Eva API Client Specification

Account setup API integration and usage billing

Client edition • October 22, 2026

Use this specification to create an account, obtain an API key, purchase Stripe credits, and integrate live text, image and video generation. It covers request and response contracts, pricing, limits, errors and account management. This edition replaces the earlier flat-rate specification and video-pause notice.

Create your account and key

1. Open https://evaxapi.com/developers/signup. Existing GoAI accounts can sign in at https://evaxapi.com/developers/signin.
2. Open API keys in your dashboard. Create a key for your application and copy it immediately; the complete secret is shown only once.
3. Store the key on your server as EVA_API_KEY. Never embed it in browser code or commit it to a repository. Revoke a key in the dashboard if it is exposed.
4. Use the $10 welcome credit, then open Billing & payments to buy a $10, $25 or $50 top-up through Stripe. Saving a card alone does not add credit.

Your rate card

Generation	Eva price	Selected comparison
Text input	$1 / million tokens	Half GPT-6.1 Sol standard input
Text output	$5 / million tokens	Half GPT-6.1 Sol standard output
Image	$0.0265 each	Half GPT Image 2 medium output
Video	$0.05 / requested second	Half historical Sora 2 720p

Credit is a dollar balance, not a fixed request allowance. Rates are benchmarked as of October 5, 2026. The accompanying Pricing Guide explains the comparison, worked examples and sources. Models, output quality and resolutions differ; these are unit-price comparisons.

Current availability

Text, image and video generation are enabled. Video uses FastWan2.1-T2V-1.3B and supports 3, 5, 10 and 15 seconds. Check GET /v1/capabilities for current availability before offering a generation type. Request examples and response examples are provided below.

Base URL and dashboard

```
API base:   https://evaxapi.com/v1
Dashboard:  https://evaxapi.com/developers/dashboard
Docs:       https://evaxapi.com/developers/docs
```

The /developers/v1 prefix remains supported. Returned video status and download URLs may use that prefix; resolve relative paths against https://evaxapi.com without adding /v1 a second time.

Authentication and text generation

Every generation call requires Authorization: Bearer <your Eva key> and Content-Type: application/json. Supply a unique Idempotency-Key for each new text, image or video generation. Use 8–128 letters, digits, dots, colons, underscores or hyphens.

Send a text request

```
curl https://evaxapi.com/v1/generate \
  -H "Authorization: Bearer $EVA_API_KEY" \
  -H "Content-Type: application/json" \
  -H "Idempotency-Key: example-text-20261005-001" \
  -d '{"prompt":"Explain what an API is.","max_new":256}'
```

Field	Accepted value	Default
prompt	Nonempty text; up to 24,000 UTF-8 bytes	Required
max_new	Integer from 1 to 8,192	256
temperature	Number from 0 to 1.3	0

Only these request fields are supported. Streaming and tool calls are not exposed. A successful response includes id, object, model, text, usage and prompt_truncated. Handle 200 results and 202 queued jobs. See the queue contract at the end of this specification.

Read the billing fields

Usage field	Meaning
cost_usd	Final wallet deduction for this successful request
pricing_version	openai-half-2026-10-05 for the current schedule
billing_input_tokens	Billable prompt tokens counted with o200k_base
billing_output_tokens	Billable returned-text tokens, capped at max_new
output_tokens	Native model token count; may differ from billing count
input_tokens	null; native input-token count is not reported
free_credit_used_cents	Legacy field name; decimal cents spent from the shared wallet

In prepaid mode, billable_requests is 0 because the call creates no separate Stripe usage charge. It does not mean generation is free: cost_usd is deducted from your welcome or purchased balance.

Eva reserves prompt cost plus the output cap before generating. Unused reserved credit is released on success. Truncated prompts have no input charge. For example, 1,000 billable input tokens and 500 billable output tokens cost $0.0035.

Images and video jobs

Generate an image

```
curl https://evaxapi.com/v1/images/generations \
  -H "Authorization: Bearer $EVA_API_KEY" \
  -H "Content-Type: application/json" \
  -H "Idempotency-Key: example-image-20261005-001" \
  -d '{"prompt":"A sailboat on a blue lake","seed":901}'
```

Supply prompt with 1–1,000 characters. Optional fields are seed (integer 0–2,147,483,647; default 901), steps (integer 10–100; default 30), and guidance (number 1–10; default 7). The model may use fixed sampling steps.

Decode data[0].b64_json to save the PNG. Read mime_type, width, height and seed from the response. The verified output is 512 × 512; use returned dimensions rather than assuming a size. One successful image costs $0.0265, including its prompt. Image bytes are returned directly and are not stored in the billing ledger.

