Episode Book

EPISODE #6617

Why CVAT Beats Generalists at Video Annotation

PUBLISHED	15 November 2026 at 20:20 UTC
DURATION	21:55
LISTEN	myweirdprompts.com/episode/cvat-video-annotation-deep-dive-67857/
HOSTS	Corn, Herman, Larry, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	omnivoice
PIPELINE	V5.33
SOURCES	25 cited by the research (listed at the end)
BOOK GENERATED	15 November 2026 at 23:00 UTC

SYNOPSIS

CVAT has fewer GitHub stars than Label Studio, yet it remains the reference tool for video annotation. Originating inside Intel in 2017, its narrow focus on computer vision delivers keyframe interpolation, Outside-Merge-Split track primitives, and 3D point-cloud support that generalists lack. This deep dive examines how CVAT's architecture handles model-assisted labelling via Nuclio and CVAT AI Agents, why Safari isn't supported, and how its recent $2 million investment and metered SAM 2 features are reshaping the economics of human-in-the-loop annotation.

TRANSCRIPT



Corn

Why does the narrower tool keep winning? That's the question underneath Daniel's prompt this week. He's asking for a proper deep dive on CVAT, the Computer Vision Annotation Tool, and specifically what it does that a generalist platform like Label Studio doesn't. Tracking objects across frames, interpolation, AI-assisted labelling, what happens at team scale, and where you'd honestly reach for something else.



Herman

And the framing I'd add is that this tool started as an internal Intel utility in twenty seventeen, built on top of an older open-source video tool called VATIC, and it's now the answer that comes back whenever anyone in computer vision asks how to label a video. Seventeen thousand stars on GitHub, more than three million downloads, MIT licence.



Corn

Label Studio has twenty-six thousand stars, for the record. More popularity, less specialisation.



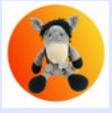
Herman

Which is the whole tension. Two tools, both permissive open source, both free to self-host, and the one with fewer stars is the reference tool in its lane. The interesting question isn't which has more features. It's why a tool that only does images and video beats a generalist at video.



Corn

Start with where it came from. You said Intel, VATIC.



Herman

Right, so twenty seventeen, internal Intel tool, built so Intel's own computer vision engineers didn't have to hand-label training data with a stick in the mud. It sat on top of VATIC, which was a research-grade video annotation tool out of Irvine, and CVAT essentially industrialised it. Intel open-sourced it in June twenty eighteen, and then in July twenty twenty-two the core team spun it out as CVAT.ai Corporation, alongside OpenCV.AI. Two founders, Nikita Manovich and Boris Sekachev. And just a couple of



Herman (continued)

weeks ago, October twenty-seventh, a venture firm called OH.io, founded by Ratmir Timashev, who co-founded Veeam, put two million dollars in and set up a Columbus-based go-to-market team targeting manufacturing QA, robotics, geospatial, precision agriculture.



Corn

Two million dollars to go sell an annotation tool to tractor companies.



Herman

That's the part I find fascinating. The CEO's own quote in the announcement is that CVAT grew into one of the most widely used tools of its kind largely through word of mouth, and the partnership gives it, quote, the sales and marketing muscle to reach companies that don't yet know how our technology can help them. Which is a candid admission that the thing scaled on reputation alone.



Corn

Hold that thought for the end, because whether a word-of-mouth tool survives being sold is exactly the open question. Current version?



Herman

v2.83.0, released November twelfth. Page-scoped state retention, some quality control niceties, a security fix. It's an actively maintained repo, five hundred and ninety-something open issues and pull requests, last push today.



Corn

Fine. So the framing for anyone who's used both: CVAT does images, video, PDFs, point clouds, and now audio intervals. Label Studio does all of that plus text, audio transcription, time series, mixed modality. CVAT's own comparison says it's specialised focus, quote, aligning perfectly with Computer Vision projects, versus Label Studio's versatility. That's a competitor writing the comparison, but it's also just true.



Herman

It is true. And what specialisation buys you is the thing I want to spend the first half of this on, because it's the mechanism people underrate. Keyframe interpolation.



Corn

Explain it like I'm staring at a ten-thousand-frame video.



Herman

You draw a box around an object in Track mode. Let's use CVAT's own documentation example, because it's concrete: a cyclist appears at frame two thousand two hundred seventy. You draw the box on that frame, then you jump forward thirty frames, to two thousand three hundred, and adjust the box to where the cyclist now is. The twenty-eight frames in between are generated for you. Linear interpolation on the box's position and size between the keyframes.



