Episode Book

EPISODE #6614

How Humans Actually Grade Generated Video

PUBLISHED	15 November 2026 at 18:21 UTC
DURATION	26:07
LISTEN	myweirdprompts.com/episode/evaluating-generated-video-interfaces-95246/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	omnivoice
PIPELINE	V5.32
SOURCES	24 cited by the research (listed at the end)
BOOK GENERATED	15 November 2026 at 23:00 UTC

SYNOPSIS

Computer vision platforms like CVAT, Encord, and Label Studio offer keyframe interpolation, SAM 2 object tracking, and frame-level timelines for real footage. Yet nobody has built a widely-adopted annotation environment for grading generated video. Instead, AI video evaluation relies on thin arena voting pages like Artificial Analysis's Video Arena with its five-second watch gate, rubric experiments from iMerit and Claru, and an open-source NeurIPS protocol with just five stars.

TRANSCRIPT



Corn

We can generate video now. Sixty seconds of a robot riding a horse through Jerusalem, no camera anywhere. And the state of the art for grading it is a webpage with two buttons.



Herman

Prefer left, prefer right. That's the peak of the instrument.



Corn

Which is roughly what Daniel noticed. He's been thinking about image-to-video models specifically, and what he wants to know is: if you're going to have humans rate or label the clips those models produce, what do you actually run that on? Are there annotation environments built for video in the first place? Are people using dedicated video labelling tools, general platforms that happen to have a video mode, or interfaces they rolled themselves? And he's got a specific itch about the interface layer, the primitives, scrubbing, frame-level timelines, tracking an object across frames, the stuff an image labeller never needs. Then the last piece: how is human evaluation of generated video actually being done today?



Herman

And the answer to the last piece is where it gets almost embarrassing, because there are two completely separate worlds here. One of them is a mature, crowded commercial market with keyframe interpolation and object tracking and event spans, and the other is a video player with a watch-time gate. They barely reference each other.



Corn

So we should probably say the negative finding out loud early, so people know what they're listening for.



Herman

Yes. Nobody we could find has built a widely-adopted annotation environment specifically for grading generated video. Not one. What exists falls into four buckets, computer-vision platforms built for real footage, thin arena voting pages, research protocols nobody reuses, and crowdsourced pairwise data collection. The closest thing to a purpose-built



Herman (continued)

image-to-video evaluation harness is a camera-movement benchmark, and even that is pairwise voting, not annotation.



Corn

Two buttons with a very large sample size.



Herman

So the shape of this is: first the world that built the machinery, then the world that didn't, then the gap.



Corn

Start with the machinery. Who's in it?



Herman

The names you'd expect. CVAT, Encord, Label Studio, V7, SuperAnnotate, Voxel51, Labelbox, Supervisely. And the first thing to know about this market is that its map is drawn by its residents. Every roundup of the best video annotation tools in twenty twenty-six is published by a vendor, and the vendor ranks itself first or near-first. Encord's list of eight tools puts Encord in it. Unitlab's top seven puts Unitlab in it. Labellerr's twelve includes Labellerr.



Corn

The list of best lists is itself a list by a listed party.



Herman

It's the whole industry in miniature. Nobody disinterested has ever counted.



Corn

Which tells you the customers don't shop by reading reviews. Who are the customers?



Herman

Autonomous driving, surveillance, detection and segmentation training data. That's the framing across every single vendor's marketing. And that matters for later, because none of them, not one page we looked at, mentions grading generated clips. This is a market with a hundred-thousand-frame problem, and it solved it properly.



Corn

Give me the number.



Herman

One hour of video footage can run past a hundred thousand frames, each one needing temporally consistent labels. You cannot hand-label that. A human annotator marking boxes at even two frames a second would take fourteen hours per hour of footage, and the labels would drift. So the entire UI vocabulary exists to let a human specify a sparse set of decisions and delegate the rest.



Corn

That's the actual design constraint. You're not annotating a video, you're annotating maybe one frame in fifty and trusting interpolation.



