Episode Book

EPISODE #6611

Image Annotation When Experts Are the Bottleneck

PUBLISHED	15 November 2026 at 18:19 UTC
DURATION	21:56
LISTEN	myweirdprompts.com/episode/image-annotation-domain-experts-33439/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	omnivoice
PIPELINE	V5.32
SOURCES	17 cited by the research (listed at the end)
BOOK GENERATED	15 November 2026 at 23:00 UTC

SYNOPSIS

Annotating architectural styles in photos is far harder than audio or text because a label can hinge on a 40-pixel carved fleuron, and styles overlap across centuries. Drawing on the WikiChurches paper, the discussion explores why domain experts should curate prototypical examples—narrowing 117 classes to four—and pair bounding boxes with free-text descriptions. It also compares CVAT, Label Studio, Labelme, Supervisely, and Roboflow Annotate, highlighting the UI features that matter most: keyboard-first zoom and category-first browsing.

TRANSCRIPT



Corn

Daniel has spent months annotating audio and text, and now Hannah wants in on the training-data game. She's an architect, and the project is a model that classifies architectural styles. His question: what annotation environments do people actually use for detailed image work, cloud and self-hostable, and what UI features matter when a domain expert is labelling thousands of images. The part I like is that last bit. When the annotator is the expert, the tool isn't just catching keystrokes. It's handling the scarce resource.



Herman

And it's a different job from what he's been doing. I keep coming back to the audio case, because he knows it. You place a boundary in time on a one-dimensional signal, and once you've heard the thing, there's usually no argument about where it starts and stops. Text is the same, just highlighted instead of cut. Architecture gives you none of that comfort.



Corn

None.



Herman

There's a paper called WikiChurches, from NeurIPS, which is literally this task. Nine and a half thousand images of churches, expert bounding boxes, architectural style classification. They say determining a building's style is hard even for experts, because the categorization depends on a combination of subtle and sometimes tiny visual features, and they name quatrefoils and fleurons. A fleuron is a carved flower at the top of a finial. It can be forty pixels wide in a photo of an entire facade.



Corn

Forty pixels. The whole label decision can hang on something the size of a thumbnail on a five-megapixel image.



Herman

That's the first difference. The second is that the labels aren't mutually exclusive. Roskilde Cathedral in Denmark is Brick Gothic in the catalogue, but parts of it were rebuilt across centuries, so sections read as Renaissance, Neoclassicism, Byzantine Revival, even



Herman (continued)

Modern. The paper concludes that style recognition is actually a multi-label problem. And the third difference: the labels are hierarchical, and people are inconsistent about depth. Thousands of buildings labelled simply Gothic, only eighty-one labelled at the finest level of the hierarchy, because most people stop at the coarse label.



Corn

So in audio, the tool needs a waveform and a pair of buttons. Here, the tool has to support zooming into a multi-megapixel image, judgement calls about whether a feature is diagnostic, and a decision about which level of a taxonomy to commit to, with no objective answer. That's what Hannah's signing up for.



Herman

Which is why the choice of environment actually matters here more than it did for his audio work. Let's do the landscape first, then the differences, then the interface features, then the workflow.



Corn

Go.



Herman

Self-hostable side, four names keep coming up. CVAT is the reference computer-vision annotator. MIT licence, seventeen thousand stars, runs with a docker compose up, supports boxes, polygons, masks, keypoints, even 3D point clouds. Twenty-plus export formats, COCO, YOLO, Pascal VOC, so the model side is never a problem. Python SDK, REST API.



Corn

The catches.



Herman

Two, and they're specific. The good interactors, click a point and get a mask from Segment Anything, are in the free version, but the newer SAM two and three auto-labelling and the AI agents are paid Enterprise only. And it officially supports Chromium browsers only, Safari is not supported. So check your browser before you set Hannah up.



Corn

I like that the free version's price is a browser.



Herman

Label Studio is the multimodal one, Apache licence, twenty-eight thousand stars. It does audio, text, images, video, time series. This matters for Daniel specifically, because one self-hosted instance could hold his audio spans and Hannah's facade boxes side by side. Installs with pip or docker.



Corn

You have a but face.



