Gemini API Referentie — juni 2026

> Samengesteld door Sjakie op 2026-06-12 uit 18 officiële docs-pagina's op ai.google.dev.
> Focus: (a) virtual-makeover web-app (Nano Banana edits + Veo-video via Netlify Functions), (b) AI Studio Build-modus workflow.

---

1. Quickstart

- **API key:** gratis aanmaken via https://aistudio.google.com/apikey
- **SDK's:** Python `pip install -U google-genai` (Python 3.9+), JS `npm install @google/genai` (Node 18+)
- **Auth header (REST):** `x-goog-api-key: $GEMINI_API_KEY`
- **Endpoint:** `https://generativelanguage.googleapis.com/v1beta/models/<model>:generateContent` (streaming: `:streamGenerateContent`)
- **Kernmethodes:** Python `client.models.generate_content()` / `generate_content_stream()` / `client.chats.create()` + `chat.send_message()`. JS: `ai.models.generateContent()` / `generateContentStream()` / `ai.chats.create()` + `chat.sendMessage()`.
- **Valkuilen:**
- REST is stateless — bij multi-turn altijd de volledige history meesturen.
- Google adviseert nieuwe projecten de **Interactions API (Beta)** i.p.v. `generateContent`.
- Function calling vereist expliciete `functionDeclarations`.

---

2. Modellen (overzicht)

Tekst/algemeen:

Model-ID	Status	Positionering
`gemini-3.5-flash`	Stabiel (GA)	Slimste model voor agentic/coding
`gemini-3.1-flash-lite`	Stabiel	Goedkoop frontier-class

`gemini-3.1-pro-preview`	Preview	Complexe problemen
`gemini-3-flash-preview`	Preview	(o.a. nog Computer Use)
`gemini-2.5-pro` / `gemini-2.5-flash` / `gemini-2.5-flash-lite`	Stabiel	Oudere generatie, goedkoper

Beeld (Nano Banana familie):
- `gemini-3.1-flash-image` — **Nano Banana 2** (snel, high-volume) — stabiel
- `gemini-3-pro-image` — **Nano Banana Pro** (professioneel, 4K) — stabiel
- `gemini-2.5-flash-image` — **Nano Banana** (origineel, snel/goedkoop) — stabiel
- Imagen 4 (`imagen-4.0-*`) — pure text-to-image tot 2K

Video: `veo-3.1-generate-preview`, `veo-3.1-fast-generate-preview`, `veo-3.1-lite-generate-preview`, `veo-3.0-generate-001`, `veo-3.0-fast-generate-001`, `veo-2.0-generate-001` (stabiel).

Overig: TTS (`gemini-3.1-flash-tts-preview`, `gemini-2.5-*-tts`), Live audio (`gemini-3.1-flash-live-preview`), live translate (`gemini-3.5-live-translate-preview`, 70+ talen), muziek (`lyria-3-*`), embeddings (`gemini-embedding-2` multimodaal, `gemini-embedding-001`), computer use (`gemini-2.5-computer-use-preview-10-2025`), deep research (`deep-research-preview-04-2026`), agent (`antigravity-preview-05-2026`).

- Alias `gemini-flash-latest` bestaat; preview-modellen hebben datum-suffixen.

---

3. Wat is nieuw in Gemini 3.5 (Flash)

- **`gemini-3.5-flash` is GA/stabiel.** 1M input-context, 65k max output, kennis-cutoff jan 2025.
- **`thinking_level`** vervangt `thinking_budget`: `minimal` | `low` | `medium` (nieuwe default) | `high`. Default ging van high → medium.
- **`temperature`, `top_p`, `top_k` zijn DEPRECATED** — verwijderen uit requests.
- Thought signatures: redenering wordt automatisch meegenomen over beurten (SDK regelt dit).
- Function calling: `FunctionResponse` moet nu matchende `id` én `name` bevatten.
- `media_resolution` per content-item: `low` / `medium` / `high` / `ultra_high`.
- Tools: Google Search, Maps grounding, code execution, URL context, File Search, function calling (combineerbaar). **NIET:** Computer Use (blijf daarvoor op `gemini-3-flash-preview`). Image segmentation niet in 3.x.
- Migratie van `gemini-3-flash-preview`: modelnaam wijzigen, sampling-params weg, thinking_budget → thinking_level, prompts hertesten (minder verbose output).

