if you use video extension. Referencing a video to extend it resets its 2-day timer. Extended videos count as newly generated, so their clock starts fresh. If a 48-hour window does not fit how you work, that is a strong reason to use the Agent Platform route in step 11. There, output lands in your own storage bucket and stays until you delete it.

ASK CLAUDE

New standing rule for all my Veo work. Every script you write must download the finished video to my computer in the same run. Never leave a file sitting only on Google's server. Name each file with the date and a short version of the prompt so I can find it later. Confirm the file exists on disk before the script finishes, and tell me if it does not.

You know it worked when: You can open the video file from your own drive, with the internet off.

WATCH OUT

This catches people who generate a batch on Friday afternoon and open their laptop on Monday morning.

11 The Agent Platform route, if you want to own the output

Use this instead of steps 6 through 10 if you want the finished models rather than preview ones, or you want your video written straight into storage you own. What you do: in the Cloud console, find Agent Platform and enable it on your project. There is an Enable the API button. You will also get a browser window asking you to sign in and allow access. Click Allow yourself. That part is yours, not Claude's. Everything else is Claude's. It turns on the API, signs the machine in, makes the storage bucket, sets the settings and runs the first generation. One trap worth knowing. The setting is now `GOOGLE_GENAI_USE_ENTERPRISE` set to `True`. Older code and older tutorials use `GOOGLE_GENAI_USE_VERTEXAI`. If you copied a script from before the rename, that is why it does nothing. This route is fiddlier. More moving parts, more ways to hit a permission error. It is worth it only if the two-day expiry or the preview model status is a real problem for you.

ASK CLAUDE

```
Set my project up on the Gemini Enterprise Agent Platform instead of the
Gemini API. Enable the API with gcloud services enable
aiplatform.googleapis.com. Sign this machine in with gcloud auth
application-default login, and tell me when a browser window opens so I
can click Allow myself. Create a Cloud Storage bucket for my video output
and tell me its address. Set GOOGLE_CLOUD_PROJECT to my project ID,
GOOGLE_CLOUD_LOCATION to global, and GOOGLE_GENAI_USE_ENTERPRISE to True.
Do not use GOOGLE_GENAI_USE_VERTEXAI, that name is dead. Then write and
run a script that makes one video with model veo-3.1-generate-001, aspect
ratio 16:9, writing the output to my bucket, checking every 15 seconds
until it is done. Print the finished video address when it lands.
```

You know it worked when: A video file appears in your Cloud Storage bucket, and it is still there tomorrow, and the day after.

WATCH OUT

The Agent Platform pricing page writes prices as dollars per 1 count, which reads as though a whole clip costs 40 cents. It does not. The Gemini API pricing page states the same figures as per second in USD. Budget per second.

12 Know the hard limits before you promise a client anything

These are the walls. No prompt gets around them. All of this comes from Google's current model specification table. Clip length. 4, 6 or 8 seconds. That is the whole range. It must be 8 seconds if you want 1080p, 4K, or reference images. Veo 3.1 and 3.1 Fast can extend a clip you already made by 7 seconds at a time, up to 20 times. That tops out near 148 seconds of finished video. Extension is 720p only, so you cannot extend a 1080p clip. If you need a two-minute continuous 1080p shot, Veo cannot do it. Aspect ratios. Two. 16:9 landscape, which is the default, and 9:16 portrait. There is no 1:1 and no 4:5. Square posts get cropped in the edit. Frame rate. 24fps. Audio. Built in and always on for every Veo 3.1 version. You cannot switch it off on the Gemini API. On Agent Platform there is a cheaper silent price, currently \$0.20 per second at 720p and 1080p, against \$0.40 with audio. Text in frame. This is the one that ruins ads. Veo renders on-screen text badly. It will also try to draw your quotation marks as text in the video. Google's best-practice doc says to use a colon after the speaker's action to show speech, and to avoid quotation marks. So write: A woman says: My name is Clara. Put your logo, your price and your call to action on in the edit instead. Watermarking. Every Veo output carries a SynthID watermark marking it as AI-generated. One video per request. English is the only prompt language Google has fully tested. What you do: read this before you quote a job, not after.

ASK CLAUDE

Check this job against Veo's real limits before I quote it. Here is what the client wants: [describe the video, the length, the shape and where it is going]. The limits are: clips of 4, 6 or 8 seconds only, 16:9 or 9:16 only, 24fps, audio always on, 1080p and 4K need 8 seconds, extension adds 7 seconds at a time up to 20 times at 720p only, and no on-screen text. Tell me whether this job is possible. If it is, tell me how many separate clips it needs and roughly what it costs at 4 generations per kept shot.

You know it worked when: You have a yes or no, a clip count, and a dollar figure before you send the quote.

WATCH OUT

The 8-second ceiling is the real limit on commercial work. A 30-second ad is four clips minimum, and you will not get four usable clips in four tries. Budget three to five generations for every shot you keep.