Corn

Twenty-eight frames I never touch.



Herman

Twenty-eight frames you never touch. You mark a keyframe with the K key or the star button, and the docs are explicit that it isn't necessary to change the rectangle on each frame. Update a handful of keyframes, everything between them is automatic.



Corn

For a sloth, this is not a feature. This is a constitutional right.



Herman

And the honest version of the speedup claim, because you'll see a number floating around, ninety percent reduction in manual video labelling time. That traces to a third-party comparison site, Inference Systems, not to anything CVAT has published. No first-party benchmark exists.



Corn

So ninety percent is a marketing number wearing a lab coat.



Herman

Directionally, sure. For simple tracking with a box that doesn't change shape much, interpolation saves most of the work. The number itself, treat with salt. What's not in dispute is the mechanism.



Corn

The mechanism I believe. What I want to know about is the edges. Objects don't cooperate. The cyclist goes behind a van.



Herman

That's where the track lifecycle primitives come in, and this is the vocabulary generalist tools don't have. Three of them. Outside, on the O key, ends a track when an object leaves the frame or gets too small to label. Merge, M key, stitches two tracks back together when the same object vanishes and reappears, which is the behind-the-van case exactly. And Split breaks one track into two, for when one object becomes two or you realise you've been tracking two things as one.



Corn

Plus navigation.



Herman

Buttons to jump to the previous keyframe, the next keyframe, the initial frame, the last keyframe. Small stuff, but when you're doing this eight hours a day, that's the difference between a tool designed by people who've done the job and a tool designed by people who've read about the job.



Corn

That's the whole Outside-Merge-Split cluster though. Objects in real video are continuously appearing, disappearing, occluding.



Herman

Continuously. And here's the shape vocabulary on top: bounding boxes, rotated boxes, ellipses, two-D cuboids, polygons, polylines, pixel masks with brush tools, keypoints and skeletons for pose, tags, and three-D cuboid detection and tracking through point-cloud sequences with perspective and projection views. That last one is not a toy. Point-cloud annotation is how autonomous vehicle and robotics datasets get built.



Corn

And the export side?



Herman

Over twenty import and export formats. COCO, YOLO, Ultralytics YOLO, Pascal VOC, KITTI, MOT, MOTS, Datumaro, Cityscapes, Open Images, on down to things like Wider Face and Market-1501. And the detail I'd flag is MOT and MOTS specifically, because those are the multiple object tracking benchmark formats. The tool isn't just built for the labelling step, it's built for the evaluation pipeline. You can pull your annotations out in the exact format the tracking benchmark expects.



Corn

So the frame budget goes to making output that downstream measurement understands.



Herman

Yes, and that's no accident, because the whole thing came out of a place that was building models. Labelling tools that come out of research groups have that shape.



Corn

One caveat I want on record, because it's the kind of thing no comparison matrix ever mentions. The README says CVAT is primarily tested on Chromium-based browsers. Safari, WebKit, not supported.



Herman

Not supported, in the README, in those words. Firefox may work with caveats. And for a Mac-heavy team, that's a real IT conversation. You're either pushing Chrome to every annotator or you're having a very dull meeting about browser policy before you've labelled a single frame.



Corn

A tool so specialised it has opinions about your browser. I respect the commitment.



Herman

Specialisation cuts both ways.

L

Larry

Friends, are your object tracks drifting out of frame, vanishing, and reappearing like they never loved you? Then ask your doctor about Ceevatol, the once-daily annotation remedy that interpolates your entire life. Ceevatol draws a box around the problem on Monday, nudges it on Wednesday, and quietly fills in every awkward day between, so you never have to touch a single frame again.

L

Larry

Ceevatol is not for everyone. Side effects may include sudden keyframes behind the wheel, involuntary merging with strangers at the supermarket, splitting into two separate people during meetings, and a persistent feeling that your outside state has been mislabelled. Some patients report three million downloads of their personal memories, a rotating cuboid where their left ear used to be, and a polite but firm message from a serverless function. Ask your doctor if Ceevatol is right for you, and remember, if you experience a track that will not close, stop annotating and seek perspective view immediately.



Corn

How much is a month of Ceevatol.



Herman

You don't have a bicycle, a video of a bicycle, or a dataset. You nap through the frames you're supposed to be checking anyway.



Corn

The doctor said seek perspective view. I'm seeking it horizontally.