Submit and manage video jobs

Video generation is live. The worker creates silent MP4 clips with FastWan2.1-T2V-1.3B. Jobs can be accessed only with an active API key belonging to the submitting account; other accounts receive 404.

```
POST /v1/videos/generations
{"prompt":"Ocean waves at sunset","seconds":3}
```

Send the same authentication and idempotency headers as above. The API contract accepts prompt of 1–1,000 characters and seconds of 3, 5, 10 or 15 (default 3). A 202 response returns id and status_url. Check live capabilities before use; the worker must support the requested duration.

Method and path	Purpose
GET /v1/videos/{id}	Poll status every 15–30 seconds
GET /v1/videos/{id}/file	Download completed MP4 with bearer authentication
POST /v1/videos/{id}/cancel	Cancel queued or running work
DELETE /v1/videos/{id}	Delete a finished or cancelled video

A completed job includes download_url. The output format is silent 832 × 480 MP4 at 16 fps. Video prompts and files are retained by the worker for up to 30 days unless deleted sooner. Save completed media in your application.

Video reserves $0.05 per requested second and consumes the reservation on completion. Polls, downloads and deletion add no generation charge. Failure or cancellation releases the reservation; deleting a completed clip does not refund it. Completion time depends on duration and queue load. A 3-second production test completed in about 63 seconds; this is an observation, not a latency guarantee.

Balances limits and recovery

Manage your balance

Billing & payments shows available and reserved credit and recent purchases. Credits are shared by all your keys and do not expire or renew each month. Stripe confirms payment before purchased funds are added. There is no automatic refill or automatic conversion to subscription billing.

The legacy free_credit object represents the shared wallet. Its *_micros fields are exact microdollars (1,000,000 = $1); cents fields may contain decimals. Historical jobs keep their original price and model identity.

Respect service and account limits

Generation attempts are limited to 30 per minute per account. The monthly request limit defaults to 1,000 and can be changed in Usage up to 100,000; this cap is separate from dollars remaining. Pending and uncertain requests count toward admission limits. Text and images use separate queues; video permits one active job per account. Worker capacity and storage limits also apply.

Handle errors without duplicate work

Status	What to do
400	Correct invalid fields or a missing or malformed idempotency key.
401	Check that the bearer key is valid and has not been revoked.
402	Add credit or reduce the required reservation, such as max_new.
409	Inspect the existing request ID and state. Do not assume a new generation occurred.
404 or 410	Video is unknown, belongs to another account, expired or deleted.
413 or 416	Reduce request body size or correct the download byte range.
422	The model rejected the request. Correct the input before a new generation.
429	Wait for capacity or rate limits, or adjust the monthly cap.
502 or 503	Inspect the error code. Uncertain work may retain a reservation; unavailable video is not admitted.

Retry an unconfirmed submission with the same Idempotency-Key. Queued text/image retries return the existing job or retained result; changed bodies and repeated video submissions return 409. For uncertain video submissions, poll the returned status URL. Avoid a new key simply to retry a timeout, because that creates a separate request.

Error response shape

```
{"error":{"code":"invalid_request","message":"Details here"}}
```

Use HTTP status and error.code for recovery. Messages are explanatory and may change. Video errors may also include id and status_url; retain them for polling.

Pricing and comparison with OpenAI

All prices are in USD. Eva uses a prepaid dollar balance. Each new account receives $10 in welcome credit without a card; there is no fixed number of free requests. Published rates are 50% of the selected October 5, 2026 OpenAI benchmark unit rates, as shown below.

Usage unit	Eva	OpenAI benchmark
1 million input tokens	$1.00	$2.00 · GPT-6.1 Sol
1 million output tokens	$5.00	$10.00 · GPT-6.1 Sol
One completed image	$0.0265	$0.053 · GPT Image 2 medium
One requested video second	$0.05	$0.10 · Historical Sora 2

Text compares standard uncached short-context pricing. The image benchmark is 1024 × 1024 output only; OpenAI input charges are additional. Eva includes the image prompt in its price and currently returns 512 × 512 output. Video compares historical Sora 2 720p pricing; that API retired September 24, 2026. Eva video is silent 832 × 480. These comparisons do not imply equal models, quality, resolution, speed or token counts, or a discount against every competitor.

Cached-input, Batch, Flex, tool and long-context rates are outside this comparison. Eva rates are fixed to the selected benchmark and do not automatically track later OpenAI changes.