Herman

And the primitives fall out of that. Keyframe interpolation is the big one: you draw the box on frame one, you redraw it on frame thirty, and the tool fills everything between. CVAT has that natively, and Encord, V7, SuperAnnotate, Unitlab all advertise it. Then persistent object IDs, so the pedestrian you labeled at frame one is still pedestrian seventeen four hundred frames later, same identity. CVAT is the one people single out for tracking object IDs across long sequences, which is exactly why it dominates driving and surveillance datasets.



Corn

And the tracking got weird recently, didn't it.



Herman

It got good. SAM 2 changed the economics. Voxel51's pitch is almost poetry, they say: draw once, and SAM 2 locks onto your object's visual features and tracks it across frames automatically, adjusting for movement, scale, and lighting. And then the honest clause, when the model loses its lock due to occlusion or blur, the human takes over again. That's the whole division of labour in one product description. The model does the ninety-nine percent, the annotator handles the frame where the cyclist passes behind a van.



Corn

The tool is a slot machine that pays out until it doesn't.



Herman

Beyond that you've got timeline navigation with frame-by-frame scrubbing, multi-view annotation, temporal context, Encord's interface is built around granular control over video sequences. Timestamped event spans, so you're not just boxing objects, you're marking that a thing happened from frame four-ten to frame six-fifty. And dynamic per-object properties, state that changes over time, a light that turns red, a person who sits down.



Corn

So scrubbing, frame-level timelines, cross-frame tracking, event spans, per-object state. Daniel's list of primitives basically matches the CV world's feature sheet one for one.



Herman

His list reads like a CVAT spec sheet, which is a good sign he was looking at the right things. And the third bucket in his question, general platforms with a video mode, is a bigger part of the answer than people assume. Label Studio is the clean example. It's open source, it started as a multi-modal labeller, text, audio, images, and it grew a serious video mode. There's a control tag called TimelineLabels, merged in September twenty twenty-four, and the description is exactly Daniel's primitive list: you select a label, click once to annotate a single frame, click and drag to annotate a span of frames. The values land in a ranges field. There's a dedicated player with frame-by-frame navigation underneath it.



Corn

And that's the same repo that runs annotation for half the ML teams on the planet.



Herman

The same repo. And they went further, there's a TimelineLabels ML backend that combines YOLO features with an LSTM for temporal multi-label classification, and you need on the order of ten to twenty annotated videos of roughly five hundred frames each to get it into trainable mode. So the loop closes: the timeline spans humans draw become training data for a model that pre-fills the next timeline spans.



Corn

The annotator trains her own replacement during coffee breaks.



Herman

In a limited domain, but yes, that's literally the architecture. The pricing tells you who all this is for, by the way. CVAT is free for one or two people, then twenty-three dollars a month on annual billing, thirty-three monthly, and if you want to self-host the enterprise edition it starts at twelve thousand dollars a year. V7's starter tier runs around nine thousand a year for fifty thousand files and three seats. Encord won't quote you at all, it's enterprise sales, with SOC 2 and HIPAA compliance badges on the page. Nobody buying a twenty-three dollar solo plan is grading video model outputs in their bedroom.



Corn

Twelve thousand a year to self-host. The word free in that pricing table is doing Olympic-level work.



Herman

And Unitlab's practical advice is the tell about who actually uses any of this: pilot on one hundred to five hundred frames before committing. That's fleet-scale procurement thinking. And that gets us to the other world, because none of this machinery is pointed at generated video.



Herman

So, generated video. Four distinct modes of human evaluation visible out there, and the first is the one everybody knows: arena voting.



Corn

The two buttons.



Herman

Artificial Analysis runs a Video Arena. You submit a prompt, you get two anonymous videos side by side, you vote. And the interface details are where the methodology hides. There's a gate, watch for five more seconds to vote, the vote button literally doesn't unlock until you've watched. Keyboard shortcuts for prefer left, prefer right, play, pause, restart, and a key that swaps in different models. Separate tabs for text-to-video, image-to-video and video editing, and a toggle for with audio or no audio, with the instruction, sound on, you're judging the sound too, use headphones in a quiet space.



Herman

It is, because it's an admission. The whole UI has fewer controls than a YouTube embed. No frame stepping, no timeline, no marking where things went wrong. It's a preference instrument, not a measurement instrument.



Corn

And the five-second gate. I keep turning that over.