Herman

I have a big but. The September release turned security defaults on, and one of them blocks ML backends on localhost and private networks. Which is exactly where a self-hoster's backend runs. You have to know the flag exists and deliberately disable it. Otherwise you spend an afternoon wondering why your model-assisted labelling is silent.



Corn

A security default whose practical effect is to break the thing you installed the software for. Herman, you must have hit that one personally.



Herman

I hit it within an hour, yes. Clean install, backend up, no suggestions arriving. I assumed it was my firewall, my ports, my karma. It was a default.



Corn

Labelme is the third one, right? The desktop one.



Herman

Labelme, yes. Python and Qt, GPL, a lightweight desktop application rather than a server. Polygons, rectangles, points, plus something I find interesting. It does point-to-polygon with SAM, so you click inside a window tracery and get the outline. And it does text-to-annotation with newer models, so you can literally type what you want outlined.



Corn

Type quatrefoil, receive geometry.



Herman

Roughly. And in this month's release it added AI-Find All, which finds every object matching a text prompt or one example box across a batch, and you refine it with counter-examples while holding Shift. Draw one rose window, get every rose window proposed, then draw over the mistakes. For an architect who knows exactly what she's looking for, that could collapse a week of work into an afternoon of reviewing.



Corn

Then it's doing the finding and she's doing the confirming.



Herman

Which is the right division of labour for the scarce resource. Fourth, Supervisely, enterprise-grade, Apache SDK, handles images, video, 3D, medical imaging. Its interesting feature is live training sessions, you start one mid-project, pick your classes, seed it with the annotations you've already made, and the model starts suggesting while



Herman (continued)

you label. The model improves during the labelling instead of after it.



Corn

So those are the environments. The self-hosted ones are free, they run on his hardware, and each has a landmine with his name on it.



Herman

Cloud side, briefly. Roboflow Annotate is the strongest hosted end-to-end option, but it's cloud-only, no true self-hosting on the free or pro tiers, and on the free Public plan your data is visible in Roboflow Universe. Hannah's photographs of client buildings, in a public gallery, in exchange for the free tier.



Corn

That's the actual price. Not the zero dollars.



Herman

Right. Hosted CVAT exists, third-party listings put it around thirty-three dollars per user per month, Enterprise from twelve thousand a year, but those are roundups, not the vendor, so treat them as approximate. Label Studio paid starts around ninety-nine a month, Supervisely Pro around a hundred and ninety-nine euros. The real cost of self-hosting, though, isn't the licence, it's the ownership. Someone has to handle upgrades, access control, backups, performance, and the integrations that move data into training. That's Daniel, or it's nobody.



Corn

He'll do it. He'll enjoy it. The landmine question is whether Hannah enjoys it.



Herman

She won't touch it, and she shouldn't have to.



Corn

Okay, that's the landscape. Now the differences, because I want to push on this. The interesting claim in that paper isn't that images are harder than audio. It's where the difficulty sits.



Herman

It sits in three places. Tiny features, which we covered. Non-exclusive labels. And uneven label precision. The last one is sneaky, because it's not a tooling problem at all, it's a data problem wearing a tooling costume. If your classes come from a catalogue like Wikidata, you inherit every inconsistency in the catalogue. The paper found about six and a half percent label noise, and thirty-one of the sixty-four level-one styles had exactly one image each.



Corn

One image. A class with a population of one.



Herman

One church, one style label, one photo. That's not a class, that's an anecdote. And this is where it connects to a failure mode I've read about in other domains, a historical-document field test that found deep learning accuracy with more than a thousand one-hot output classes drops to nearly zero, while an older bag-of-visual-words approach held around eighty-seven percent.



Corn

Wait. The paper's own dataset has a hundred and seventeen style labels. If Hannah builds a full taxonomy, she walks straight into that cliff.



Herman

She does if she tries to use the whole taxonomy. The WikiChurches team didn't. For the expert annotation stage they deliberately narrowed to four well-studied styles, Romanesque, Gothic, Renaissance, Baroque, which together cover about eighty-nine percent of the images anyway. They had the courage to throw away a hundred and thirteen classes.



Corn

So the expert wasn't annotating the dataset. The expert was annotating a curated subset of it.



Herman

And here's the number that should set expectations for Hannah. All of that expert knowledge, an art historian, given free rein, produced six hundred and thirty-one bounding boxes across a hundred and thirty-nine churches. Not labels for nine and a half thousand images. Six hundred and thirty-one boxes. Weeks of expert time, deliberately spent on the prototypical cases only. The protocol literally asked for churches that were, quote, pure in style and prototypical for that period.