4. Prijzen (per 1M tokens, standaard tier, tenzij anders)

Tekstmodellen:

Model	Input	Output	Batch (in/out)
gemini-3.5-flash	$1.50	$9.00	$0.75 / $4.50
gemini-3.1-flash-lite	$0.25 (audio $0.50)	$1.50	$0.125 / $0.75
gemini-3.1-pro-preview	$2.00 (≤200k) / $4.00	$12.00 / $18.00	50%
gemini-3-flash-preview	$0.50 (audio $1.00)	$3.00	$0.25 / $1.50
gemini-2.5-pro	$1.25 (≤200k) / $2.50	$10.00 / $15.00	50%
gemini-2.5-flash	$0.30 (audio $1.00)	$2.50	$0.15 / $1.25
gemini-2.5-flash-lite	$0.10 (audio $0.30)	$0.40	$0.05 / $0.20

Beeldmodellen (belangrijkst voor makeover-app):

Model	Input	Output per afbeelding
`gemini-2.5-flash-image` (Nano Banana)	$0.30	**$0.039/img** (batch $0.0195, priority $0.0702)
`gemini-3.1-flash-image` (NB 2)	$0.50	$0.022 (0.5K) / $0.034 (1K) / $0.050 (2K) / $0.076 (4K)
`gemini-3-pro-image` (NB Pro)	$2.00	$0.134 (1K/2K) / $0.24 (4K)
Imagen 4	—	$0.02 (Fast) / $0.04 (Standard) / $0.06 (Ultra)

Video (per seconde):

Model	Prijs/sec
Veo 3.1 standaard	$0.40 (720p/1080p), $0.60 (4K)
Veo 3.1 Fast	$0.10 (720p) / $0.12 (1080p) / $0.30 (4K)
Veo 3.1 Lite	**$0.05 (720p) / $0.08 (1080p)**
Veo 3	$0.40; Fast $0.10–0.30
Veo 2	$0.35

→ Een 8-sec draaivideo: Veo 3.1 Lite 720p = **$0.40**; Veo 3.1 standaard = $3.20.

Overig: TTS 3.1 $1.00 in / $20.00 audio-out; embeddings `gemini-embedding-2` $0.20 tekst; Lyria $0.04 per 30s-clip. Context caching 3.5-flash: $0.15 read + $1.00 storage per 1M/uur. Google Search/Maps grounding: 5.000 prompts/maand gratis (Gemini 3), daarna $14 per 1.000. File Search $0.15 per 1M embedding-tokens.

Gemini 2.0-serie: shutdown 1 juni 2026 — niet meer gebruiken.

Free tier: beperkte modellen, gratis tokens, maar **data wordt gebruikt om Google-producten te verbeteren**. Paid tier: niet.

5. Image generation / editing (Nano Banana)

- **Editing = gewoon `generate_content` met `[prompt, image]` in contents.** Beeld als base64 inline of via Files API.
- `config.response_modalities=['TEXT','IMAGE']` ; output in `part.inline_data` (base64).
- **Aspect ratios:** 1:1, 2:3, 3:2, 3:4, 4:3, 4:5, 5:4, 9:16, 16:9, 21:9, 1:4, 4:1, 1:8, 8:1. **Sizes:** 512, 1K, 2K, 4K (hoofdletter K): `response_format={"image": {"aspect_ratio":"3:4","image_size":"2K"}}` .
- **Referentiebeelden:** 3.1 Flash max 10 object + 4 character (14 totaal); 3 Pro max 6 object + 5 character (11 totaal). Character-referenties = gezichtsconsistentie.
- Multi-turn editing via chat (iteratief bijsturen) — ideaal voor makeover-iteraties.
- **Thinking staat default AAN bij Gemini 3-beeldmodellen en kan niet uit** — wel sturen met `thinking_config.thinking_level` ("High"/"Minimal"). Thinking-tokens worden altijd gefactureerd.
- Grounding met Google Search mogelijk; image search alleen 3.1 Flash.
- Video-naar-image (frames analyseren) alleen 3.1 Flash; max 131.072 input-tokens.
- **Alle output krijgt SynthID-watermerk** (onzichtbaar).
- SDK handelt `thought_signature` automatisch af — bij raw REST zelf terugsturen in history.