Herman

It's an evaluation-methodology decision disguised as a UI detail. It's an anti-skimming gate, and the fact that it's needed tells you what the default rater behaviour is, which is voting on the first frame. Compare that with the other end of the spectrum, trained annotators with rubrics and inter-annotator agreement. iMerit argues against a single overall quality score and for dimension-specific rubrics with real anchors: temporal consistency, one is severe instability, three is noticeable artifacts, five is stable throughout. Prompt adherence, one is substantially incorrect, five is fully executed. Motion naturalness, same shape.



Corn

Two theories of the human. One says the human is a coin that needs to be watched for five seconds so it doesn't cheat, the other says the human is an instrument you calibrate.



Herman

Claru did the cleanest version of that calibration I've seen. They ran a pilot with five trained annotators testing six candidate metrics, and they kept only the ones where agreement, Krippendorff's alpha, cleared point seven. Two metrics survived, overall quality and prompt adherence. The other four, temporal consistency, visual fidelity, motion naturalness, aesthetic appeal, all got dropped, because trained humans couldn't agree on them.



Corn

So the dimensions Daniel most wants marked, temporal consistency, motion naturalness, are the ones that don't survive contact with a rubric.



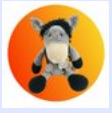
Herman

In that study, at that sample, yes. Then the main evaluation was an Elo tournament of about six thousand videos from a hundred and twenty prompts. Which is the interesting part, the trained rubric didn't replace pairwise voting, it fed it.



Corn

The rubric picks what the coin flips on.



Herman

Third mode, academic protocols. There's a NeurIPS twenty twenty-four paper that is the most directly on-topic artifact in existence, it's literally titled Rethinking Human Evaluation Protocol for Text-to-Video Models. The authors surveyed the field and found that few papers employ consistent protocols, some use Likert scales, some comparative, many don't document their methods at all, and most rely on laboratory-recruited annotators with no quality checks. So they built a proper protocol: defined metrics, annotator training, a dynamic evaluation module, an annotation interface where annotators choose the superior video per metric. They claim to cut evaluation costs nearly in half, and they tested it on two thousand video pairs across Gen2, Pika, Latte, VideoCrafter.



Corn

And they open-sourced it, I assume, since that's the whole point of a protocol.



Herman

They did. The repo exists, and it has five stars and zero forks.



Corn

Five stars.



Herman

Five stars, and the last commit was this past September. The single most directly reusable annotation interface for this exact problem, published at the biggest ML venue there is, and the entire field collectively shrugged. That's the reuse rate on shared evaluation UI.



Corn

We did an episode once about benchmarking ASR beyond word error rate, and the exact same thing showed up, everyone agrees the metric is broken, nobody adopts the fix. It's apparently the natural state of evaluation tooling.



Herman

Fourth mode, crowdsourced pairwise at industrial scale. Rapidata published a dataset of nine hundred thousand human responses, nine hundred thousand, collected through their SDK, comparing fifteen image-to-video and world models on camera movement. Same still, same instruction, two models' clips head to head. And the task design is clever: the instructions are scene-agnostic, tilt thirty degrees down, truck left, so the judgment is purely whether the camera obeyed.



Corn

Truck left.



Herman

It's the industry term, moving the camera laterally. But notice what even this, the closest thing to a purpose-built image-to-video evaluation harness, actually is. It's pairwise comparison. Not frame-level. Nine hundred thousand humans clicking left or right.



Corn

Nine hundred thousand responses and not one of them says where the hand melted.



Herman

Not one. And that's the real finding of the whole episode, I think. Every failure mode people complain about in generated video is frame-localized. A hand that's fine at second two and a puddle by second four. An object that teleports. A face that changes identity mid-shot. All of these have an exact frame address. And the CV world solved frame-addressed annotation years ago, timeline spans, event markers, click and drag over the frames where the thing happened. The generated-video world has borrowed none of it. Everything is clip-level: prefer this clip, or score this clip one to five per dimension.



Corn

Label Studio's timeline tag could mark the melt span this afternoon.



Herman

It could. The primitive exists, open source, merged, documented. Nobody has wired it to generated video grading at any scale anybody's heard of.