What it costs

Veo bills per second of video made. You pay whether you keep the clip or bin it. These are the published rates in USD, paid tier only. There is no free tier on any Veo model. Veo 3.1 Standard with audio: \$0.40 per second at 720p and 1080p, \$0.60 at 4K. An 8-second clip is \$3.20, or \$4.80 at 4K. Veo 3.1 Fast with audio: \$0.10 per second at 720p, \$0.12 at 1080p, \$0.30 at 4K. An 8-second 720p clip is \$0.80. Veo 3.1 Lite with audio: \$0.05 per second at 720p, \$0.08 at 1080p. No 4K. An 8-second 720p clip is \$0.40. On the Agent Platform there is a cheaper silent tier. That is \$0.20 per second for Veo 3.1 without audio at 720p and 1080p, and \$0.08 per second for Fast at 720p. The number that actually decides your bill is not the price

per clip. It is how many clips you throw away. Plan on three to five generations for every shot you use. A 30-second ad built from four 8-second shots really costs between \$13 and \$16 on Fast, or \$50 to \$65 on Standard 1080p. That is before anyone edits anything. There is one documented case where you are not charged. If an audio processing issue stops the video generating, it is free. Google's wording is that you will only be charged if your video is successfully generated. A clip that generates fine and looks terrible is a clip you paid for. For comparison, Google AI Pro at CA\$26.99 a month includes 1,000 Flow credits. That is 10 Veo 3.1 Quality generations, or 50 Fast, or 100 Lite, plus the editor. All of this changes without much notice. Check the pricing pages before you put a number in a proposal.

When it goes wrong

The tutorial I found says Vertex AI and none of the links work	Vertex AI was renamed the Gemini Enterprise Agent Platform. It was announced at Google Cloud Next in April 2026 and finished in May 2026. Old cloud.google.com/vertex-ai links now redirect to docs.cloud.google.com/gemini-enterprise-agent-platform. The API address aiplatform.googleapis.com did not change, so existing code still runs. In the console, look for Agent Platform wherever the guide says Vertex AI.
My Python script worked last year and now returns nothing	Two likely causes. First, the setting name changed. Agent Platform now uses <code>GOOGLE_GENAI_USE_ENTERPRISE</code> set to True, where older code used <code>GOOGLE_GENAI_USE_VERTEXAI</code>. Second, your model ID may be dead. Veo 2 and Veo 3 were flagged for shutdown on June 30, 2026. Move to Veo 3.1. Paste the script to Claude and ask it to fix both.
429 RESOURCE_EXHAUSTED, even though billing is on	You have hit a spend-based rate limit, not a quota. Tier 1 allows \$10 of spend per rolling 10-minute window. At Veo 3.1 Standard that is about three 8-second clips. Wait it out, switch to Fast or Lite, or spread the batch over time. Tier 2 lifts it to \$50 per 10 minutes. It needs \$100 paid plus 3 days from your first successful payment. Ask Claude to add a pause between clips so the script rides it out instead of failing.
My API key is rejected and it used to work	Google now rejects unrestricted standard API keys, and will reject all standard keys in September 2026. Make a fresh key at aistudio.google.com/apikey. New keys are created as auth keys, tied to a service account. Paste the new key into your .env file yourself. Never into a chat.
My video disappeared	On the Gemini API route, videos are stored for 2 days and then removed. There is no recovery. Download inside the window. Or move to the Agent Platform route and write output to your own Cloud Storage bucket. Ask Claude to add the download step to every script you own.

There is garbled text floating in my video	Veo is trying to draw your quotation marks. Drop them. Write: A woman says: My name is Clara. Do not put the line in quote marks. Add any real on-screen text, logos and prices in your editor afterwards.
I asked for a 15-second clip and got 8 seconds	Veo only makes 4, 6 or 8 seconds. Anything longer is built by making an 8-second clip and extending it 7 seconds at a time, up to 20 times, at 720p only. Extension works on Veo 3.1 and 3.1 Fast, not on Lite.
1080p or 4K is rejected	Those resolutions need durationSeconds set to 8. Reference images also force 8 seconds. And 4K is not available on Veo 3.1 Lite at all.
My bill is higher than my budget alert said it would be	Budget alerts do not cap anything. Google's docs say an alerts-only budget does not automatically cap usage or spending. The alert is an email after the fact. Your real ceiling is the tier spend limit and the cap written into your own script. Ask Claude to add one if it is missing.
Generation is slow and I think it has hung	It probably has not. Google documents a wait of 11 seconds at the fastest and up to 6 minutes at busy times. The script needs to survive six minutes of waiting without giving up. Tell Claude to set the timeout past six minutes.

Where this came from

Last verified: 22 August 2026. Vendor screens change constantly. If a step does not match what you see, the tool moved.

<https://ai.google.dev/gemini-api/docs/veo>

<https://ai.google.dev/gemini-api/docs/pricing>

<https://ai.google.dev/gemini-api/docs/rate-limits>

<https://ai.google.dev/gemini-api/docs/api-key>

<https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/video/generate-videos-from-text>

<https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/video/best-practice>

<https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/veo/3-1-generate>

<https://cloud.google.com/gemini-enterprise-agent-platform/generative-ai/pricing>

<https://docs.cloud.google.com/billing/docs/how-to/budgets>

<https://support.google.com/flow/answer/16526234>

<https://gemini.google/subscriptions/>

<https://cloud.google.com/blog/products/ai-machine-learning/veo-3-1-lite-and-a-new-veo-upscaling-capability-on-vertex-ai>

<https://labs.google/fx/tools/flow>

<https://aistudio.google.com/apikey>

WHO MADE THIS

Built by ATP AI, Sales and Marketing. We set this up inside businesses for a living, so if you would rather not do it yourself, that is what we are for. atpsalesandmarketing.com/contact-us/