🎥 Hollywood + Open Source

How to make high quality VFX edits with open models

Greg Schoeninger, CEO oxen.ai

The $500,000 Problem

Oops, we filmed in the wrong jacket…

- Editors find a costume mistake
- Re-shoots are expensive (people, sets, location, etc)
- VFX is expensive, and takes a long time
- Can AI become our new VFX tool?

Goal of this talk

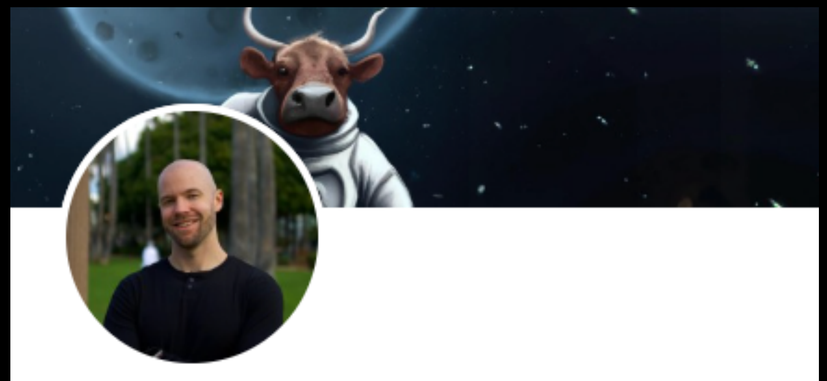
✈️ 10,000 ft view of video generation models

- Show a real world example (The $500k Problem)
- Give you a lay of the land
  - T2V, I2V, R2V, V2V, Fine-Tuning LoRAs, WAN 2.1 VACE, LTX-2, ComfyUI
- Inspire you with what’s possible
  - Shorter timelines, realistic budgets, real workflows
- Emphasize curation + artistic talent still needed

$ whoami

Greg Schoeninger

- Founder and CEO of Oxen.ai
- Training language models since 2013
- Early in generative image/video/3D models (GANs, Diffusion Models)
- Open source/weights enthusiast



Greg Schoeninger

CEO of Oxen.ai 🐂🌾

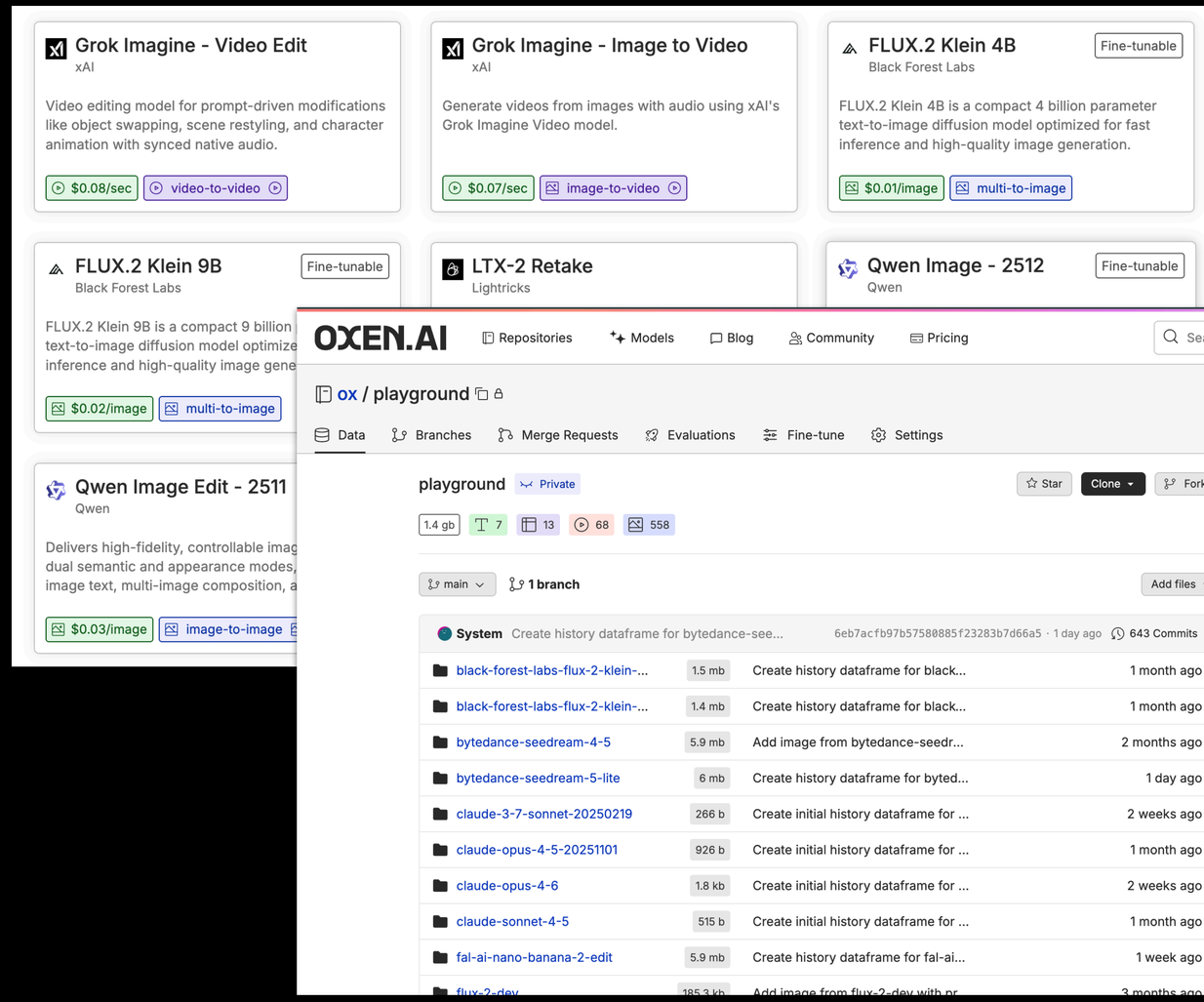
Los Angeles, California

OXEN.AI **Oxen**

Oxen.ai: Datasets, Fine-Tuning, & Inference

🐂🌾 AI infrastructure without the headache

- One-Click Fine-Tuning
- Spin Up 8xH100 in seconds
- Automatically spin down GPUs
- Deploy to on-demand inference APIs
- Version control TB of datasets
- Own your AI don’t rent it
- Try it out! (If my talk is boring)