Corn

Which is strange, because the leagues can't agree with each other, and that's supposed to be the thing pairwise voting solves.



Herman

It's worse than disagreeing. Artificial Analysis has Wan 3.0 at number one on text-to-video, Elo eleven seventy-five, Gemini Omni nowhere near the top three. Arena.ai has Gemini Omni 1.2 Flash at number one at fifteen forty, with Wan 3.0 down at eighth. Same models, opposite verdicts, different prompts, different voter populations, different Elo anchors, Artificial Analysis pins Kling 3.0 at a thousand and builds everything off that.



Corn

If human preference is the ground truth, the ground truth has a different answer depending on which website you ask.



Herman

And the vote-count asymmetry is the part that should embarrass somebody. Arena.ai's number one model sits on seventeen hundred votes and is flagged preliminary. The number three has forty-six thousand. Saying Gemini Omni 1.2 Flash is the best video model is quoting a number with a plus-minus interval built on seventeen hundred clicks.



Corn

Seventeen hundred people choosing a lunch spot would still be arguing.



Herman

Meanwhile the benchmark people are racing to replace the humans entirely. MSVBench claims ninety-four percent Spearman correlation with human judgments, against fifty-eight for VBench, and its own framing is blunt, benchmarks cannot yet replace the costly and non-scalable practice of human evaluation in multi-shot video generation. Video-Bench went the other direction and recruited ten annotators for thirty-five thousand evaluations under shared human and model guidelines. The small academic studies just do the default: MechVerse had six evaluators rating forty-three outputs on twelve dimensions, one to five. Agreement around point seven.



Hilbert

Hilbert's Anteater Trivia. Enlarged salivary glands coat the tongue in a sticky saliva. I have glands too, apparently, and nobody has ever asked them for a rating.



Corn

Not even a preliminary one.



Herman

So the benchmark numbers are the thing to watch, because the moment a benchmark sustains ninety-plus correlation, the arenas stop being load-bearing. And there's a quieter consequence: the annotation environment becomes the training data. MSVBench fine-tuned a small vision model on human judgments until it matched commercial ones. The annotators are teaching the thing that replaces them.



Corn

Which is the Label Studio YOLO loop again, one rung up the ladder.



Herman

Same shape, bigger ladder.



Corn

Herman, can I register the thing that still bothers me about the arena side? The five-second gate. The defense of it is that it forces engagement. But five seconds of a ten-second clip is half the evidence, and five seconds of a minute-long clip is a trailer.



Herman

That's fair, and I don't actually know what the number is doing. My guess is they measured voting time distributions and picked a threshold that catches the fastest cheaters without annoying everyone else. It's a blunt instrument for an honest problem.



Corn

An honest problem with a turnstile.



Herman

A turnstile. And you can see the whole tension by asking what the rater is trusted with. The arena trusts the rater with a binary preference and nothing else. The rubric world trusts the rater with five-point judgments on named dimensions, then audits the rater's agreement with four other raters. The CV world trusts the annotator with frame-level authorship, mark exactly where, and hands them a timeline to do it on. Trust goes up, UI complexity goes up, and the generated-video world sits at the bottom of that staircase calling it scale.



Corn

Because the staircase is expensive to climb.



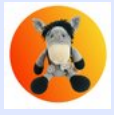
Herman

Because it is, and because pairwise scales for free. Nine hundred thousand responses costs Rapidata a lot less than nine hundred thousand frame-level annotations would. Frame-level work needs trained people, agreement checks, the whole apparatus. That's the whole story of why the tool doesn't exist, I think. It's not that nobody thought of it. It's that the cheap instrument answers the question the leaderboard wants answered, which model wins, and nobody has made the business case for the expensive instrument that answers the question the model builders actually need, which is where, specifically, it breaks.



Corn

Which is a strange economy. We spend twelve dollars a minute generating the video, on models that cost hundreds of millions to train, and then grade the output with the cheapest possible measurement of it.



Herman

The measurement is always the afterthought. And Corn, honest question, since you're the one who keeps arguing verification is the real frontier, do you actually buy that the frame-level tool gets built, or does the benchmark crowd get there first?