---

6. Veo video

- **Image-to-video:** `image` -parameter = startframe (perfect voor makeover-resultaatfoto → draaivideo).
- **First+last frame interpolatie:** `image + lastFrame` (alleen Veo 3.1) — before/after-overgang mogelijk.
- **Referentiebeelden:** max 3 (`reference_images` , `reference_type:"asset"` , alleen 3.1) — subject-behoud.
- **Video-extensie:** +7 sec per stap, max 20 extensies, tot 148 sec totaal (3.1 & 3.1 Fast, niet Lite; alleen 720p).
- **Config:** `aspectRatio` "16:9"/"9:16"; `durationSeconds` "4"/"6"/"8" (1080p/4K en extensie/referenties vereisen "8"); `resolution` 720p/1080p/4K; `personGeneration` — bij image-to-video alleen `"allow_adult"` ; **EU/UK/CH/MENA: uitsluitend `allow_adult` .**

- **Async pattern:** `generate_videos()` geeft operation terug → elke 10 sec pollen via `client.operations.get(operation)` tot `operation.done` → `client.files.download(file=video.video)`. REST: POST `:predictLongRunning`, dan GET `/v1beta/{operation_name}`.
- **Latency: 11 sec tot 6 minuten** — past NIET in een Netlify Function van 10 sec. Architectuur: function start de job en geeft operation-name terug; aparte poll-endpoint of background function checkt status.
- **Video's worden na 2 dagen van de server verwijderd** — direct downloaden en zelf opslaan.
- Prompt max 1.024 tokens; audio native (altijd aan bij 3/3.1); 24fps; SynthID-watermerk; safety-block = geen kosten.
- Geen webhooks — alleen pollen.

---

7. Imagen 4

- Model-ID's: `imagen-4.0-generate-001`, `imagen-4.0-ultra-generate-001`, `imagen-4.0-fast-generate-001`. Imagen 3 is uitgezet.
- Eigen methode: `client.models.generate_images(...)` met `GenerateImagesConfig`.
- Params: `numberOfImages` 1–4 (default 4!), `aspectRatio` 1:1/3:4/4:3/9:16/16:9, `imageSize` 1K/2K (niet bij Fast), `personGeneration` (default `allow_adult`; `allow_all` verboden in EU/UK/CH/MENA).
- **Alleen tekst-input, alleen Engels, max 480 tokens** — géén editing. Voor de makeover-app dus irrelevant; Nano Banana is het juiste gereedschap (editing + referenties).
- Tekst-in-beeld: max ~25 tekens betrouwbaar. Prijs: $0.02–0.06 per beeld. SynthID.

---

8. Structured output

- Config: `response_mime_type: "application/json"` + schema (nieuwere docs: `response_format` / `response_json_schema`).
- Python: Pydantic `BaseModel` (via `model_json_schema()`); JS: Zod + `zodToJsonSchema()`.
- Ondersteund: types, `enum`, `format` (date-time), min/max, minItems/maxItems, required, anyOf (unions), recursie via `$ref:"#"`, `prefixItems`.
- Ondersteunende modellen: 3.5 Flash, 3.1 Flash-Lite/Pro, 2.5-serie; 2.0 vereist `propertyOrdering`.

- Combineerbaar met tools (Search, function calling) — alleen Gemini 3-serie.
- **Valkuilen:** garandeert syntactisch geldige JSON, niet semantisch juiste inhoud — altijd valideren. Te grote/diepe schema's worden geweigerd; onbekende properties stilzwijgend genegeerd. Streaming levert geldige partiële JSON-chunks.