Our Fake Movie Set

Because we can’t show the real one…

Jacket Swap

🧥 Our Training Data…

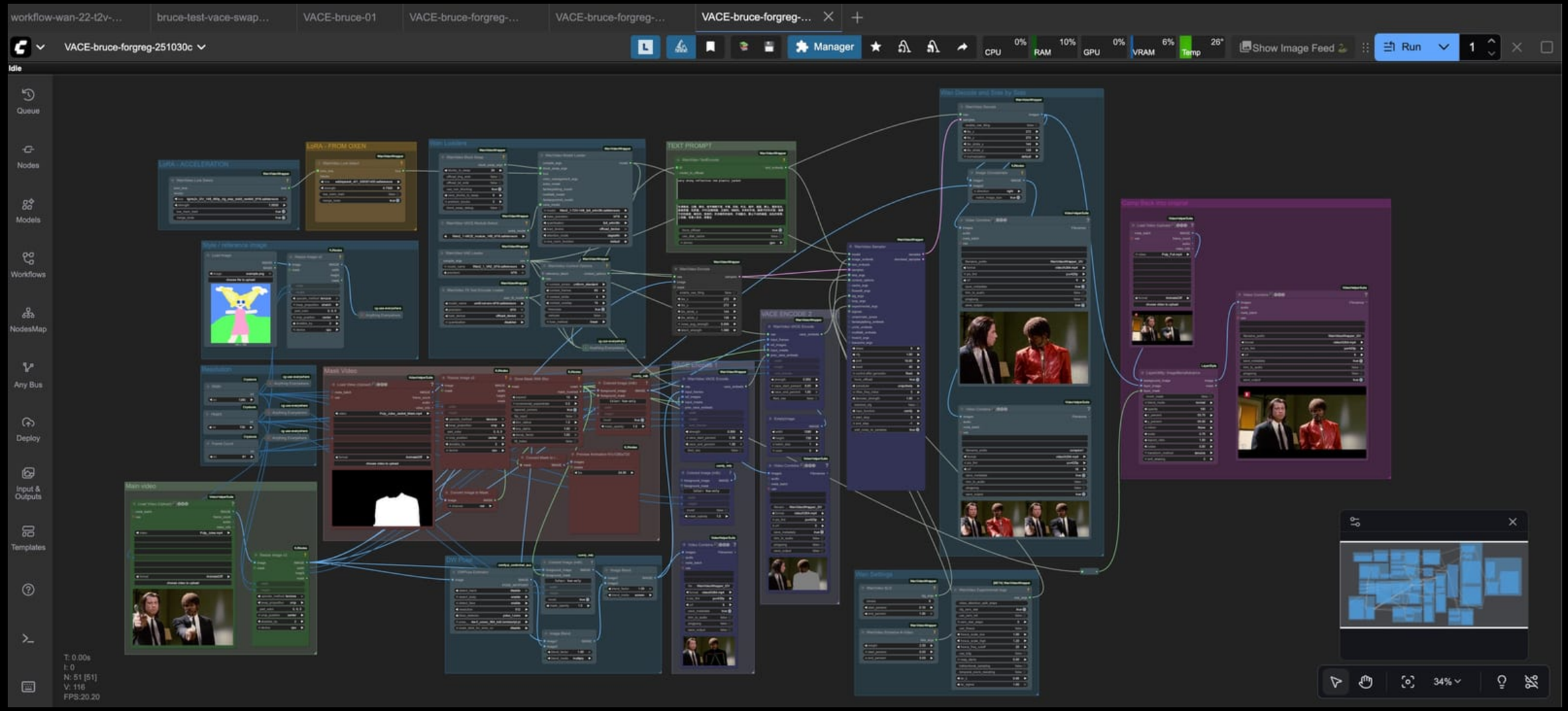
The Result 🎉

With 100% Open Weight Models

A lot has changed
in 5 months…

Professional Workflow

Are all these nodes still needed…?



 Lay of the land

Wild Wild West of Video Generation

SOTA Video Generation Models

Can be hard to keep up with….

- Open Source
  - WAN 2.1 VACE
  - WAN 2.2 A14B
  - LTX-2
  - HunyuanVideo
  - LTX-2.3 (yesterday!)
  - Helios 14B (this week!)
- Closed Source
  - Kling O3, v3
  - SORA 2
  - Runway 4.5
  - Google VEO 3.1
  - Luma / Ray
  - Seedance 2.0 (wen?)

🤫 *PS: You can try them all on Oxen.ai…DM me for free credits*

Why Open Source/Open Weights?

- Local
- Private
- Cheap
- Trainable/Customizable
- Own your IP
- Get around content filters…
- Distill from closed source 😏

Why fine-tune?

Cost, speed, accuracy, privacy

But generating image assets with Nano Banana has a few big issues:

- **Consistency:** Even with reference images, examples, and tons of prompt engineering, Nano Banana still struggles mightily to generate images with the preferred style consistently. I'd guess that it's *at best* 50/50, which is far from good enough for the estimated 40k tiles I'll need to generate.
- **Cost & Speed:** Simply put, Nano Banana is slow and quite expensive. It simply won't be possible to generate all of the tiles I'll need to generate given the cost and speed of a powerful model.

So I decided to fine-tune a smaller, faster, cheaper model. I opted to try fine-tuning a Qwen/Image-Edit model on the (wonderful) oxen.ai service and created a training dataset of ~40 input/output pairs. The fine-tuning took ~4 hours and cost ~12 bucks, and I was pretty happy with the results!

Acronyms & Principles

T2V, I2V, R2V, WAN 2.1 VACE + LoRA

What does it all mean?

- T2V: Text to video, your traditional prompt to video
- I2V: Image to video, start frame to continue into a video
  - Start frame + end frame
- R2V: Reference to video
  - “@Element1 walks into a bar with @Element2 on the table, @Element3 walks in and sits down next to them”
- A2V: Audio to video
  - Lip syncing, a tree falling in the forest, etc
- Training a LoRA
  - None of the above working? Let’s fine-tune

T2V - Text to Video (LTX-2.3)

Less used, still great for brainstorming

LTX 2.3 Pro: Text to Video

$0.12/sec | text-to-video

About | Playground | Pricing | API

Input

Prompt *

A cinematic medium close-up shot in widescreen with gritty 1990s film grain. Two men stand face-to-face in a dim, run-down room with olive-green walls and worn textures. The lighting is moody and warm, casting a yellow-green tint across the space. A faint lamp glows softly in the background between them, creating a hazy halo of light and subtle shadows.

The camera holds steady at chest height, framing both men from the chest up.

On the left stands a white man with shoulder-length dark brown hair parted in the middle, slightly wavy and greasy. His face has angular features, a strong jaw, and light stubble. He wears a black suit jacket over a white dress shirt and a loosely knotted narrow black tie. He stands in profile facing right, still and tense, staring directly at the other man with a serious, contemplative expression.

Opposite him stands a Black man with dark skin and a large, voluminous afro. He has a goatee and mustache, and a small earring in his left ear. He wears a bright red leather zip-up jacket with a mandarin collar, partially unzipped to reveal a gold chain necklace with a small pendant. He looks slightly downward at first, jaw tight, then slowly raises his gaze toward the man in front of him.

The camera subtly pushes in, tightening the frame as the two men hold eye contact. The background lamp flickers faintly, and the room's greenish walls reflect the warm light, heightening the tense, confrontational atmosphere.

Text prompt describing the video to generate

I2V - Image to Video

A picture is worth a thousand words

LTX 2.3 Pro: Image to Video

$0.12/sec | multi-to-video

About | **Playground** | Pricing | API

Input

Prompt *

The two men look up, and point their guns at the camera.

Text prompt describing the video to generate

Image

First frame image for image-to-video generation

Drag and drop or click to upload

Image Editing Models

👨‍🎨 Perfect the first and last frames

OXEN.AI Repositories Models Blog Community Pricing Search Docs $140.07

Nano Banana 2 - Image Edit

$0.08/image image-to-image

About **Playground** Pricing API

Input

Form

Prompt *

Change the the man with the afro's suit (on the right) to be the red jacket from the comedian

Text description of what you want to generate, or the instruction on how to edit t...

