

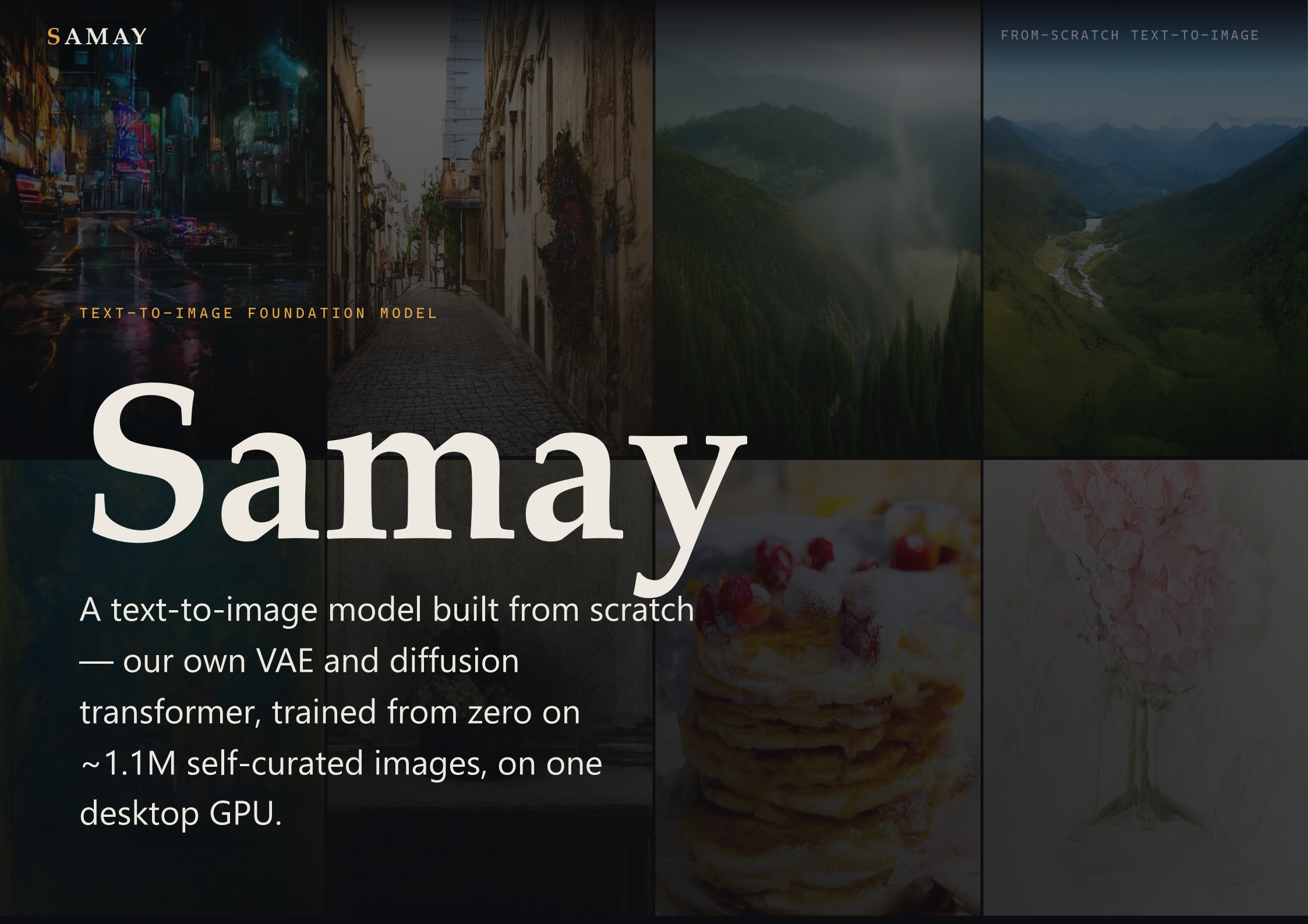
SAMAY

FROM-SCRATCH TEXT-TO-IMAGE

TEXT-TO-IMAGE FOUNDATION MODEL

Samay

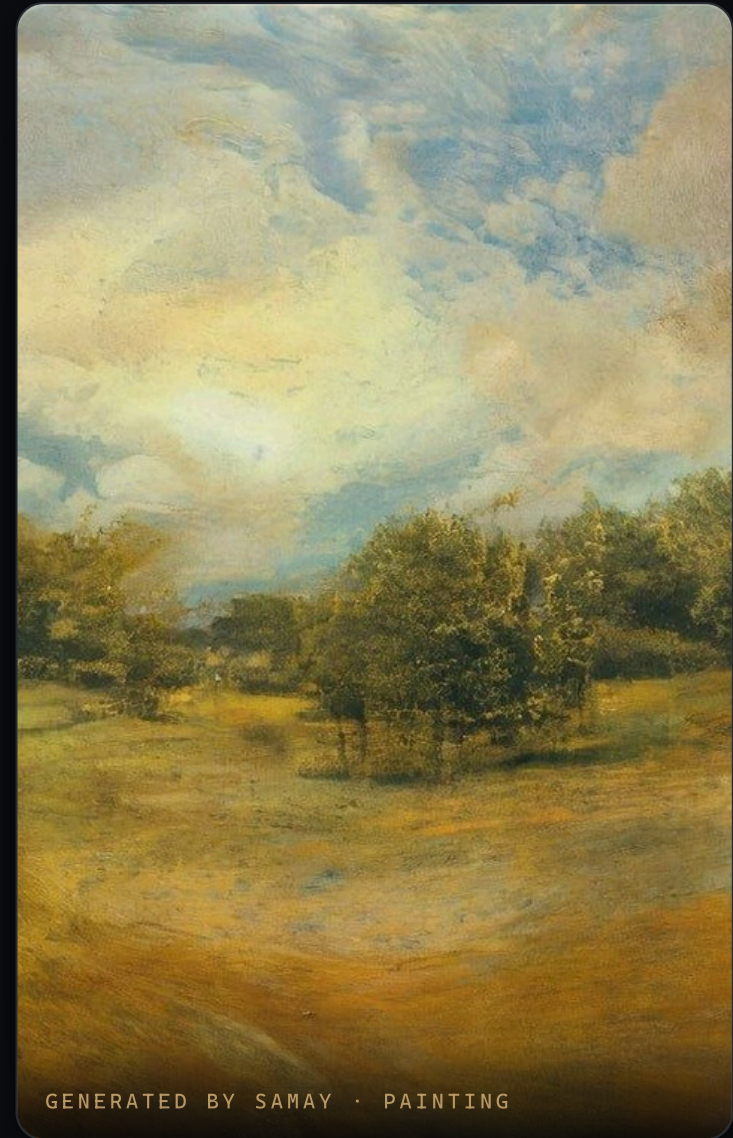
A text-to-image model built from scratch
— our own VAE and diffusion
transformer, trained from zero on
~1.1M self-curated images, on one
desktop GPU.



THE THESIS

One person. One GPU.
A foundation image
model built **from zero**
— no fine-tune, no
borrowed backbone.

EVERY IMAGE HERE WAS GENERATED BY OUR MODEL — TRAINED
FROM SCRATCH ON ~1.1M IMAGES. NOTHING IS STOCK,
NOTHING FINE-TUNED.



A full pipeline, trained end to end.

01

Our own VAE

A 16-channel latent autoencoder with self-attention — ~12:1 compression, not a borrowed Stable-Diffusion VAE.

02

Our own DiT

An attention-based diffusion transformer, flow-matching, dual-stream text conditioning — written and trained from scratch.

03

Our own data

1.1M images curated from dozens of sources, captioned and filtered in-house — the corpus is ours too.

04

Stability tricks

EMA weights + REPA representation alignment for clean, stable convergence at long horizons.