Estimate your usage

Quantity	Eva cost
1,000 billable input + 500 output tokens	$0.0035
1,000 requests of that size	$3.50
100 completed images	$2.65
3 / 5 / 10 / 15 video seconds	$0.15 / $0.25 / $0.50 / $0.75

Text cost = (billable input tokens × $1 + billable output tokens × $5) ÷ 1,000,000. A text request with 1,000 input tokens and max_new=1,024 reserves $0.00612. If output uses 500 billable tokens, it costs $0.0035 and releases $0.00262. There is no one-cent minimum per request. Examples exclude taxes and are arithmetic estimates, not forecasts of model output.

Reference sources

Eva live rate card https://evaxapi.com/v1/pricing

OpenAI text benchmark https://developers.openai.com/api/docs/pricing

OpenAI image benchmark https://developers.openai.com/api/docs/guides/image-generation

Historical Sora comparison https://developers.openai.com/api/docs/models/sora-2-pro

OpenAI retirement notice https://developers.openai.com/api/docs/deprecations

Generation and discovery endpoint index

Use https://evaxapi.com/v1 as the API base. All endpoints below require an Eva bearer key except GET /pricing. Generation POST requests also require a valid Idempotency-Key and JSON body. Endpoint paths are relative to /v1.

Method and path	Successful response
GET /pricing	200 · Published versioned rate card
GET /models	200 · Enabled text-model records in data[]
GET /capabilities	200 · Text, image and video availability and formats
POST /generate	200 · Text result or 202 · Queued job
POST /images/generations	200 · PNG JSON or 202 · Queued job
POST /images/generate	200 / 202 · Image compatibility alias
POST /videos/generations	202 · Video job, state and status_url
GET /videos/{id}	200 · Current job state and usage
GET /videos/{id}/file	200 · MP4; 206 for a valid byte range
POST /videos/{id}/cancel	200 · Updated job state; send {}
DELETE /videos/{id}	200 · {"deleted":true,"id":"..."}

Discover capabilities

```
curl https://evaxapi.com/v1/capabilities \
  -H "Authorization: Bearer $EVA_API_KEY"
```

The response contains text, images and videos objects, each with supported and endpoint. Images also describe output_format, response_format and images_per_request. Video adds asynchronous, model, durations_seconds, width, height, audio and retention_days. GET /models lists the enabled text release, not a selectable catalog of all media models. The generation APIs do not accept a model parameter.

Transport and compatibility

Use HTTPS and JSON request bodies. API errors are JSON; video downloads are binary MP4. Text/image generation waits up to 15 seconds, then returns a 202 job if unfinished. Prefer: respond-async returns a job immediately. Video submission returns a job promptly; use polling rather than holding a generation connection open.

The /developers/v1 prefix remains supported on evaxapi.com. The legacy goai.evax.live/developers/v1 host is also retained; new integrations should use evaxapi.com/v1. Returned video URLs may use the compatible prefix. Do not concatenate /v1 onto a returned path that already contains it.

This is Eva's own REST contract. It is not a drop-in OpenAI Chat Completions or Responses endpoint. Streaming, client-supplied tools, image editing and batch generation are not exposed by this developer API.

Text and image response examples

These examples illustrate field structure. IDs and generated content are examples; token counts, model identifiers and media dimensions must be read from each actual response.

Text response

```
{
  "id": "00000000-0000-4000-8000-000000000001",
  "object": "generation",
  "model": "eva-mvp-derived-english-v9",
  "text": "Generated answer appears here.",
  "usage": {
    "input_tokens": null, "output_tokens": 20,
    "billable_requests": 0,
    "free_credit_used_cents": 0.011,
    "cost_usd": 0.00011,
    "pricing_version": "openai-half-2026-10-05",
    "billing_input_tokens": 10, "billing_output_tokens": 20
  },
  "prompt_truncated": false
}
```

The monetary example uses 10 input and 20 output billing tokens: $0.00001 + $0.00010 = $0.00011. Native output_tokens can differ. input_tokens is null because the worker does not report native input counts. If prompt_truncated is true, input billing is waived. Always use cost_usd and billing_* fields for reconciliation.