Input Images

Input images to transform or use as reference (supports multiple image urls)

Drag and drop or click to upload

– 1/4 +

Clear Generate Image

Generations

View Dataset

$0.08 auto 1K

Created in 29s

Change the the man with the afro's suit (on the right) to be the red jacket from the comedian

Reuse Download Delete

$0.08 auto 2K

Created in 45s

Change the the man with the afro's suit (on the right) to be the red jacket from the comedian

Reuse Download Delete

Open Source Image Editing Models

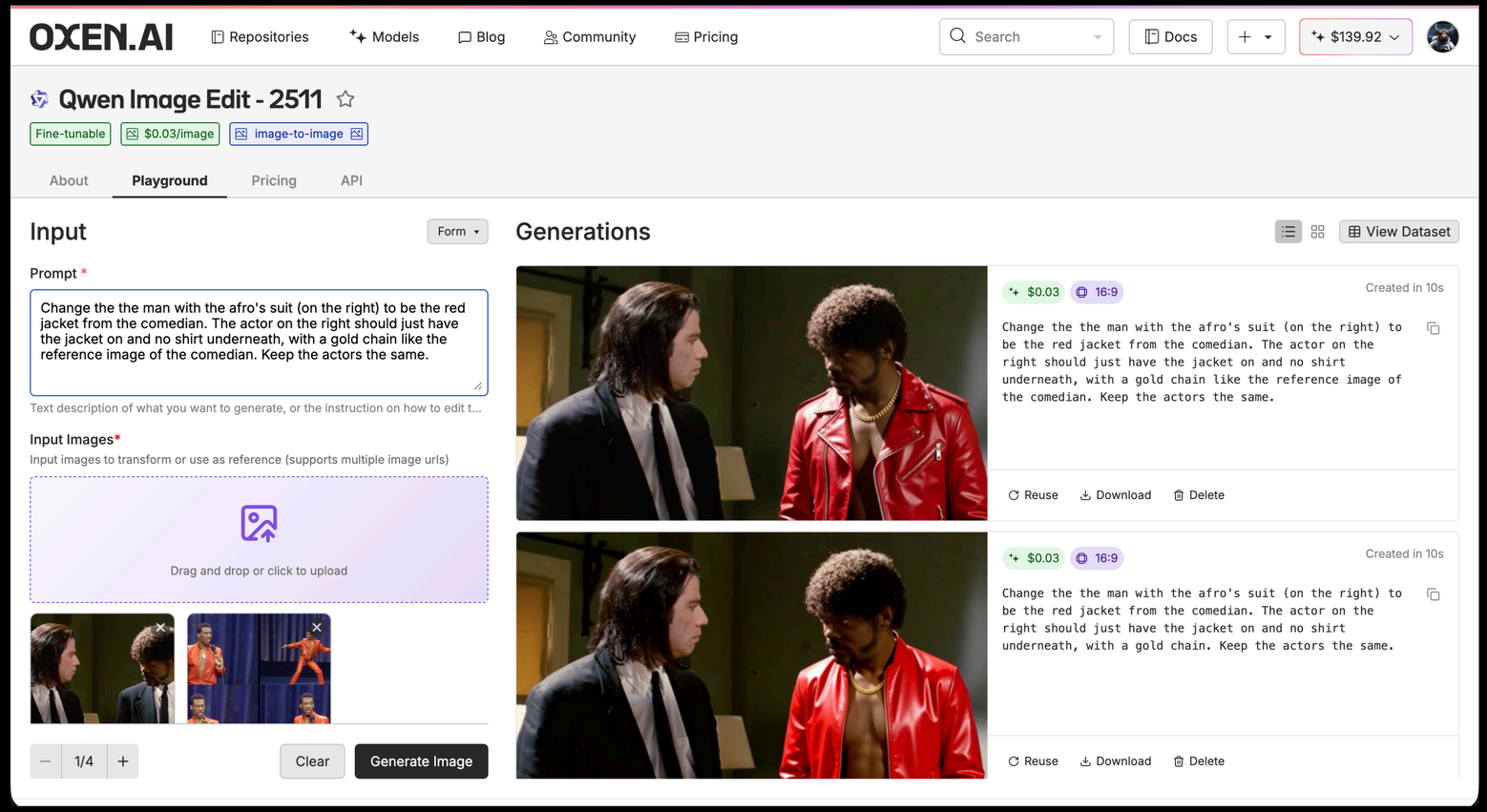
What models can we run/train ourselves?

- Qwen-Image-Edit
- Qwen-Image-Edit-2509
- Qwen-Image-Edit-2511
- Flux-2 [dev]
- Flux-2 Klein 4B & 9B

Open Source Image Editing Models

Qwen-Image-Edit - 20B Parameters

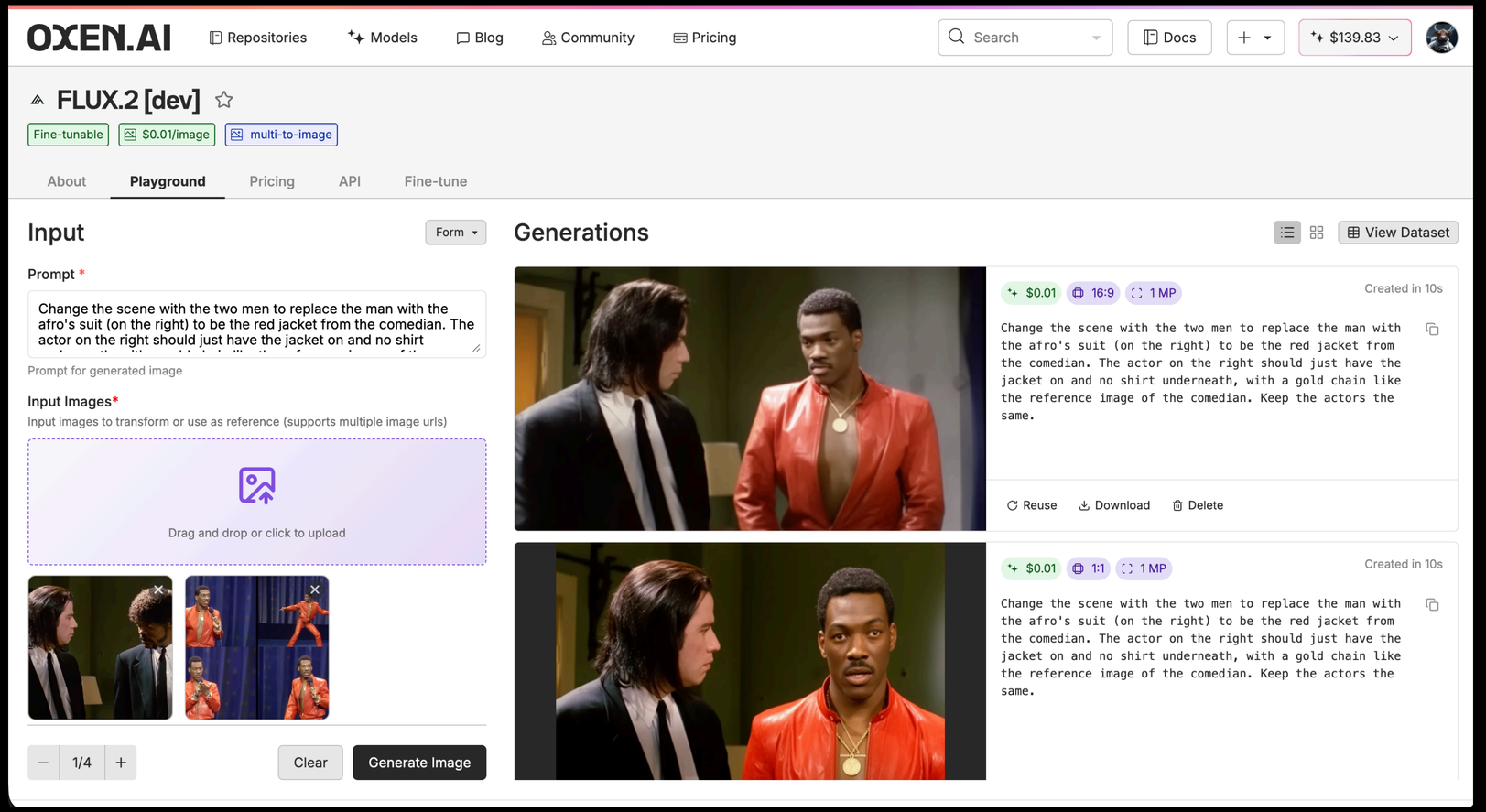
- Faster
- Cheaper
- Run Locally
- Private
- Fine-tuneable 🎉