---

9. Files API

- `client.files.upload(file="...")` / JS `ai.files.upload({file, config: {mimeType}})` ; REST = resumable upload (`X-Goog-Upload-Protocol: resumable`).
- **Limieten: 2 GB per bestand, 20 GB per project, PDF max 50 MB. Bestanden worden na 48 uur automatisch verwijderd.** Gratis in alle regio's.
- Verplicht wanneer totale request > 100 MB (anders mag inline base64).
- Gebruik: file-object direct in `contents` meegeven; REST via `file_data` (mime_type + file_uri).
- `files.get/list/delete` beschikbaar; je kunt metadata lezen maar bestanden **niet terugdownloaden** (uitzondering: Veo-output via files.download).
- Voor de makeover-app: klantfoto's < ~7 MB kunnen gewoon inline base64 (let op Netlify's eigen 6 MB request-limiet!).

---

10. Rate limits

- **4 tiers:** Free → Tier 1 (billing actief, cap $250/mnd) → Tier 2 ($100+ besteed, 3+ dagen; cap $2.000) → Tier 3 ($1.000+, 30+ dagen; cap $20.000–100.000+). Upgrade Free→1 direct, daarna ~10 min.
- Dimensies: **RPM, TPM, RPD** — één overschrijding = 429. RPD reset om middernacht Pacific.
- **Limieten gelden per project, niet per API key** — meerdere keys helpen niet.
- Actuele per-model-getallen: https://aistudio.google.com/rate-limit (niet meer statisch in docs).
- Batch: 100 gelijktijdige jobs, inputfile max 2 GB, opslag 20 GB; enqueued tokens Tier 1: 3M (3.5-flash), 1M (3.1-flash-image); Tier 3: 1B resp. 750M.
- Priority-verkeer = 0.3× de standaard rate limit. Limieten "niet gegarandeerd".

---

11. Billing

- **Prepay** (default sinds 23 maart 2026): credits vooraf, min $10 / max $5.000; **verlopen na 12 maanden**; niet-restitueerbaar; dienst stopt bij $0; auto-reload mogelijk. **Postpay** alleen Tier 3+ (en onomkeerbaar).
- Paid tier = "prompts en responses worden NIET gebruikt om Google-producten te verbeteren"; free tier wel.
- **Google Cloud welkomstcredits ($300) gelden NIET voor de Gemini API.**
- Spend caps per project mogelijk (experimenteel); cost-grafieken tot 24 uur vertraagd; batch/agents kunnen voorbij saldo doorlopen (negatief saldo mogelijk).
- 400/500-fouten kosten geen tokens, maar tellen wel mee in quota.
- Gekoppelde betaalde key in AI Studio = ook AI Studio-gebruik wordt belast.

---

12. Batch API

- **50% korting** op standaardprijs; target-doorlooptijd 24 uur (meestal sneller); jobs verlopen na 48 uur.
- Inline requests (< 20 MB totaal) of JSONL-file via Files API (max 2 GB), elke regel een request met eigen key.
- States: PENDING / RUNNING / SUCCEEDED / FAILED / CANCELLED / EXPIRED. Pollen op job-naam óf **webhooks** op `batch.succeeded / batch.failed`.
- Ondersteunt alles: tools, structured output, multimodaal, embeddings, **óók Nano Banana image-generatie** (batch-beeldprijs $0.0195/img bij 2.5-flash-image) en context caching.
- Geen idempotency — dubbel insturen = dubbele job. Check `failedRequestCount` in `batchStats`; per-regel-fouten in outputfile.
- Use-case Haarvisie: nachtelijke bulk-generatie van brand-content tegen halve prijs.

---

13. Context caching

- **Impliciet:** default aan voor alle 2.5+-modellen, automatisch, geen garantie. **Expliciet:** zelf cache aanmaken = gegarandeerde korting.
- Minimum tokens: 4.096 (3.5 Flash, 3.1 Pro), 2.048 (2.5 Flash/Pro).
- Workflow: content uploaden → `client.caches.create()` met system instruction + content → cache-naam meegeven in `generateContent`. TTL default 1 uur, aanpasbaar (bv. "300s"); alleen TTL/expire_time updatebaar.

- Prijs 3.5-flash: $0.15/1M read + $1.00/1M per uur opslag.
- Cache-inhoud is niet terug te lezen (alleen metadata); cached tokens tellen mee voor rate limits.