Corn

Skip everything that argues with you.



Herman

Ambiguous buildings were skipped by design. That's not laziness, it's curation.



Corn

Interesting. My instinct would have been to push the expert through the hard cases, because that's where the model is weakest. But this says the opposite, spend the expert where the signal is cleanest.



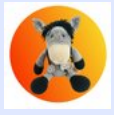
Herman

There's a second reading too, which I like even more. The six hundred and thirty-one boxes each carry a free-text description of the feature inside, written by the art historian, and boxes could be grouped when several elements form one structure. So the expert's actual reasoning, what she noticed and why it mattered, is sitting in the data. There's a recent paper on Emirati residential architecture that leans all the way into that, they embedded the expert's textual descriptions alongside the images with CLIP and got ninety-eight percent accuracy across eight style clusters. The words aren't decoration. They're a training input.



Corn

Which answers a question Daniel didn't ask. The expert's descriptions aren't metadata around the labels. They might be the labels.



Herman

That's the transition I wanted. If the expert is the bottleneck, and she is, she's the only person in the building who can tell a fleuron from a finial, then the interface features that support expert judgement aren't nice-to-haves. They're the whole game. So let's do those.



Corn

Before we do, we have our producer's segment.



Hilbert

Hilbert's Mid-Show Haiku. The Badge Is Forty Pixels Wide. One box drawn at four hundred percent. The expert squints. The label waits.



Herman

The badge is forty pixels wide. So was the fleuron.



Corn

So the interface features. Zoom first, because everything in that protocol depended on it.



Herman

The WikiChurches expert drew boxes in a popup window, by dragging the mouse over a larger version of the image. A bespoke interface built for one art historian, and the enlargement is treated as part of the annotation design, not a viewer convenience. Any tool Hannah uses has to make zooming a one-gesture, keyboard-first thing, not a menu.



Corn

And my claw situation says something about that, incidentally. Pinch-to-zoom on a trackpad was designed for fingers. Give me keys and a wheel or give me death.



Herman

Second feature, category-first browsing. The interface let the expert browse all images per category, including sub-categories. Read that again. The taxonomy is the navigation. She moves through the collection by style, the thing she knows, not through a flat folder of nine thousand JPEGs, the thing the file system knows.



Corn

That's backwards from how every annotation tool defaults. They're built around the image queue, because the annotator is assumed to be interchangeable. The moment the annotator is the expert, the queue is the wrong shape.



Herman

Third, the AI-assisted drawing. We touched it with Labelme, but it belongs in the list. SAM as an interactor, click a point, get a mask, is in CVAT's free tier too. The expert confirms rather than traces. Fourth, example-based finding, the AI-Find All pattern, find every rose window from one example, corrected with counter-examples. Fifth, quality control that understands geometry, and this is more interesting than it sounds. CVAT's October release added a variable overlap threshold depending on the area of the ground-truth box.



Corn

Because a forty-pixel fleuron and a whole facade can't be judged by the same tolerance.



Herman

A one-pixel miss on a fleuron is a catastrophic error in relative terms, and invisible on a facade. A single fixed threshold would call one of them perfect and the other wrong. Sixth, consensus. CVAT added replica jobs, multiple annotators on the same job. Which matters because of a finding from medical annotation I keep quoting, six radiologists segmenting CT scans came out with a kappa of about zero point six nine, and the study's conclusion was that the inter-reader variation is itself a challenge for the definition of ground truth. The disagreement among experts isn't noise in the data. It's data.



Corn

Though Hannah is a practice of one. Who's her second radiologist?



Herman

She has Daniel, and it's useful even if he's wrong in a consistent direction, because then you learn which judgements are his. But if she can't get a second expert, the honest answer is her ground truth is an opinion with excellent credentials.



Corn

Seventh, seventh. Model in the loop.



Herman

Seventh, model in the loop, the Supervisely live-training pattern, seed it with her first hundred boxes and let it suggest. And there's evidence people behave well under this. That CT study found users who leaned on the suggestions, prioritizing learning speed over model accuracy in their words, gave the tool higher usability scores. Nobody is honoured for tracing every outline by hand.