Image response

```
{
  "id": "00000000-0000-4000-8000-000000000002",
  "created": 1791298800,
  "model": "eva-mvp-derived-english-v9",
  "data": [{
    "b64_json": "<base64 PNG bytes>", "mime_type": "image/png",
    "width": 512, "height": 512, "seed": 901
  }],
  "usage": {
    "billable_requests": 0,
    "free_credit_used_cents": 2.65, "cost_usd": 0.0265,
    "pricing_version": "openai-half-2026-10-05",
    "billing_input_tokens": 0, "billing_output_tokens": 0
  }
}
```

Decode b64_json as base64 binary, not a URL. Save the PNG in your application. Queued text/image results can be retrieved or replayed for 24 hours after completion. created and other timestamp fields use Unix seconds unless their names indicate otherwise.

Video response and lifecycle examples

Submit a video

```
curl https://evaxapi.com/v1/videos/generations \
  -H "Authorization: Bearer $EVA_API_KEY" \
  -H "Content-Type: application/json" \
  -H "Idempotency-Key: example-video-20261006-001" \
  -d '{"prompt":"Ocean waves at sunset","seconds":3}'
```

A successful 202 submission returns the same job structure used by status polling. state starts as queued or running. usage.cost_usd is 0 before completion, and free_credit_reserved_cents is 15 for a three-second clip. The configured $0.15 is charged only on completion.

Completed job example

```
{
  "id": "00000000-0000-4000-8000-000000000003",
  "object": "video", "model": "FastWan2.1-T2V-1.3B",
  "state": "completed", "width": 832, "height": 480,
  "frames": 49, "fps": 16, "expires_at": 1793890800,
  "step": 3, "steps": 3,
  "status_url": "/developers/v1/videos/<id>",
  "download_url": "/developers/v1/videos/<id>/file",
  "usage": {
    "billable_requests": 0, "cost_usd": 0.15,
    "pricing_version": "openai-half-2026-10-05",
    "video_seconds": 3, "free_credit_used_cents": 15,
    "free_credit_reserved_cents": 0
  }
}
```

Substitute the returned job ID into URLs; <id> is an explanatory placeholder. Progress fields may be omitted while queued. expires_at may be null when unavailable. Frame quantization can make the encoded duration slightly different from the request; billing uses requested seconds. The verified three-second clip contained 49 frames at 16 fps, approximately 3.063 encoded seconds.

Download or cancel

```
curl "https://evaxapi.com/v1/videos/$JOB_ID/file" \
  -H "Authorization: Bearer $EVA_API_KEY" -o video.mp4
curl -X POST "https://evaxapi.com/v1/videos/$JOB_ID/cancel" \
  -H "Authorization: Bearer $EVA_API_KEY" \
  -H "Content-Type: application/json" -d '{}'
```

For partial downloads, send Range: bytes=0-999. Valid ranges return 206 with Content-Range; invalid ranges return 416. Downloading before completion returns 409. Cancel before deleting a queued/running job. A deleted job returns 410; unknown or other-account jobs return 404. Read, download, cancel and delete requests are limited to 120 per minute per account.

On video_submission_uncertain, retain the returned id/status_url and poll that job. Automatic reconciliation continues even if the client disconnects. Do not submit a second job with a new key merely to recover from a timeout.

Account and Stripe payment reference

Use the dashboard for account, key and payment management. The routes below describe its browser-session contract; they do not accept developer bearer keys as a substitute for login. Paths are relative to https://evaxapi.com/developers/api. Mutations require the trusted website Origin and the session cookie. Never put a customer password, cookie or API key in browser code for a public third-party app.

Method and path	Request and response
POST /auth/signup or /auth/login	{email,password}; creates a secure session cookie
POST /auth/logout	{}; ends the browser session
GET /account	Account, keys, usage, daily/recent activity, wallet, capabilities
PATCH /account	{monthly_limit:1000}; returns updated limit
POST /keys	{name:"My app"}; 201 with id and one-time secret
DELETE /keys	{id:"key ID"}; {revoked:true}
GET /billing/credits	enabled, packs, balance, recent purchases
POST /billing/credits	{amount_cents:1000}, 2500 or 5000; checkout url
POST /billing/credits	{session_id:"cs_..."}; verifies payment and fulfillment
GET /billing/payment-method	enabled and masked card details or null
POST /billing/setup	{} creates/reuses card setup; returns url
POST /billing/setup	{session_id:"cs_..."}; confirms setup as {saved:true}
POST /billing/portal	{}; returns Stripe customer portal url

GET /account includes keys (metadata only), current UTC month, usage totals, daily activity and the 25 most recent requests. Recent entries show request_kind, state, native and billing tokens, pricing_version and cost_or_reserved_usd. The free_credit object includes available_micros, spent_micros and reserved_micros. A pending cost is a reservation, not a completed charge.