---

14. Tokens

- ~4 tekens per token; 100 tokens ≈ 60–80 Engelse woorden.
- **Beeld: ≤384px beide kanten = 258 tokens; groter = tegels van 768×768 à 258 tokens.** Video: 263 tokens/sec. Audio: 32 tokens/sec.
- `count_tokens` vooraf; achteraf `usage_metadata`: `prompt_token_count`, `candidates_token_count`, `total_token_count`, `cached_content_token_count`, `thoughts_token_count`.
- **Thinking-tokens worden volledig gefactureerd als output** — `thoughts_token_count` monitoren voor kostenbeheersing.
- `media_resolution` (Gemini 3) bepaalt max tokens per beeld/videoframe — hoger = beter kleine tekst lezen, duurder en trager.
- Contextlimieten per model opvraagbaar via `models.get` (`input_token_limit` / `output_token_limit`).

---

15. AI Studio Build-modus

- Natuurlijke-taal-prompt → **Antigravity Agent** genereert complete werkende app met live preview. Ook "I'm Feeling Lucky" en remixen uit de App Gallery.
- Output: web-app (React frontend + Node.js server) of Android-app (Kotlin/Jetpack Compose, emulator-preview, Play Store-publicatie mogelijk).
- **Gemini API key wordt automatisch als server-side secret geconfigureerd** — nooit in client-code, niet zichtbaar voor anderen bij delen. Eigen secrets via het Secrets-panel.
- npm-packages worden automatisch geïnstalleerd; Firebase (Firestore + Auth) en 12+ Google Workspace API's (incl. OAuth) automatisch aansluitbaar.
- Itereren: chat, directe code-editing in Code-tab, of **annotation mode** (UI-element aanwijzen + wijziging vragen).
- **Code eruit halen: Download als ZIP of push naar GitHub.** Lokaal draaien = zelf `GEMINI_API_KEY` env var zetten.

- **Beperkingen:** geen import terug van lokaal bewerkte apps; geen gedeelde edit-toegang; geen `git pull` van remote changes; gedeelde apps tellen mee in jouw usage; **max prompts/dag staat niet in de docs** (limieten via je normale API-tier).
- Browser-permissies (camera/mic/locatie) via `metadata.json` → `requestFramePermissions`.

---

16. AI Studio fullstack apps

- Architectuur: React/Angular client + Node.js server. UI-logica client-side, business-logica/keys server-side.
- **Alle Gemini-calls lopen via server-side code** met de vooraf geconfigureerde `GEMINI_API_KEY` — zelfde patroon als wij met Netlify Functions doen.
- Third-party keys via Secrets panel → als env vars in server-code; nooit hardcoden.
- Database: prompt "add a database to my app" → automatische Firestore + Firebase Auth ("Sign in with Google"). Alternatieven (Supabase, MongoDB Atlas) kunnen ook maar handmatiger.
- Multiplayer/real-time state via de server; testen met meerdere tabs of de Share-URL.

---

17. AI Studio deployen

- **Starter Tier: max 2 fullstack-apps publiceren ZONDER Google Cloud project of billing.** Eén Cloud Run-regio. Flow: Publish → Get Started → Publish App → je krijgt een Cloud Run-URL.
- Meer dan 2 apps of meer controle: Cloud Billing account + upgrade naar standaard Google Cloud project (Cloud Run-prijzen gelden).
- Verwijderen via Apps-pagina → prullenbak-icoon.
- Custom domains / env vars / GitHub-export staan niet op deze pagina (export = ZIP/GitHub vanuit Build-modus zelf).

---

18. Safety settings

- 4 instelbare categorieën: harassment, hate speech, sexually explicit, dangerous. Kinderveiligheid e.d. = ingebouwd, niet uitschakelbaar.
- Thresholds: `OFF` / `BLOCK_NONE` / `BLOCK_ONLY_HIGH` / `BLOCK_MEDIUM_AND_ABOVE` / `BLOCK_LOW_AND_ABOVE`.

- **Default bij Gemini 2.5/3-modellen is `OFF`** — extra filters dus zelf aanzetten indien gewenst.
- Configuratie: `safety_settings=[SafetySetting(category=..., threshold=...)]` in `GenerateContentConfig`.
- Geblokkeerd: `promptFeedback.blockReason` (prompt) of `candidate.finishReason == "SAFETY"` (response) + `safetyRatings` (probability HIGH/MEDIUM/LOW/NEGLIGIBLE).
- Let op: blokkering op *waarschijnlijkheid*, niet ernst. Lossere instellingen kunnen ToS-review triggeren.