Corn

Which is a cousin of the skip lesson. Let the human spend herself only where she's needed.



Herman

Eighth, and this is the one I'd put money on. Interaction count. There was a survey of fifty-nine video annotation tools, and its finding was blunt, most have overwhelming interfaces, poor interaction techniques, and difficult-to-understand features, leading to longer annotation time, label inconsistencies, and user fatigue. The survey's own redesign cut required interactions by thirty-six percent against five popular tools. Thirty-six percent, just from fewer clicks per decision.



Corn

So the misconception is pick the tool with the most features. The truth is closer to pick the tool with the fewest clicks.



Herman

For an expert, yes. A crowdsourced annotator needs guardrails and checkboxes. Hannah needs speed, zoom, and a skip button.



Hilbert

The cast-iron storefronts. That was the winter of eighty-five, a small downtown, and I photographed every building for the historical society and marked the ironwork on paper prints, in pencil, because that's what they'd asked for and the pencil came out of my own pocket.



Corn

On paper prints.



Hilbert

Photographic prints, Herman, letter size, and I drew on them with a chinagraph pencil, which doesn't smudge when the print sweats in a cold building. The marking was the easy part. You'd think the catalogue was the buildings we kept. It wasn't. The work was the shoebox. Every building I decided was not interesting enough, I printed anyway and put it in a separate box, labelled REJECTED in the same chinagraph, and when I finished the winter I had more in that box than in the catalogue.



Herman

How many more?



Hilbert

I didn't count them then and I won't now. By ninety-two, four of the rejected ones had been demolished, and the shoebox was the only photograph anyone had. I still have it. Corn, I bring one photograph from it to family gatherings and ask if anyone remembers the building, and mostly nobody does, and I put it back.



Corn

You're running a small mercy mission for buildings nobody fought for.



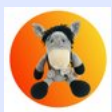
Hilbert

I'm running nothing. I put it back in the box.



Corn

He's right, though, that the rejects are a record. Half the judgement in annotation is what you refuse to label.



Herman

That's the part the tools don't store. The skip is a click that goes nowhere. Imagine if every skipped image carried the reason, and that's, no, I'm just going to say it plainly. The skip log is the expert's negative space, and no environment I know of keeps it.



Corn

So, the workflow end to end, because Daniel asked what it actually looks like.



Herman

WikiChurches is the honest template, and it's less glamorous than people expect. Collect, and their collection was a query against Wikidata, ten thousand churches. Filter, they cut about three percent interiors with a pretrained classifier and then checked it by hand. Resize, smaller side to a maximum of about twelve hundred and eighty pixels. Then the expert stage, and that's the six hundred and thirty-one boxes we talked about, on a hundred and thirty-nine hand-picked churches, four styles, free-text on every box.



Corn

Then the model, which by your historical-document caution can't be a hundred and seventeen classes.



Herman

It can't. The bench isn't huge either. There's a recent benchmark of about seventeen hundred architectural images with a vision-language framework that reports ninety-two percent consistency with expert annotations, and the CLIP approach with the eight style clusters at ninety-eight. Small, well-defined class sets, expert text as a signal, part-level boxes for the small diagnostic features. That's the whole recipe.



Corn

We did an episode once on architectural AI with ControlNet and ComfyUI, on the generation side of that world, and I remember the same asymmetry. Everyone wants to talk about the model. Nobody wants to talk about the curation.



Herman

And curation is where Hannah's actual skill lives. So here's the one thing. If you take one thing from this episode, let it be this: when the annotator is the expert, the problem isn't how to get more labels, it's how to spend her attention. Four styles, not a hundred and seventeen. Prototypical buildings only. Free-text on every box. Six hundred and thirty-one boxes from an art historian was a complete, publishable dataset.



Corn

Narrow the problem until the expert's judgement fits in it.



Herman

And the open question this leaves me with, the thing I'd like to see built, is an annotation environment organised around the expert's taxonomy rather than the model's output format. Hannah's label tree as the front page, her skip decisions logged, her free-text treated as signal rather than decoration. The tools are all pointed at the model. The interesting one points at the architect.