05

Attention, engineered

Full self-attention with 2D axial RoPE positions and dual-stream text↔image attention, plus register tokens — memory-efficient SDPA kernels keep every layer inside one card's VRAM.



ONE PIPELINE, FOUR DOMAINS — EVERY TILE GENERATED BY SAMAY

THE NUMBERS

Small model. Serious range.

288.65_M

Diffusion transformer
parameters

46.0_M

VAE parameters, 16-channel
latent

1.12_M

Training images, self-curated

1_{GPU}

Single ~32 GB desktop card

768_{PX}

512 stage done — 768 stage
now training

0

Pretrained weights borrowed

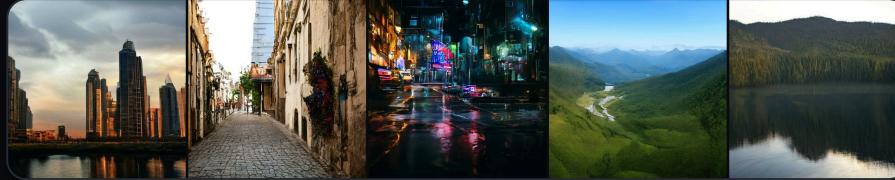
TEXT CONDITIONING VIA A FROZEN QWEN ENCODER. EVERY TRAINED WEIGHT ABOVE IS OURS.

ONE MODEL, ANY DOMAIN

It didn't specialize. It generalized.

Landscape & city

PHOTOREAL



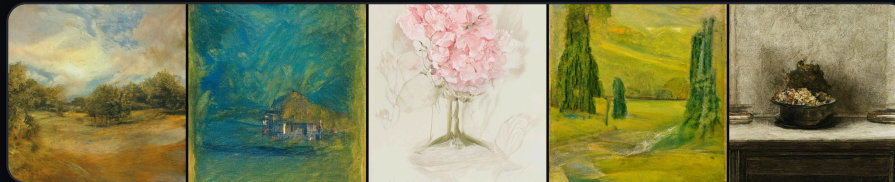
Food

PHOTOREAL



Art & illustration

MULTI-STYLE



People & portrait

EDITORIAL



AND THIS IS ONLY THE BEGINNING

YOU ARE HERE

512 finished · 768 stage training now · quality still climbing

512px · 1 GPU · solo

DONE

768px

TRAINING NOW

1024px+

NEXT

Bigger model · team

WITH FUNDING

WHY THIS IS UNUSUAL

This is the work of
a **lab and a cluster**
— done on a desk,
by one hand.

- **Memory-tuned to the last gigabyte** — training fits under the card's VRAM ceiling, no cluster to hide inefficiency.
- **Every layer is understood** — VAE, latent stats, transformer, sampler, data — all authored here, all debuggable.
- **Quality is climbing, not plateaued** — 512→768 now, EMA on, still on the steep part of the curve.



GENERATED BY SAMAY · LANDSCAPE

The recipe is proven. Give it compute and a team.

This entire stack — VAE, transformer, and data pipeline — was built and trained from scratch by one person on a single GPU. The hard, uncertain part — proving the architecture learns from zero — is already done. And this is only an early checkpoint: quality climbs with every step, and it scales directly with more compute and more data. What's left is no longer a research risk. It's resources and people.



A bigger model

The same recipe carries to billions of parameters and a richer latent — sharper faces, hands and fine detail. It just needs the compute.



Higher resolution, faster

Multi-GPU training reaches 1024px+ and cuts weeks of convergence to days.