https://www.oxen.ai/ai/models/qwen-image-edit-2511

Open Source Image Editing Models

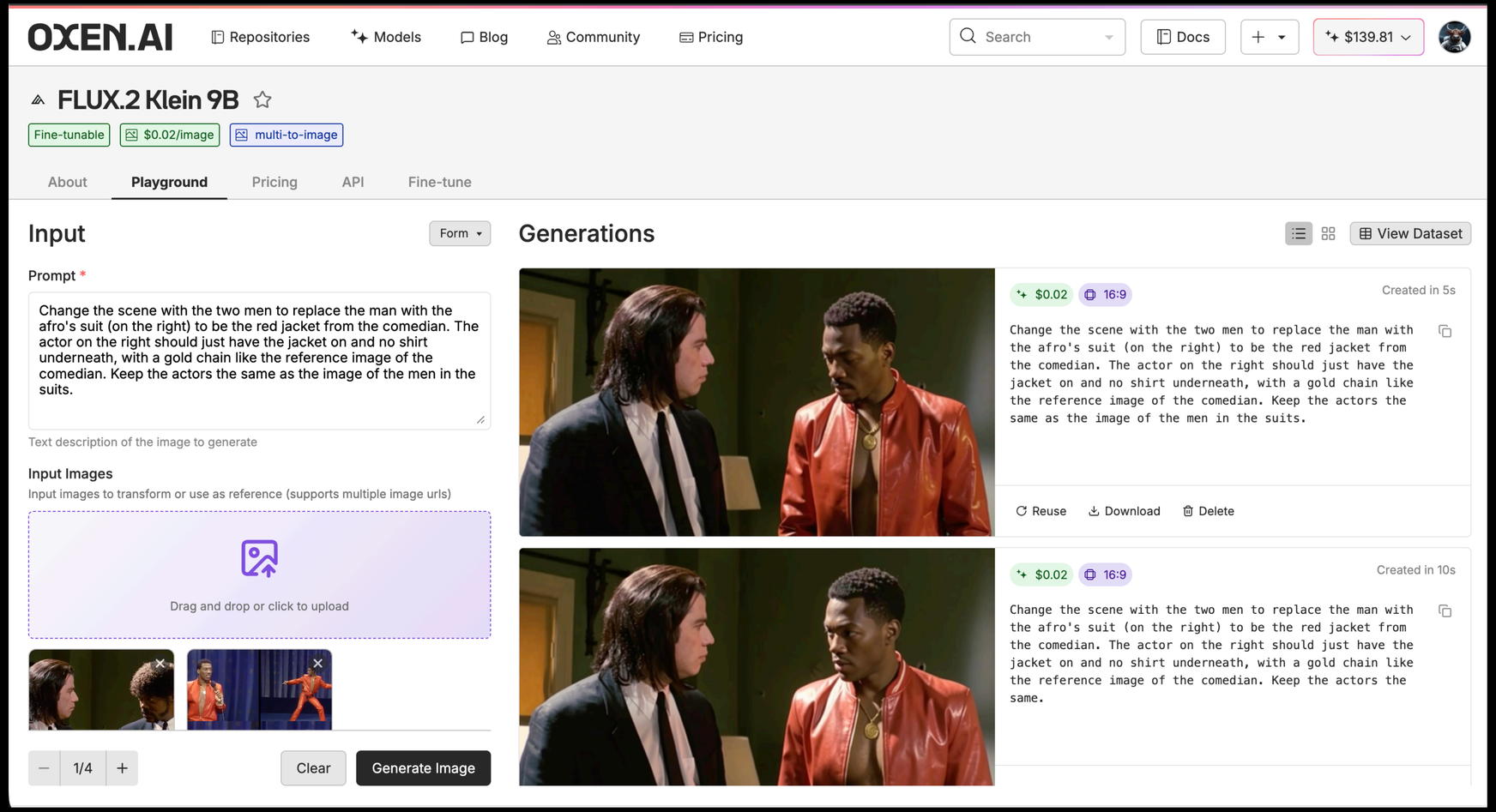
Flux-2 [dev] - 32B Parameter



https://www.oxen.ai/ai/models/flux-2-dev

Open Source Image Editing Models

Flux-2 Klein - 9B Parameter



https://www.oxen.ai/ai/models/black-forest-labs-flux-2-klein-9b

We trained our own model…
more on that later

FLF2V - First Frame Last Frame to Video

Perfect the start and end frames (LTX-2.3)

Open Source FLF2V Models

First frame to last frame, that you can train!

- LTX-2.3 (22B) (released yesterday 🎉)
- LTX-2 (19B)
- WAN 2.2 A14B
- WAN 2.1 VACE (14B)

You are 90% there…
How do you get to 100%?

WAN 2.1 VACE

Video tasks within an **A**ll-in-one framework for **C**reation and **E**diting

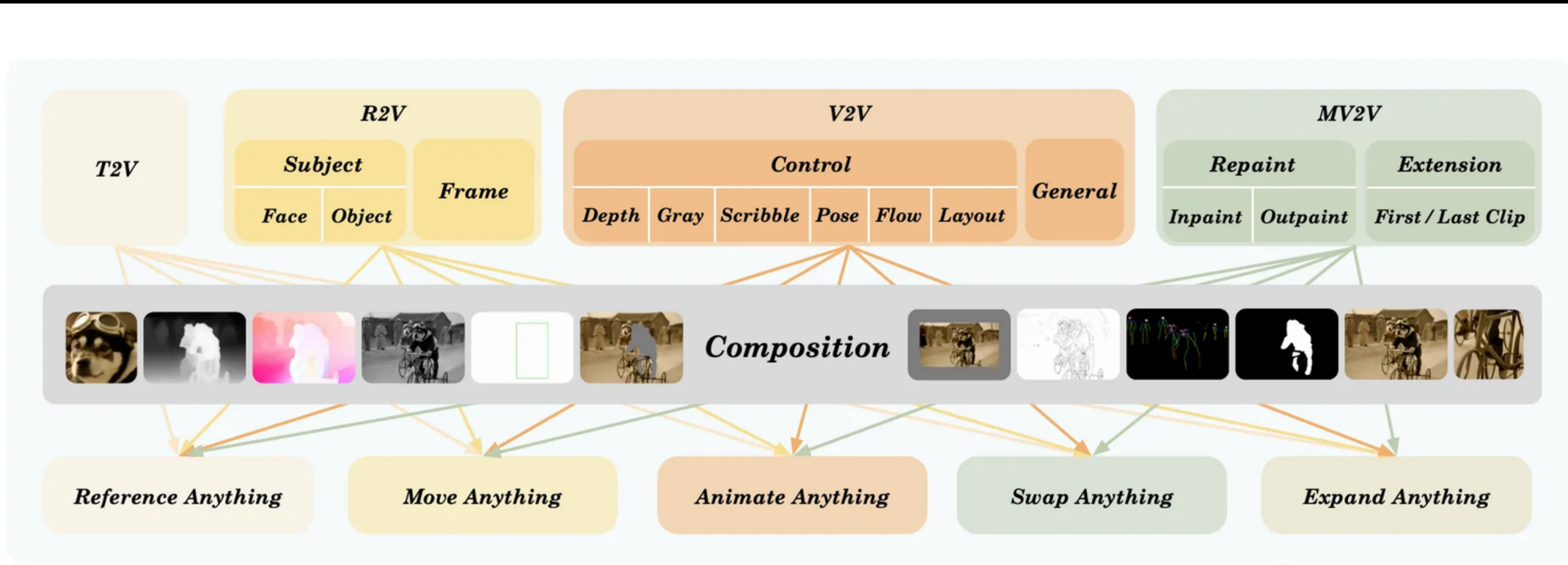


Figure 2. **Task categories covered by VACE.** Four basic tasks can be combined to create a vast number of possibilities.