Corn

Our producer is Hilbert Flumingtop, who keeps a shoebox we will never see. If you enjoyed this one, try episode twenty, Architectural AI; episode twenty-one ninety-six, The Invisible Workforce Behind AI; and episode eight, Building Your Own Whisper. This has been My Weird Prompts. If the show earns a corner of your week, a review helps other people find us.



Herman

We'll be back soon. See you tomorrow.

SOURCES

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

1. **CVAT repo, README and releases v2.81.0 [primary]**
   v2.83.0 ; retrieved 2026-11-15; last push 2026-11-15; latest release v2.83.0 (2026-11-12).
   https://github.com/cvat-ai/cvat

2. **Label Studio repo, README and releases 1.23.1/1.23.2/nightly ; retrieved 2026-11-15; last push 2026-11-14; stable 1.23.2 (2026-09-29). [primary]**
   https://github.com/HumanSignal/label-studio

3. **Labelme repo, README and releases v7.9.1 [primary]**
   v7.11.0 ; retrieved 2026-11-15; last push 2026-11-14; latest v7.11.0 (2026-11-14).
   https://github.com/wkentaro/labelme

4. **Supervisely SDK repo, README and releases v6.74.54 [primary]**
   v6.74.56 ; retrieved 2026-11-15; last push 2026-11-12.
   https://github.com/supervisely/supervisely

5. **Barz & Denzler, WikiChurches: A Fine-Grained Dataset of Architectural Styles with Real-World Challenges, NeurIPS 2021 Datasets & Benchmarks Track, v2 revised 2021-10-18; full text read (pp. 1 [primary]**

   5) ; retrieved 2026-11-15.
   https://arxiv.org/abs/2108.06959

6. **CVAT Manual Annotation docs, v2.83.0 ; retrieved 2026-11-15 (page returned mostly navigation; feature detail taken from the repo README and release notes instead). [primary]**

   https://docs.cvat.ai/docs/annotation/manual-annotation/

7. **Zhong et al., ArchiLense: A Framework for Quantitative Analysis of Architectural Styles Based on Vision Large Language Models, v3 revised 2025-08-02; retrieved 2026-11-15.**

   https://arxiv.org/abs/2506.07739v3

8. **Kubba et al., Multimodal Cultural Heritage Architectural Style Classification for Residential Buildings in the UAE Based on CLIP Embeddings and SVM, submitted 2026-09-15 (SITA 2025); retrieved 2026-11-15.**

   https://arxiv.org/abs/2609.17181v1

9. **Wei, Li & Shamir, Computer Analysis of Architecture Using Automatic Image Understanding, J. Data Mining & Digital Humanities, v4 revised 2020-07-09; retrieved 2026-11-15.**

   https://arxiv.org/abs/1807.04892v4

10. **Shrestha et al., FEVA: Fast Event Video Annotation Tool, v2 revised 2023-01-07; retrieved 2026-11-15.**

    https://arxiv.org/abs/2301.00482v2

11. **Banerjee et al., SWAN**
    Enabling Fast and Mobile Histopathology Image Annotation through Swipeable Interfaces, submitted 2025-11-11; retrieved 2026-11-15.
    https://arxiv.org/abs/2511.08271v1

12. **Längkvist et al., Interactive user interface based on Convolutional Auto-encoders for annotating CT-scans, submitted 2019-04-26; retrieved 2026-11-15.**

    https://arxiv.org/abs/1904.11701v1

13. **Schomaker, A large-scale field test on word-image classification in large historical document collections, submitted 2019-04-17; retrieved 2026-11-15.**

    https://arxiv.org/abs/1904.08421v1

14. **Image Annotation Tools Compared: CVAT, Label Studio, Roboflow, V7 and Potato; retrieved 2026-11-15.**

    https://www.potatoannotator.com/docs/guides/image-annotation-tools-compared

15. **12 Best Image Annotation Tools in 2026: CVAT, Roboflow & More; pricing as displayed August 18, 2026; retrieved 2026-11-15.**

    https://sekin.in/12-best-image-annotation-tools-for-efficient-labeling-in-2026/

16. **Best Computer Vision Annotation Tools 2026; retrieved 2026-11-15.**
    https://humansintheloop.org/best-computer-vision-annotation-tools-2026/

17. **Ultralytics, Best Image Annotation Platforms for Vision Teams; retrieved 2026-11-15.**
    https://www.ultralytics.com/blog/best-image-annotation-platforms