Herman

Moving on. Part two, because interpolation gets you the geometry, but the next question Daniel's asking is what model-assisted labelling looks like. There are three routes in CVAT. The oldest is Nuclio serverless functions: you package your model script into a container, expose it over HTTP, Nuclio manages the function, and the CVAT server calls it with the right arguments. Once it's deployed it's available to every user.



Corn

The architecture behind that?



Herman

Django backend, PostgreSQL, Redis as the job queue, an Open Policy Agent service for authorisation, and the auto-annotation Nuclio microservice sitting alongside. It's a proper little distributed system. The second route is third-party services, Hugging Face, Roboflow. The third is the newest one, CVAT AI Agents, from version 2.25 onward, and that's the one I find clever.



Corn

The pitch there is keep your weights.



Herman

The pitch is the answer to a question every team with a proprietary model asks, which is how do I use my own model without handing my weights to a vendor. The agent is a process running on your hardware or your infrastructure. It doesn't require CVAT administrator control, it needs only HTTPS connectivity out, requests are queued first-come-first-serve, and if an agent crashes the request gets reassigned. It's your GPU, your model, their orchestration.



Corn

I'd bet on that as the model. Nobody who spent two years training a detector hands it to an annotation vendor.



Herman

Available on CVAT Online and Enterprise, note, not Community. In Community you get the Nuclio route, and the built-in models are real: Segment Anything as an interactor, Inside-Outside Guidance, RetinaNet as a detector, HRNet for pose. But here's where the paid line sits, and it's the most strategically interesting fact in the whole tool. SAM 2 and SAM 3 segmentation, batch auto-annotation across an entire task, the advanced quality control UI. Paid tier.



Corn

Say that plainly. The famous feature is free and its modern replacement costs money.



Herman

The interpolation that made CVAT's name is geometric and cheap to serve, so it lives on the free tier. SAM 2 does video segmentation semantically, prompt it once and it propagates the mask across frames, and that's metered. One hundred calls a month on the free plan, ten thousand on Solo, ten thousand times paid seats on Team.



Corn

The moat becomes the upsell.



Herman

Which tells you the moat was never the interpolation itself, it was that annotators learned the workflow. Now the question is whether the workflow survives the transition from boxes you adjust by hand to masks a model suggests and you approve. I don't know. The hybrid's been the argument for a while now, agent first pass plus human finishing, and we've made that argument on this show, but the tool's economics now assume it.



Corn

Team scale. The bit I actually think about is what happens with twenty annotators and a deadline.



Herman

Project, task, job. A project holds a dataset and the shared label configuration. A task is part of the dataset. Jobs are the unit you actually assign to an annotator. Split a ten-thousand-frame video into twenty jobs, twenty annotators, no collisions, because nobody's editing the same frames.



Corn

That answers the collaboration critique people make.



Herman

The critique is that CVAT's collaboration is moderate, task-based, versus Label Studio's project-based real-time collaboration with user roles. And I think the critique is aimed at the wrong thing. If you're running outsourced labelling at scale, real-time co-editing is not what you want. You want isolation, and you want agreement measurement.



Corn

Because the failure isn't two annotators fighting over one box.



Herman

The failure is twenty annotators quietly producing inconsistent boxes, and nobody noticing until the model trained on them is bad, and nobody able to say which annotator. That's the actual risk at scale. And CVAT's answer is a quality control stack that I'd call the real differentiator over Label Studio, more than the tracking.



Corn

Ground Truth jobs.



Herman

Ground Truth jobs and Honeypots, reference annotations and validation frames planted to measure individual annotator accuracy. Reviewer assignment with issues, comments, approve-reject. Consensus-based annotation, multiple annotators on the same data. Immediate Job Feedback, which shows an annotator their quality result the moment they finish a job, with limited validation attempts, so they can't just retry until they pass.



Corn

That last bit's a small design decision with teeth.



Herman

It is. And the recent releases doubled down. Version 2.81 added replica jobs, N copies of the same job for consensus, creatable via API or a bulk action on selected jobs. Version 2.82 added a variable IoU threshold depending on the ground truth annotation area, so small objects get judged fairly instead of failing a fixed overlap bar, plus a Boundary F1 metric for how precisely edges were drawn, and a choice between OKS and IoU for object similarity.



Corn

That's not a checkbox marked quality control.



Herman

No, it's a bet. Replica jobs are the betting slip. QA beats simultaneity. For the outsourced-annotation world, I think that bet is right, and the reviewers who ding CVAT for moderate collaboration are reviewing it as a design tool.