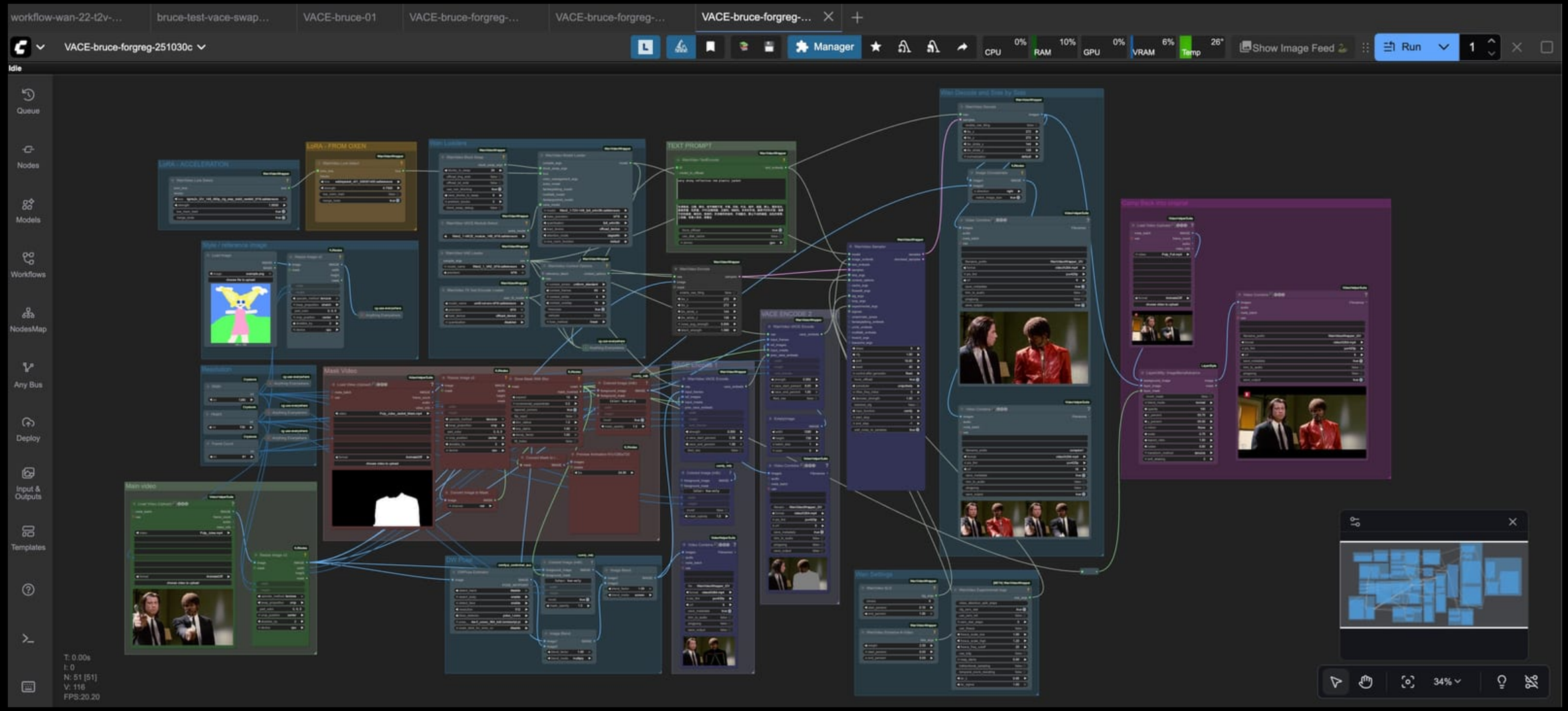
WAN 2.1 VACE is Fine-Tuneable 🎉

Teach it anything

- NOTE: Always start with prompting, it’ll be faster to iterate, explore the latent space
- Fine-tuning is important when a concept is not in the pre-training data
  - Motion
  - Characters
  - Voice
  - Style
  - Camera movements / lenses
- Labelers of massive pre-training datasets did not have great cinematic vocabulary