Payments security and operational limits

Payment flow

Choose a $10, $25 or $50 credit pack, follow the returned Stripe Checkout URL, and complete payment. On return, the dashboard confirms the session. Stripe webhook processing also fulfills confirmed purchases. A redirect or unpaid session alone does not grant credit. Repeated confirmation cannot credit the same purchase twice; only the owning account can confirm it.

Saving a card is separate: it adds no balance, starts no subscription and makes no usage charge. Manage saved payment details through Stripe's portal after billing has been started. Refunds and payment disputes reduce credit available to spend. Current prepaid mode does not use the legacy /billing/checkout subscription route; it returns 409.

Operational and security notes

Key names contain 1–64 characters. Each account can hold 20 active keys; key creation is limited to 5/minute. Signup/login/logout attempts share a 10/minute source-IP limit. Credit checkout is limited to 10/minute/account; card setup to 5/minute/account. Key revocation prevents new requests, but work already admitted may finish.

Stripe posts signed events to /billing/webhook. This is a provider callback, not a client integration endpoint; clients do not need a Stripe secret or their own webhook. The internal reconciliation endpoint is not publicly accessible. The ledger keeps accounting metadata. Queued prompts and results use temporary encrypted storage (see queue contract); video retains prompts and files for up to 30 days.

Limits at a glance

Operation	Limit
Generation attempts	30/minute/account
Video read download cancel or delete	120/minute/account
Monthly admitted requests	Default 1,000; configurable from 1 to 100,000
Active API keys	20/account
Text or image request JSON body	32 KiB text; 16 KiB image
Video request JSON body	8 KiB

Request body limits include JSON syntax and headers are not part of this body allowance. Prompts have the tighter field limits described earlier. Separate worker queue, storage and service availability limits can reject work even when your balance and monthly allowance are sufficient.

Production checklist

Keep API keys on your server, use one key per application, and revoke exposed keys. Save text, PNG and MP4 outputs in your own storage. Persist video IDs and idempotency keys before retries. Handle 402 and 429 without creating duplicate work, and check capability availability before enabling media features. Verify completed cost_usd against your usage records; retain fractional cents rather than rounding every request.

Latest specification and documentation https://evaxapi.com/developers/docs

Durable text and image queue

Text and image submissions enter a persistent queue. Use the existing generation endpoint, body, bearer key and Idempotency-Key. Add Prefer: respond-async to return immediately. Without it, the API waits up to 15 seconds: 200 contains the normal result; 202 contains a job. Clients must handle both outcomes.

Method and path	Purpose
GET /v1/requests/{id}	Read job status; poll every 2–5 seconds
GET /v1/requests/{id}/result	Retrieve the original text or image JSON
POST /v1/requests/{id}/cancel	Cancel waiting work; running work returns 409

```
{
  "id": "<request ID>", "object": "generation_job",
  "kind": "text", "state": "queued", "attempts": 0,
  "status_url": "/developers/v1/requests/<id>",
  "result_url": "/developers/v1/requests/<id>/result",
  "result_available": false,
  "usage": {"reserved_usd": 0.001, "cost_usd": 0}
}
```

Envelopes also include created_at, started_at and expires_at (Unix seconds; started_at may be null). States are queued, running, completed, failed, cancelled and indeterminate. A pending result read returns 202; an unsuccessful terminal result returns 409; expired successful results return 410. Keep bearer authentication on every call; other accounts receive 404.

Capacity and fairness

Each lane runs one job: text accepts up to 100 outstanding jobs and images up to 50. Each account may hold 10 combined outstanding jobs. Accounts take turns; ready jobs within an account keep arrival order. A waiting job expires after 15 minutes; work started before that deadline may finish later. Queue reads and cancellations share 120 calls/minute/account. Existing generation and monthly limits still apply. Full queues return 429 with Retry-After.

Credit and retries

Credit is reserved when accepted and settled once on success. Known failure, expiry or cancellation releases it. Explicit busy rejections retry up to 12 attempts within the queue deadline; uncertain completion retains credit and requires operator review. Reusing a key with the same body returns the existing job or retained result, never new work. Async retries of completed jobs return a 200 envelope; synchronous retries return the cached result. Changed bodies return 409. Video retains its existing contract.

Storage and retention

Prompts are encrypted while waiting/running and removed from the live queue record afterward. Completed results are encrypted and retained for 24 hours; save them in your application. Accounting metadata remains. Backups have their own retention lifecycle. Queue admission, dispatch and cancellation survive API restarts. Interrupted running work is marked uncertain rather than blindly replayed.