Top-10 inzichten voor Haarvisie v3

1. **Nano Banana 2 (`gemini-3.1-flash-image`) is nu de sweet spot voor makeover-edits:** $0.034 per 1K-beeld (goedkoper dan NB classic's $0.039), stabiel/GA, en ondersteunt tot 4 character-referentiebeelden — direct bruikbaar voor onze heilige gezichtsconsistentie-regel. Nano Banana Pro ($0.134) alleen voor hero-shots/4K.
2. **Veo 3.1 Lite maakt draaivideo's bijna gratis:** $0.05/sec 720p = **$0.40 per 8-sec makeover-video**. Image-to-video met de makeover-foto als startframe; first+last-frame (alleen 3.1, niet Lite) kan zelfs een before→after-overgang animeren.
3. **Veo past NIET synchroon in een Netlify Function:** latency 11 sec–6 min, alleen polling (geen webhooks). Architectuur: function 1 start de job (geeft operation-name terug), function 2 is een poll-endpoint die de client elke ~10 sec aanroept. Video's na 2 dagen weg — direct downloaden naar eigen storage.
4. **Image-editing is gewoon `generate_content` met `[prompt, foto]`** + `response_modalities: ['TEXT','IMAGE']` + `response_format.image` voor aspect ratio (bv. 3:4 portret) en size (1K/2K). Multi-turn chat = iteratief bijsturen zonder opnieuw uploaden.
5. **personGeneration-valkuil voor NL/EU:** bij image-to-video in de EU alleen `allow_adult` — onze brand-modellen zijn volwassen, dus OK, maar nooit kinderen in video-pipelines. Alle output krijgt onzichtbaar SynthID-watermerk.
6. **Thinking-tokens bij Gemini 3-beeldmodellen staan altijd aan en worden gefactureerd** — zet `thinking_level: "Minimal"` voor snelle bulk-edits, "High" alleen waar kwaliteit het vereist. Bij 3.5-flash tekst: `temperature/top_p/top_k` zijn deprecated, `thinking_budget` → `thinking_level`.
7. **Batch API = 50% korting en werkt óók voor Nano Banana** ($0.0195/beeld bij 2.5-flash-image): ideaal voor nachtelijke brand-content-batches; 24h-target, webhooks beschikbaar. Niet voor de

interactieve makeover (die moet realtime).

8. **Free tier traint op je data; paid tier niet.** Voor klantfoto's (privacy!) dus altijd een betaald project gebruiken. Prepay-credits: min $10, verlopen na 12 maanden; Google Cloud $300-welkomstcredits gelden NIET voor Gemini API. Rate limits per project (niet per key); Tier 2 al na $100 + 3 dagen.
9. **AI Studio Build-modus is een snelle prototyper, geen eindstation:** Antigravity Agent bouwt React+Node fullstack app, API key automatisch als server-side secret, gratis publiceren tot 2 apps op Cloud Run (Starter Tier, geen billing nodig). Code als ZIP of naar GitHub exporteren → daarna in onze eigen Netlify-stack gieten. Let op: geen weg terug (import) na lokaal doorontwikkelen.
0. **Kosten-/limiet-randvoorwaarden voor de web-app:** inline foto-upload base64 kan tot ~100 MB request bij Gemini, maar **Netlify kapt al af op 6 MB** — client-side verkleinen blijft verplicht (kost ook minder tokens: ≤384px = 258 tokens, anders 258 per 768×768-tegel). Gemini 2.0-modellen gaan 1 juni 2026 uit — nergens meer gebruiken. RPD reset middernacht Pacific; safety-defaults staan op OFF (zelf aanscherpen voor consumenten-app).

Bronnen: ai.google.dev/gemini-api/docs/{quickstart, models, whats-new-gemini-3.5, pricing, image-generation, video, imagen, structured-output, files, rate-limits, billing, batch-api, caching, tokens, aistudio-build-mode, aistudio-fullstack, aistudio-deploying, safety-settings} — opgehaald 2026-06-12.