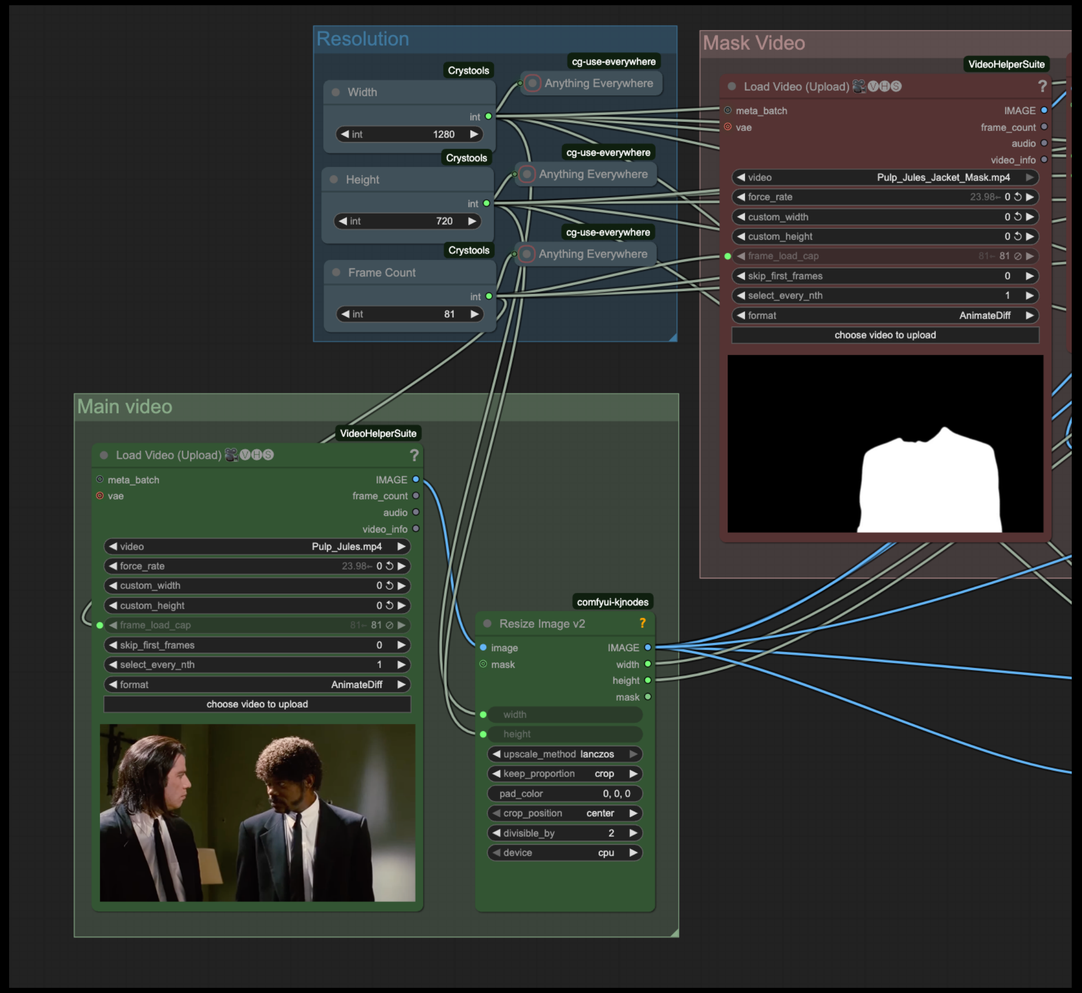
WAN 2.1 + ComfyUI

⛓️ Chaining models and workflows



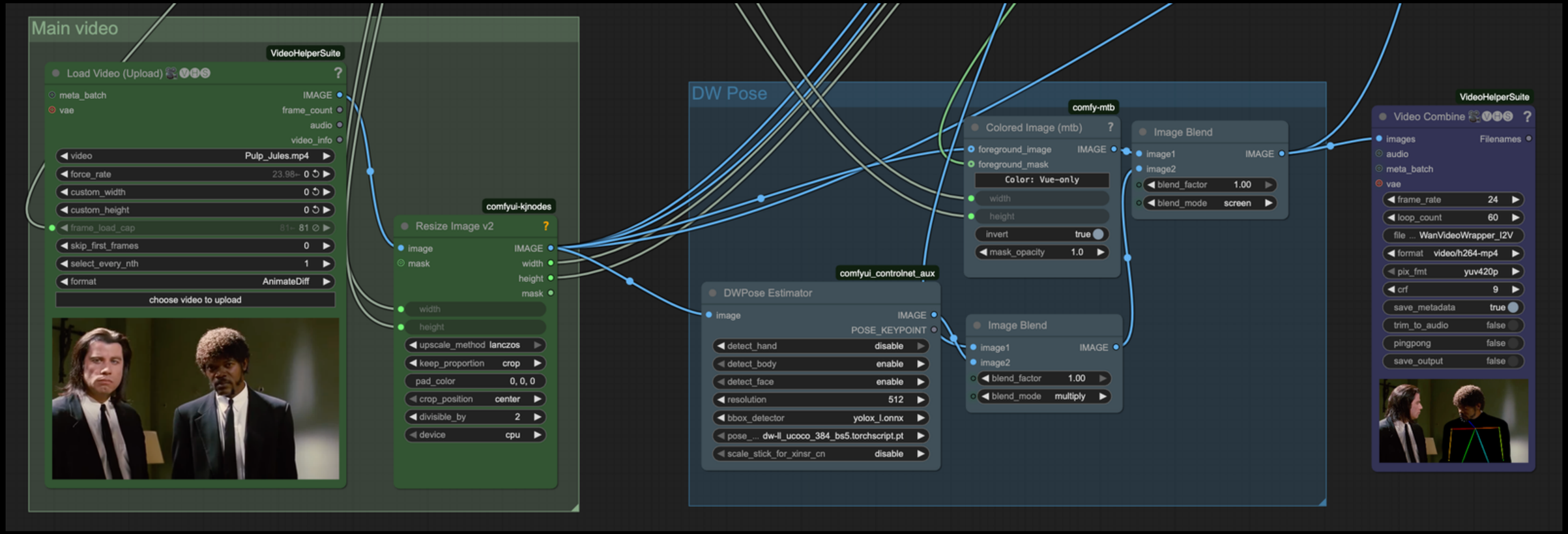
Step 1: Masking

Segment Anything 3



Step 2: Pose Estimation

OpenPose



Step 3a: Jacket Specific LoRA

Trained on Oxen.ai

OXEN.AI ox / EddieMurphyDelirious

Data | Branches | Merge Requests | Evaluations | Fine-tune | Settings

main EddieMurphyDelirious / train.parquet

ox fix row 2

Query this dataset. Ask me anything!

Schema

file_path (str)
caption (str)

2 columns, 1-100 of 137 rows

file_path	caption
frames/frame_003840.jpg	Eddie Murphy is standing confidently with one hand on his hip and the other holding a microphone, appearing relaxed ...
frames/frame_001920.jpg	Eddie Murphy is energetically dancing with arms outstretched and legs spread wide, embodying a confident and lively ...
frames/frame_000360.jpg	Eddie Murphy is standing confidently with one hand holding a microphone, appearing poised and ready to perform.
frames/frame_003930.jpg	Eddie Murphy is standing upright, holding a microphone in his right hand, and appears to be speaking or performing.
frames/frame_002190.jpg	Eddie Murphy is energetically dancing and gesturing with both arms outstretched while holding a microphone in one h...
frames/frame_003090.jpg	Eddie Murphy is standing with both hands clasped together, holding a microphone, and appears to be speaking or perf...
frames/frame_002880.jpg	Eddie Murphy is standing confidently with one hand holding a microphone and the other hand gesturing as he speaks.
frames/frame_001230.jpg	Eddie Murphy is standing with his left hand holding a microphone, looking serious and engaged as he performs.
frames/frame_003450.jpg	Eddie Murphy is standing with one hand holding a microphone and the other hand resting by his side, looking directly a...

file_path (image)