Corn

I think the benchmark crowd gets there first, and I think that's a mistake, but it's also probably right. A model grader will happily mark the melt span. The problem is you then have no human ground truth to check the model grader against, because nobody ever built the frame-level human dataset. They're automating a measurement that was never taken.



Herman

Which is exactly the trap the T2VHE authors were pointing at, in the adjacent room. Everyone automating, nobody instrumenting.



Corn

The one protocol that tried to instrument it has five stars.



Herman

Five stars.



Corn

I want to come back to something, though, the Rapidata camera benchmark, because it's the one place I'd actually push back on the pessimism. Scene-agnostic instructions, tilt thirty degrees down, truck left. That's a controlled variable in a field that otherwise has none. If I were building the frame-level tool, I'd start there, a constrained instruction where the correct answer is checkable, and mark the frames where compliance starts and stops. It's the one seed of real annotation in the whole landscape.



Herman

And it's sitting in a pairwise dataset as a byproduct. Nobody designed it as a seed, but it is one. A span-labelled version of camera-movement compliance is maybe the most tractable first dataset in generated video, because the instruction is unambiguous.



Hilbert

The gate isn't new, for what it's worth.



Corn

Which gate?



Hilbert

The watch-five-seconds thing on the arena. Video comparison sites did that years ago. There was a tape house on Jaffa Road, the man there, Ilan, ran A against B for clients on two monitors side by side, and he would not take an opinion from anyone until they'd seen both all the way through. Standing rule. He said a client who watched the whole thing argued less afterwards. It was never about skimming, Herman, whatever your anti-skimming theory is. It's about the voter feeling they've done some work before they get a say. Ilan charged by the half hour either way, and his clients were calmer, so it worked.



Herman

So the gate is a customer-satisfaction feature wearing a methodology badge.



Hilbert

It always was. And for the record, when you two were going on about nobody doing frame work by hand, I've done it. I built a frame stepper out of a spreadsheet once, to settle an argument about a render. Each row a frame, each cell labelled by hand, column for what changed. Got to around frame four hundred and the lot started losing rows to the recalculation, so I split it into sheets of two hundred and chained them.



Corn

You built a video annotation timeline in a spreadsheet.



Hilbert

It worked. The scrollbar was slow but the columns never lied. My brother-in-law has the file now, he uses it for his stock footage, and he gets better results with it than the software he pays for, though don't tell him I said that, he'll want me to maintain it. I did a similar grid for the council over a parking dispute, but that was mostly to irritate them.



Herman

Frame four hundred in a spreadsheet. That's a hundred and fifty annotator-hours a minute of video, which is honestly the cleanest demonstration of why the CV tools exist that anyone's given all episode.



Hilbert

Well, it was either that or concede the argument, and I wasn't going to do that. The label I used for the bad frames was just the word no.



Corn

The entire system, one label.



Hilbert

Two columns from frame four hundred onward, no and check again. Most of them were no.



Corn

So where does that leave us? With two annotation worlds pointed at different problems, and the one Daniel actually needs sitting unbuilt in the gap between them. The CV tools have the timeline primitives, the generated-video side has the models and the voters, and the thing that connects them, frame-level quality marks on generated clips, exists as a merged pull request nobody has aimed at the right problem.



Herman

The ground-truth question is the one I'd leave people with. If human preference is the ground truth, why can't the two biggest arenas agree on who's winning? And if the benchmarks replace the humans, what are the benchmarks being validated against, a measurement nobody ever took properly? Hilbert's spreadsheet with the no column might be the most honest instrument in the whole field. Producer credit where it's due, that's Hilbert Flumingtop at the desk, as ever. If you want more of this, try episode sixteen, On Deepfakes, SynthID, And AI Watermarking. This has been My Weird Prompts. If today's rabbit hole earned its depth, leave us a review, it's how the show finds its way to stranger questions. We'll be back soon.



Corn

Until then, mark the frame.