Corn

So where's it the obvious choice? Say it as a decision.



Herman

Long-form video tracking, full stop, that's the core competency. Multi-sensor computer vision, including the point clouds and audio. On-prem requirements, the README says no data leaves your environment, and for medical or defence data that's the whole procurement. A team with an in-house model to wire in as a pre-labeller via Nuclio or Agents. Outsourced labelling operations that need the Honeypot machinery. Anyone who needs MOT or KITTI or Cityscapes exports. The University of Illinois Chicago urology department is using it to train AI to recognise anatomy in real time during single-port robotic surgery, which is about as unforgiving a use case as exists.



Corn

And the other list.



Herman

Text and NLP at scale, audio transcription, time series, mixed-modality datasets, that's Label Studio's home ground, its breadth is the point. Safari-only shops. Anyone who wants a fully managed no-Docker experience. Very small projects, because the free cloud tier limits bite immediately: two members, one project, three tasks, one gigabyte. And teams who need batch auto-annotation or SAM 2 at volume on the free tier, because those are paid.



Corn

Three tasks. That's not a tier, that's a trial balloon.



Herman

It's an evaluation licence with a website around it. The pricing above it is flat rate, which CVAT makes a point of contrasting with Label Studio's contact-sales enterprise model.



Herman

And here's the thing about naming, Herman. People keep saying interpolation when they mean the whole keyframe discipline.



Hilbert

That word again. Interpolation.



Corn

It's the standard term.



Hilbert

It's the wrong word. You're not filling in values between measurements, the way we did on charts. Corn, you said the box adjusts itself between two thousand two hundred seventy and two thousand three hundred. Nothing adjusted. The machine drew twenty-eight guesses and everyone signed them. A guess printed in red is still a guess. The frame where the cyclist's elbow is behind the van, nobody drew that. It was agreed.



Corn

Agreed by the software and countersigned by inertia.



Hilbert

Cough on track two, I'll cut it. Where was I. I spent a month on this kind of work once, decades back, rotoscoping for an inspection outfit, and the man who owned the rostrum camera, Ismail, he would not sign an in-between frame. He'd sign the ones he touched and initial the log for those only. The client sent the reels back twice before they understood he wasn't being difficult. He was being accurate about which frames were his.



Corn

Herman, this is your... he's arguing the keyframe is the only honest signature.



Herman

No, and that's the part worth pushing on, because that's exactly what the Ground Truth machinery is for. The variable IoU threshold is CVAT conceding the point in software. Small object, the overlap bar scales down, because they know a fixed bar on a small box is a lie about accuracy.



Hilbert

I never said the tool was wrong, Herman. I said the word covers it. The technician who wires a rig like that properly is worth more than the whole software budget. It's patient work and it doesn't photograph well. Anyway, your levels on the last ten minutes are two points hot on the left channel.



Corn

The in-between frames aren't measurements. That's going to sit with me.



Herman

The misconception to kill before we wrap up, because Daniel will hear it from every comparison site: that more GitHub stars means the better tool. Label Studio, twenty-six thousand seven hundred stars. CVAT, seventeen thousand one hundred. And yet CVAT is the reference tool for tracking.



Corn

Stars measure how many people need any annotation at all. The tracking claim measures how good it is at one thing. Both true at once.



Herman

The question I'd leave you with is the one we parked earlier. A tool that grew for eight years by word of mouth just took two million dollars from a venture firm with a sales team in Columbus, at the exact moment its signature feature got old enough to be the free thing and semantic segmentation became the upsell. The open question isn't whether OH.io can sell CVAT. It's whether the tool that annotators loved on merit can stay the reference tool once it's being sold to people who don't yet know they need it. Ismail knew which frames were his. A sales team doesn't.



Corn

Producer credit to Hilbert Flumingtop, who keeps our levels honest even when he disputes our vocabulary.



Herman

If you want more of this, try episode twenty, Architectural AI. This has been My Weird Prompts. If today's episode earned its keyframes, the review that would mean the most is a short one wherever you listen.



Corn

We'll be back soon. See you then.



