FEVER TECHNICAL SPECIFICATION

Fourier Encoder for Visual Embedding and Retrieval

What images are your customers uploading?

Need to know fast? On their behalf?

FEVER is a special image retrieval model from Lowdown Labs, set into a full retrieval product via a CloudFormation stack. It runs inside your VPC, indexes your S3 asset buckets, and can provide full text -> image search functionality without endless configurations and tuning. Nothing leaves your account. Nothing is metered per call. You can use our API in your application at scale and customers can reduce time spent manually organizing, labelling, and tagging image content. You build nothing.

AT A GLANCE

Deploy	CloudFormation
Data	Never leaves your account
Compute	CPU native, GPU optional
Database	RDS, Aurora, or in stack
Interfaces	REST, MCP, console
Isolation	Per customer scoped keys
Egress	None required

THROUGHPUT AND COST, MEASURED

Embed **51** images/sec on one CPU node, **2,666** on one GPU. Query 2.9/sec on CPU and 226/sec on GPU. Hardware: c7i.4xlarge int8, g6e.xlarge L40S fp16.

COST TO INDEX A CORPUS, ORDER OF MAGNITUDE. EXTRAPOLATED FROM ONE MEASURED RUN ON 10,000 EXAMPLE IMAGES.

Corpus	Time	Cost
10 thousand	∼2 min	∼$0.03
100 thousand	minutes	cents
1 million	hours	dollars
1 billion	days	thousands of dollars

A billion images indexed for a few thousand dollars of compute. **No per-image fee.**

Order of magnitude estimates, not a performance commitment or guarantee, estimated from the measured 10,000-image run above (∼2 minutes end-to-end, ∼$0.03 at on-demand list price). Embedding itself is not typically a bottleneck: a single 1x GPU node embeds ∼2,600 images/s at default settings, 1M images embed in about 6 minutes since throughput scales near linearly per added GPU and node. End to end backfill time is more likely governed by the S3 download rate, and the database backend you choose in the CloudFormation template (managed RDS/Aurora), which fixes the index write rate. The 1B row assumes a 12 node multi GPU fleet.

WHY IT RUNS ON A CPU

Attention compares every patch to every other patch. $O(n^2)$.

That is largely why everyone needs GPUs - and can't get them.

FEVER thinks in the **Fourier domain** instead. $O(n \log n)$.

And that - is the entire reason the CPU numbers reported can even exist.

SCALING

All pairs attention	$O(n^2)$
FEVER	$O(n \log n)$

Double the grid. Attention roughly quadruples. FEVER roughly doubles.

A FULL SYSTEM, NOT JUST A MODEL

Competitors ship an embedding model. You figure it out from there. FEVER gives you the entire retrieval stack.

	FEVER this appliance	**SigLIP** embedding model	**jina-clip-v2** model or API
E2E Flickr30K text to image R@1	**81.4**	75.6	87.3
Cross resolution recovery	**97.8** R@1	–	–
Runs on	your choice, your VPC	GPU	GPU or metered API

The full pipeline measured above was scored on our patent pending low resolution / high accuracy indexing system, where JINA and SigLIP are slower at higher resolution, higher latency, and higher storage cost. Neither other vendor ships a reranker, a dedup system, or cross resolution recovery. **You would otherwise build all of it.**

FEVER numbers measured on the appliance: embed plus the bundled reranker, and the 64 px cross resolution eval (find the original of a derived copy among 10,045 images). SigLIP: Flickr30K zero shot text to image R@1, FLAIR et al. 2024. Jina publishes strong CLIP benchmark retrieval; protocols differ, and neither competitor model performs natively cross resolution on task as shipped.

WHAT IT DOES

- **Search by meaning.** Plain language to image.
- **Robust duplicate detection.** Maintained at ingest, so the view is instant.
- **Easy, minimal configurations.** No magic numbers to guess.
- **Isolate your customers.** Postgres RLS + scoped keys per `customer_id` where needed.
- **Agent ready.** Bundled MCP server, same scoped keys.
- **Stay operable.** Provided console, OpenAPI clients, background jobs, CSV exports, Telemetry setup.

MODEL CARD

Component	**Specification**
Image encoder	FEVER, 342M parameters. Proprietary Fourier operator mixing with local attention.
Text encoder	SigLIP2 so400m text tower (Apache 2.0), aligned to the FEVER image space.
Reranker	BLIP ITM (BSD 3 Clause), int8 QAT, full 577 token vision sequence.
Runtime	Custom CPU optimized server, no Triton. Torch+CUDA bundled for flexible compute setups.

DEPLOYMENT AND SECURITY

- Your VPC, your subnets, your CIDR. Four inputs.
- Private subnets, internal load balancer, no public endpoint.
- Telemetry preconfigured and ready to link with your observability stack.
- No callbacks. No outbound path to reach our models.
- The dependable security of Postgres.

WHAT YOU ARE ACTUALLY BUYING

No model to choose. No vector database to configure. No data protection review.

If something breaks, you contact us, and you reach the responsible people who wrote it and will fix it :)

Your success is our success - "The Lowdown Labs Way."

AVAILABILITY

AWS Marketplace. Container product, annual contract.

dev@gimmelowdown.com