Create a project

Set up your fine-tuning project

Upload data

Add your training images

Caption Images

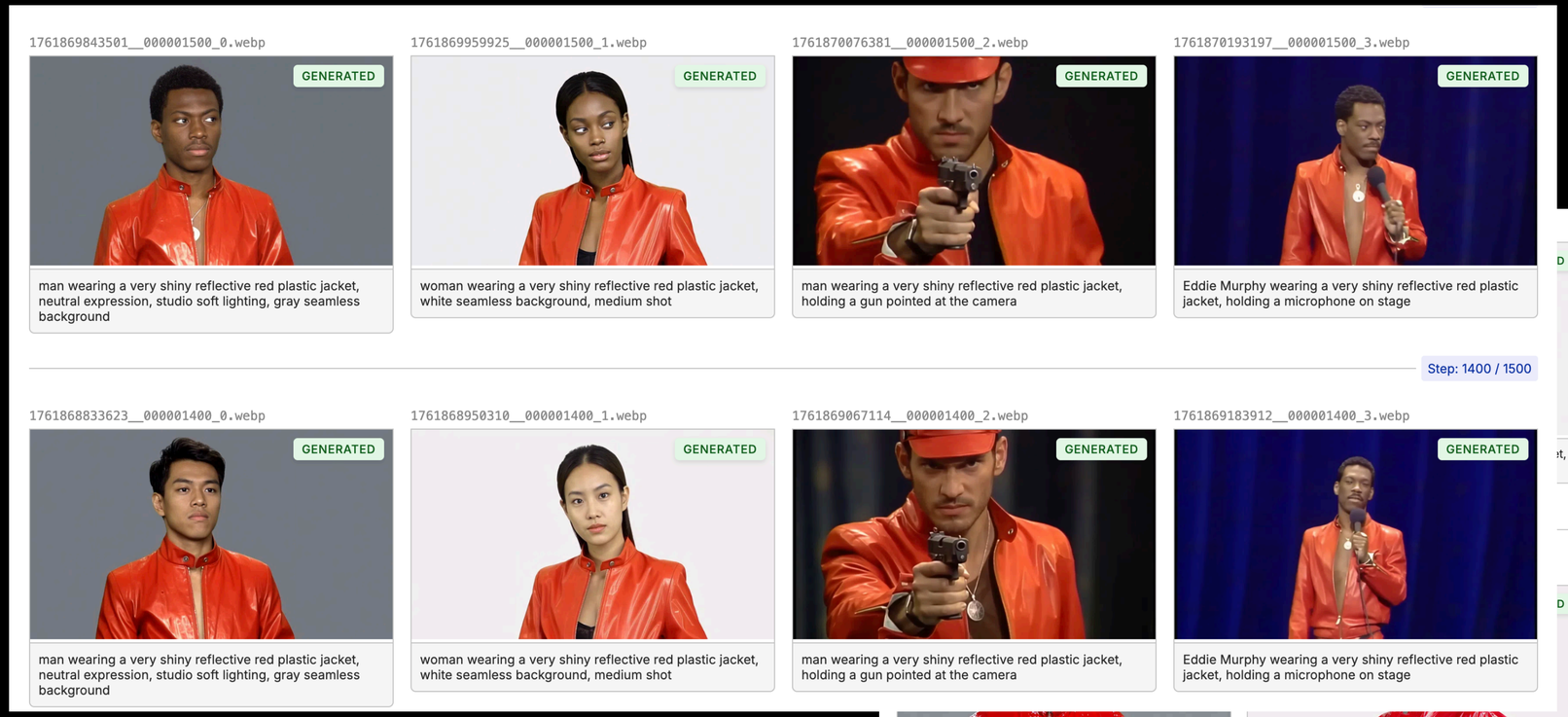
Label and annotate your data

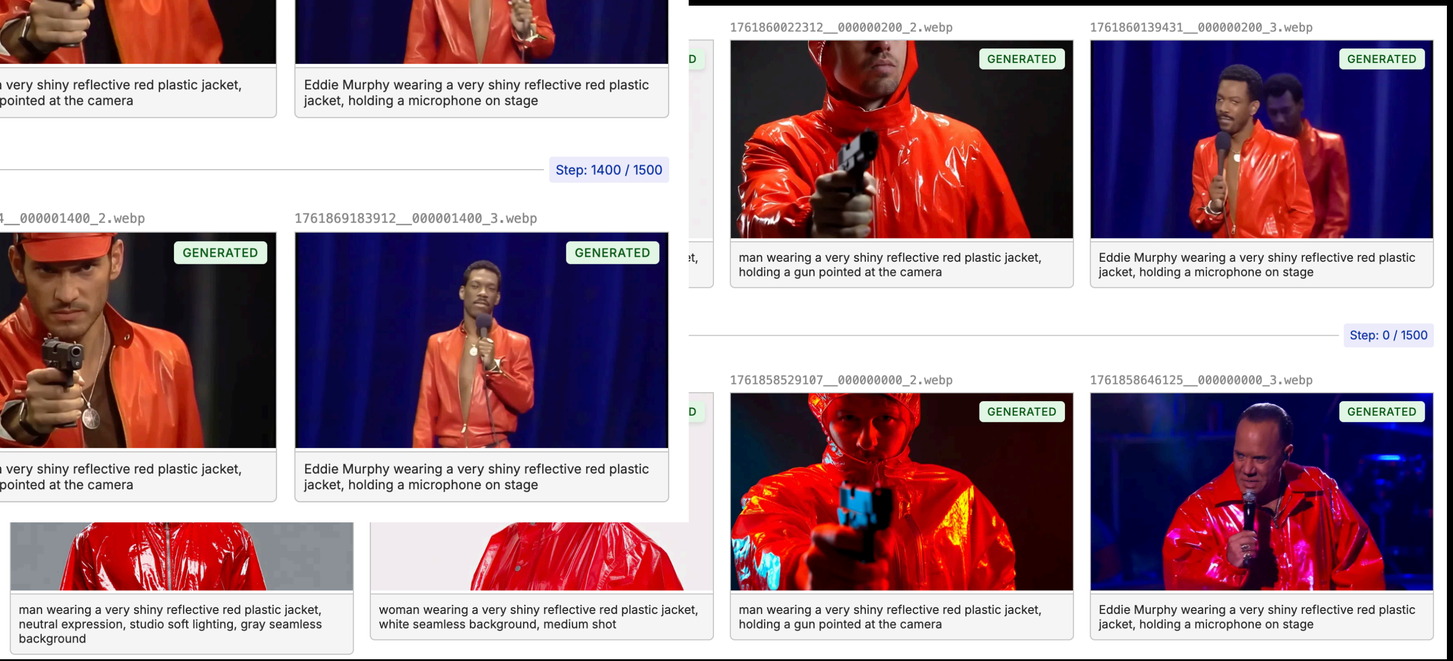
Kick off a job

Choose your hyperparameters and kick off a job

Let’s look at a training run

Debug w/ Samples

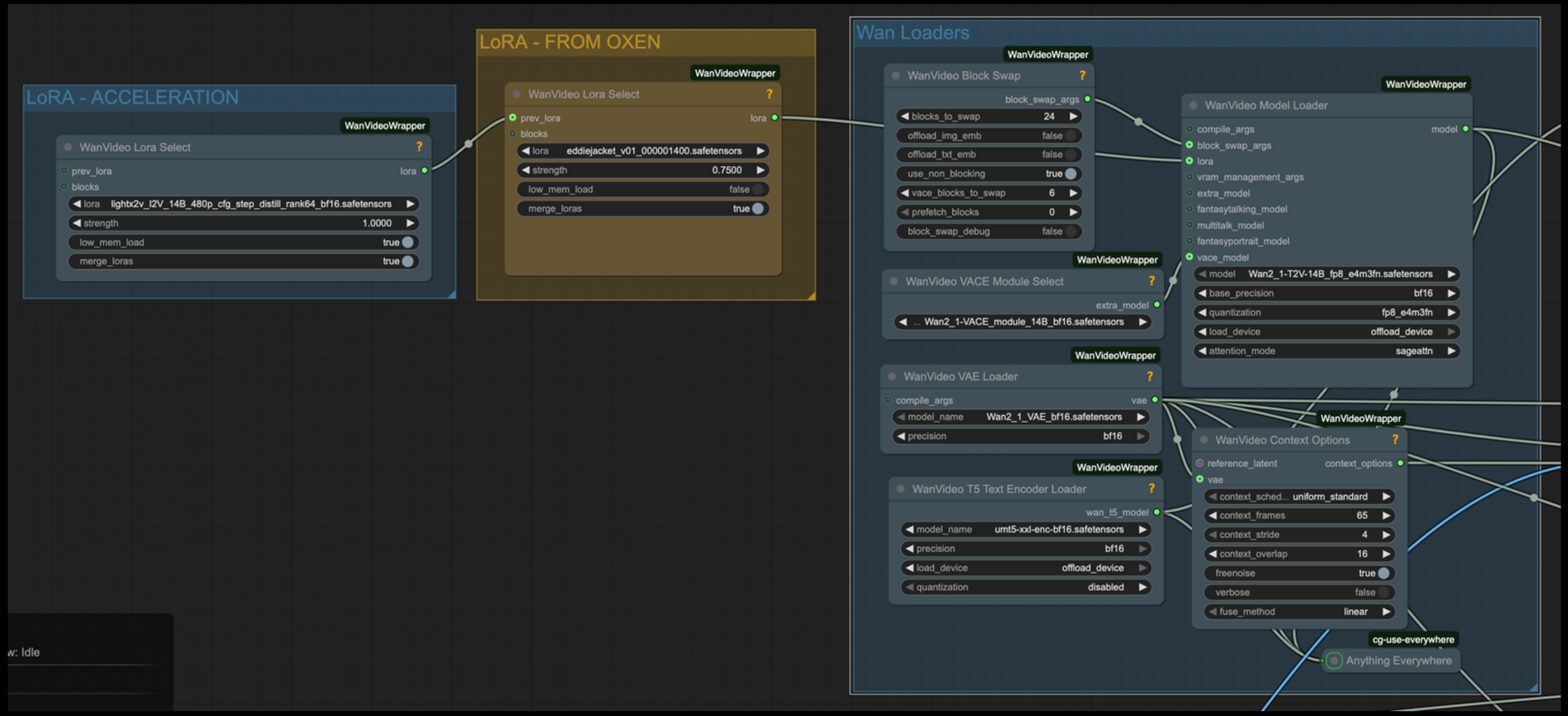




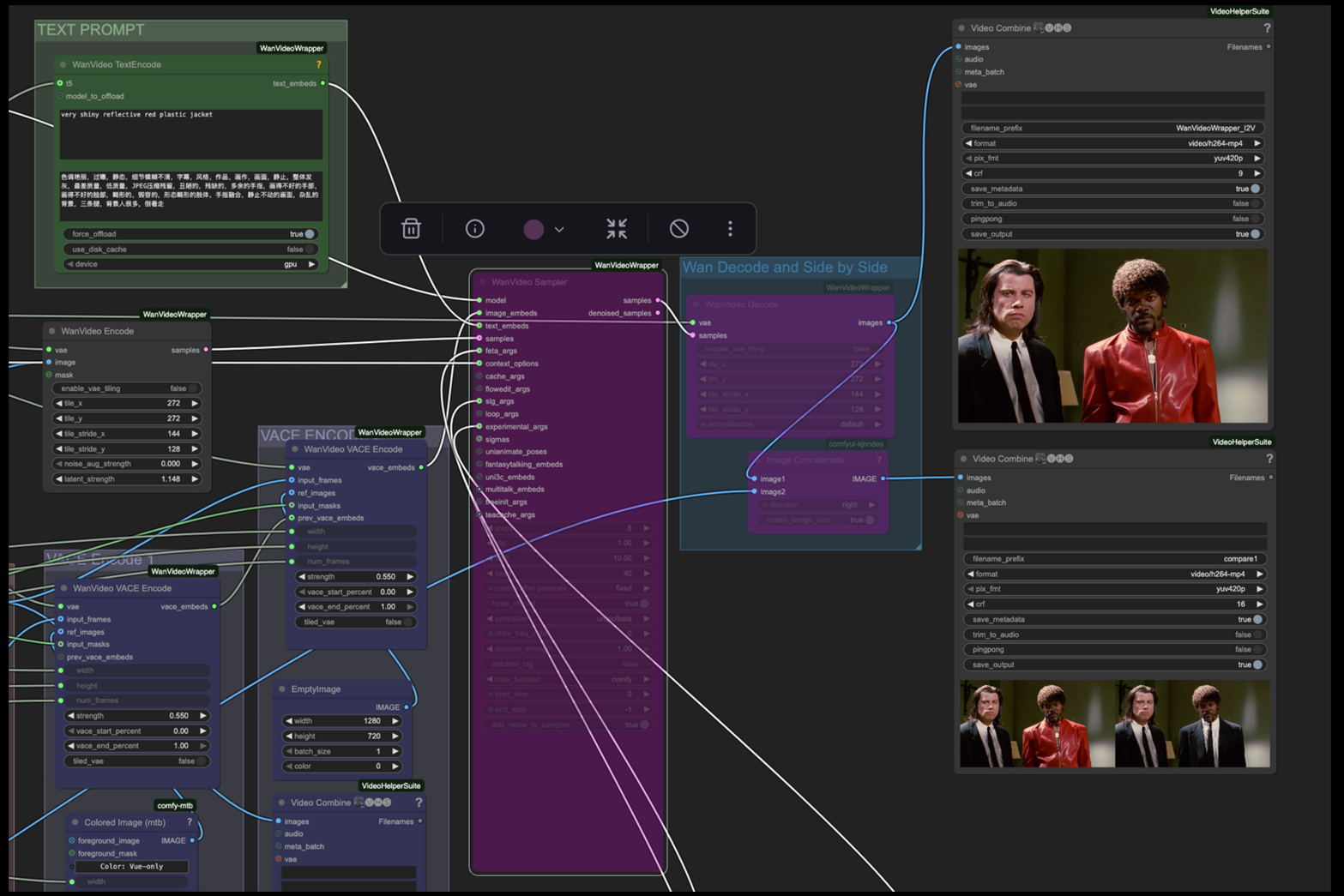
https://www.oxen.ai/ox/EddieMurphyDelirious/fine-tunes/3bb96d9a-aa7a-4c48-9f4c-5cc9be5fe580?activeTab=samples

Step 3b: Jacket Specific LoRA

Added to ComfyUI



Step 4: Sample & Generate



Stumbling through the latent space

🍺 Like a drunk actor

Stumbling through the latent space

🎲 Try a different random seed?

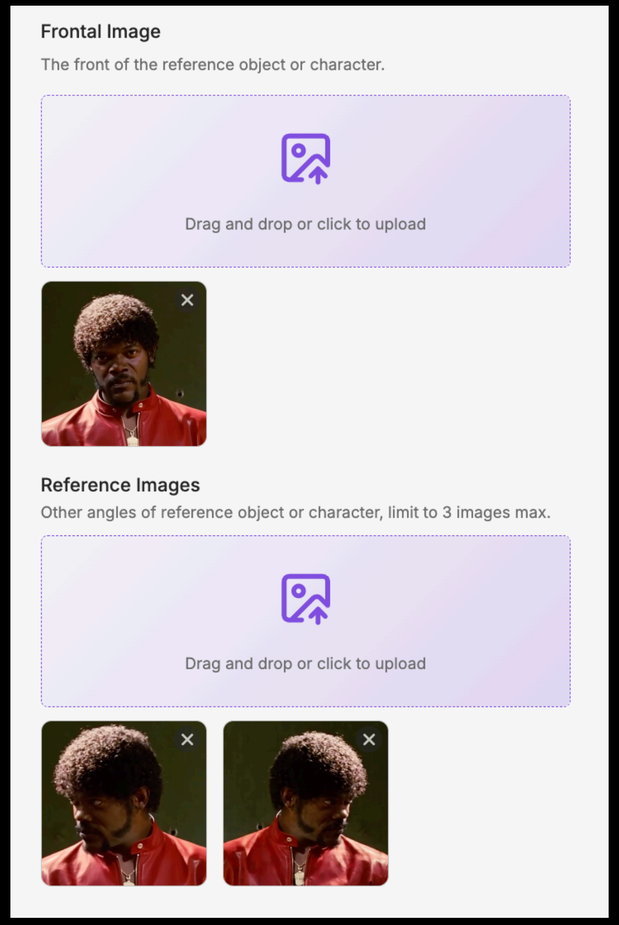
Stumbling through the latent space

🎬 Take 3

Best Closed Source…

RI2V: Reference Image to Video

Kling O3 Reference to Video (Not Open Source…)



RV2V: Reference Video to Video

Seedance 2.0

https://x.com/RuairiRobinson/status/2021394940757209134

Flipping the Narrative

~~**“I created this 30 second commercial in 4 hours with a few simple prompts!”**~~

- Real Production Pipeline / Schedules
- 8 Weeks
  - **Discovery (week 1,2)**: Brand research, story board, data collection, script writing
  - **Key Moments (week 3,4)**: Stills to animation, training LoRAs
  - **Fill in gaps (week 5,6)** - Send to editor, more animation, VFX cleanup
  - **Finish (week 7,8)** - Upres, Final Cleanup, Color Correction, VFX, Sound Design, Picture Lock

Conclusion

AI + VFX Requires Work

- Never a 1-click replacement
- Many tools, many workflows
- Artists rolling dice
- LoRAs/Reference images help with consistency
- Many fix ups in VFX (fine-details, 4K, masking, etc)
- High quality renders still takes a long time (5-10 minutes)
- Artists, actors, directors are still key!
  - Higher quality, shorter timelines, not replacement

Want to Collaborate?

🐂 Join the oxen.ai herd

- We are working with…
  - Brands
    - Franks Red Hot
    - Bell
  - Agencies
    - Car Commercials
  - Hollywood Directors
    - Household names 🤫

DM me on LinkedIn for credits 💰