SOURCES

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

1. **Voxel51, video annotation product page (SAM 2 tracking), retrieved 2026-11-15. [primary]**
   https://voxel51.com/annotation/video

2. **Encord, video annotation product page, retrieved 2026-11-15. [primary]**
   https://encord.com/video/

3. **Label Studio docs, TimelineLabels tag, retrieved 2026-11-15. [primary]**
   https://labelstud.io/tags/timelinelabels

4. **HumanSignal/label-studio repo, retrieved 2026-11-15. [primary]**
   https://github.com/HumanSignal/label-studio/

5. **PR #6191 Add TimelineLabels to mark parts of the video, created 2024-08-13, merged 2024-09-09, retrieved 2026-11-15. [primary]**
   https://github.com/HumanSignal/label-studio/pull/6191

6. **Label Studio docs, TimelineLabels ML backend (YOLO+LSTM), retrieved 2026-11-15. [primary]**
   https://labelstud.io/guide/ml_tutorials/yolo_timeline_labels

7. **Label Studio docs, video object detection/tracking template, retrieved 2026-11-15. [primary]**
   https://labelstud.io/templates/video_object_detector

8. **Zhang et al., Rethinking Human Evaluation Protocol for Text-to-Video Models (T2VHE), NeurIPS 2024, retrieved 2026-11-15. [primary]**
   https://proceedings.neurips.cc/paper_files/paper/2024/file/94a1fa3346407b2318ae7f8891d9f550-Paper-Conference.pdf

9. **T2VHE code repo, last commit 2026-09-11, retrieved 2026-11-15. [primary]**
   https://github.com/ztlmememe/T2VHE

10. **Artificial Analysis Video Arena, retrieved 2026-11-15. [primary]**
https://artificialanalysis.ai/video/arena?tab=leaderboard-text

11. **Artificial Analysis, AA-Video-T2V v2.1 leaderboard, retrieved 2026-11-15. [primary]**
https://artificialanalysis.ai/video/leaderboard/text-to-video

12. **Artificial Analysis, open-weights video leaderboard, retrieved 2026-11-15. [primary]**
https://artificialanalysis.ai/video/leaderboard/text-to-video/open-weights

13. **Arena.ai Text-to-Video Arena, data as of 2026-11-08, retrieved 2026-11-15. [primary]**
https://arena.ai/leaderboard/video/text-to-video

14. **MSVBench: Towards Human-Level Evaluation of Multi-Shot Video Generation, retrieved 2026-11-15. [primary]**
https://arxiv.org/html/2602.23969

15. **Han et al., Video-Bench, CVPR 2025, retrieved 2026-11-15. [primary]**
https://openaccess.thecvf.com/content/CVPR2025/papers/Han_Video-Bench_Human-Aligned_Video_Generation_Benchmark_CVPR_2025_paper.pdf

16. **MechVerse: Evaluating Physical Motion Consistency in Video Generation Models, retrieved 2026-11-15. [primary]**
https://arxiv.org/html/2605.14843v1

17. **Rapidata, image-to-video relative camera movement human feedback benchmark, retrieved 2026-11-15. [primary]**
https://huggingface.co/datasets/Rapidata/image-to-video-relative-camera-movement-human-feedback-benchmark

18. **NVIDIA Cosmos Lab, PhyWorldBench, retrieved 2026-11-15. [primary]**
https://research.nvidia.com/labs/cosmos-lab/phyworldbench/

19. **A Survey of AI-Generated Video Evaluation, retrieved 2026-11-15. [primary]**
https://arxiv.org/abs/2410.19884

20. **Encord, 8 Best Video Annotation Tools 2026, retrieved 2026-11-15.**
https://encord.com/blog/best-video-annotation-tools/

21. **Unitlab AI, Top 7 Video Annotation Tools & Platforms for 2026, retrieved 2026-11-15.**
https://blog.unitlab.ai/top-7-video-annotation-tools/

22. **Labellerr, 12 Best Video Annotation & Labeling Tools In 2026, retrieved 2026-11-15.**
https://www.labellerr.com/blog/top-video-annotation-tools/

23. **iMerit, Video Generation Evaluation: Beyond FID and CLIP, retrieved 2026-11-15.**
https://imerit.ai/resources/blog/beyond-fid-and-clip-what-production-grade-video-evaluation-actually-requires/

24. **Claru, video generation model evaluation case study, retrieved 2026-11-15.**
https://claru.ai/case-studies/video-model-evaluation