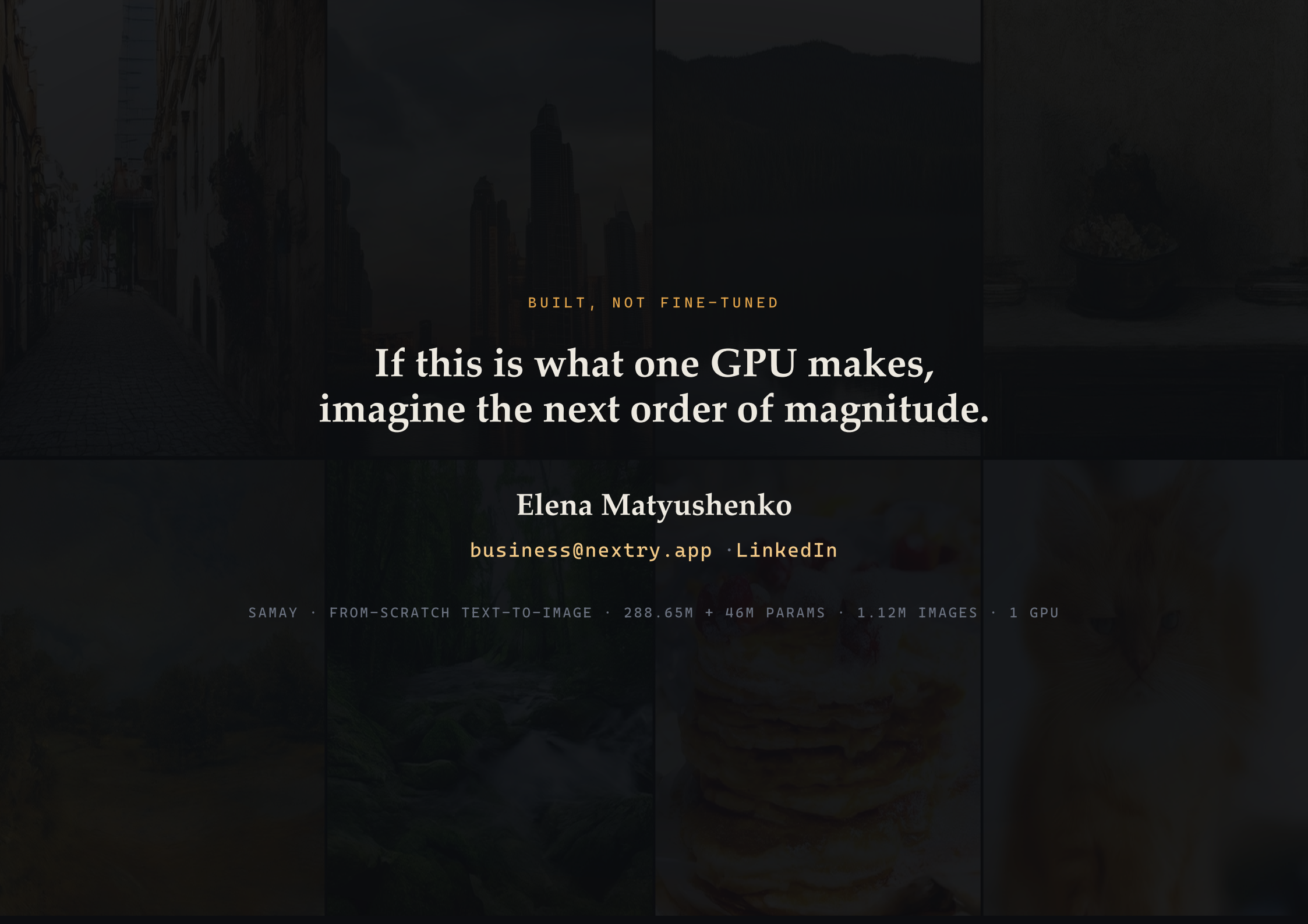
Virtual try-on & editing

The edit / reference-image stack is already architected on top of this base — a direct path to a fashion product.



A team behind it

One person got it this far. With a few strong engineers, the same stack becomes a shipping product — everything a generative-media company needs to own.



BUILT, NOT FINE-TUNED

If this is what one GPU makes,
imagine the next order of magnitude.

Elena Matyushenko

business@nextry.app · [LinkedIn](#)

SAMAY · FROM-SCRATCH TEXT-TO-IMAGE · 288.65M + 46M PARAMS · 1.12M IMAGES · 1 GPU