Herman

See you.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **GitHub, cvat-ai/cvat, retrieved 2026-11-15, last push 2026-11-15, latest release v2.83.0 (2026-11-12). [primary]**
   https://github.com/cvat-ai/cvat

2. **CVAT Docs, Track mode, retrieved 2026-11-15, page updated 2026-11-13. [primary]**
   https://docs.cvat.ai/docs/annotation/manual-annotation/modes/track-mode-basics/

3. **CVAT Docs, Manual Annotation, retrieved 2026-11-15, updated 2026-11-13. [primary]**
   https://docs.cvat.ai/docs/annotation/manual-annotation/

4. **CVAT Docs, Automated Annotation, retrieved 2026-11-15, updated 2026-11-13. [primary]**
https://docs.cvat.ai/docs/annotation/auto-annotation/

5. **CVAT, CVAT Online Plans, retrieved 2026-11-15. [primary]**
https://www.cvat.ai/pricing/cvat-online

6. **CVAT, About CVAT, retrieved 2026-11-15. [primary]**
https://www.cvat.ai/about

7. **CVAT Blog, CVAT Celebrates 14K Stars on GitHub and Its Three-Year Anniversary!, published 2025-07-24. [primary]**
https://www.cvat.ai/resources/blog/cvat-anniversary

8. **CVAT Blog, CVAT.ai Birthday is Here, published 2024-08-01. [primary]**
https://www.cvat.ai/resources/blog/cvat-celebrates-seven-year-anniversary

9. **HumanSignal, Label Studio overview, retrieved 2026-11-15. [primary]**
https://docs.humansignal.com/guide/get_started

10. **GitHub, Label Studio LICENSE (Apache 2.0), retrieved 2026-11-15. [primary]**
https://github.com/HumanSignal/label-studio/blob/develop/LICENSE

11. **CVAT Docs, Serverless tutorial, retrieved 2026-11-15. [primary]**
https://docs.cvat.ai/docs/guides/serverless-tutorial/

12. **CVAT Docs, AI models, retrieved 2026-11-15. [primary]**
https://docs.cvat.ai/docs/annotation/auto-annotation/ai-models/

13. **CVAT Changelog, Announcing CVAT AI Agents, retrieved 2026-11-15. [primary]**
https://www.cvat.ai/resources/changelog/announcing-cvat-ai-agents

14. **CVAT Docs, CVAT Architecture, retrieved 2026-11-15. [primary]**
https://docs.cvat.ai/docs/administration/community/advanced/cvat-architecture/

15. **CVAT Docs, Auto-annotation API, retrieved 2026-11-15. [primary]**
https://docs.cvat.ai/docs/api_sdk/sdk/auto-annotation/

16. **PRWeb, CVAT, Industry-Leading Open Data Annotation Platform, Joins OpenCV, published 2022-07-26.**
https://www.prweb.com/releases/cvat-industry-leading-open-data-annotation-platform-joins-opencv-866430452.html

17. **Business Wire, OH.io Partners with CVAT..., published 2026-10-27.**
https://www.businesswire.com/news/home/20261027744548/en/OH.io-Partners-with-CVAT-to-Accelerate-Growth-of-the-Open-Source-Data-Labeling-Platform-to-Train-AI-at-Scale

18. **CVAT Blog, CVAT vs. Label Studio: Which One to Choose?, retrieved 2026-11-15. [primary, but competitor-authored]**
https://www.cvat.ai/resources/blog/cvat-or-label-studio-which-one-to-choose

19. **Inference Systems, CVAT vs Label Studio: Open-Source Video Annotation Compared, retrieved 2026-11-15.**
https://inferensys.com/differences/long-form-video-intelligence-platforms/video-annotation-and-labeling-platforms/cvat-vs-label-studio

20. **Potato, Video Annotation Tools Compared: CVAT, Label Studio, ELAN, BORIS and Potato, retrieved 2026-11-15.**
https://www.potatoannotator.com/docs/guides/video-annotation-tools-compared

21. **Qwanteos, Comparing CVAT and Label Studio, published 2025-06-17.**
https://www.qwanteos.com/comparing-cvat-and-label-studio

22. **Roboflow, CVAT vs. Label Studio, retrieved 2026-11-15.**
https://roboflow.com/compare-labeling-tools/cvat-vs-label-studio

23. **HumanSignal, Compare Label Studio Editions, retrieved 2026-11-15. [primary for Label Studio]**
https://docs.humansignal.com/guide/label_studio_compare

24. **Label Studio, Label Studio 1.21.0, published 2025-09-30.**
https://labelstud.io/blog/label-studio-1-21-0/

25. **unsubbed.co, Label Studio, retrieved 2026-11-15 (26,752 stars, Apache-2.0).**
https://unsubbed.co/tools/label-studio/