Footage Is Not Supervision:
Grounding Egocentric Video for Robot Learning

Rey Pocius
DataLab at Protege

Abstract

Footage is not a dataset. An hour of egocentric video is a recording of something that happened; a robotics training set is a versioned, task-specific disposition of such recordings. What separates the two is whether anyone can see what is in the footage. A manifest reports a duration, a frame rate and a task name, and none of those says whether hands are visible, whether the wearer is touching the object they are supposed to be working on, whether that object persists across the frames a model will consume, or when the action starts and stops. This note is a tour of the layers that answer those questions, shown as output rather than as summary statistics: hand landmarks on industrial assembly work, object detection under a prompted and a fixed vocabulary, segmentation masks that carry an identity across frames, hand-object contact resolved into episodes, and language attached to a region and to an interval. For each layer we say what curation decision it makes possible, and what it still gets wrong, because both are visible in the same picture. The argument is that enrichment is not decoration on top of a corpus. It is the step that turns a pile of video into something a specification can select from, and the step whose failures are the ones most likely to be mistaken for facts about the data.

1 A Manifest Is Not a Description

An egocentric corpus at pretraining scale is not collected, it is assembled. Deliveries arrive from independent capture programmes, each with its own hardware, protocol and post-processing, and we layer our own transcodes, clip segmentation and annotation on top. What arrives with each clip is a short record: a duration, a declared frame rate, a free-text task name, a region. Every one of those can be wrong, and even when they are right they do not answer the question a training run asks, which is whether the thing to be learned is visible in the frames the loader will sample.

So we separate two objects that are usually conflated.

> The corpus is the shared evidence base; a dataset is a versioned, task-specific disposition of that evidence.

A corpus is measured, enriched and validated once. A dataset is drawn from it against a written specification, for a named consumer, with its own gates, bands and coverage constraints. The distinction is operational rather than definitional, and it complements dataset-documentation practice [6], which asks that composition, provenance, intended uses and limitations be written down: what we add is that the disposition itself is versioned, so the same evidence base can answer to several such documents at once. Figure 1 places the stages in order. This note is about the middle of that picture, the part where footage acquires a description detailed enough to select on, and it argues the case by showing the descriptions themselves.

Six questions, not one score. Before the examples, one piece of vocabulary. “Quality” is used for at least six different questions about a clip, and they have different tests and different remedies (Table 1). Most of this note lives in the third and fourth rows, task utility and grounding, because those

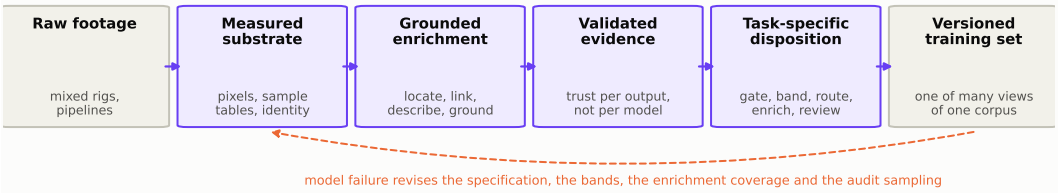

Figure 1: Where enrichment sits. Measurement establishes that the media are what they claim to be. Enrichment establishes what is in them. Only then can a specification select, and only then does a downstream failure have anything to point at.

Dimension	Governing question	Available action
Integrity	Are the media, timing, streams and identities physically valid?	Repair, quarantine, gate
Authenticity	Is this the intended viewpoint and content type?	Gate, route
Task utility	Does the learner see the interaction or transition it needs?	Band, route, enrich
Grounding	Are the generated annotations aligned, interpretable and validated?	Accept, relabel, enrich, review
Composition	Does the set cover the intended distribution?	Balance, cap, constrain
Governance	Is the material permitted and appropriate to use?	Restrict, redact, exclude

Table 1: Six dimensions a single quality score merges. Integrity failures are repairable or disqualifying. Task utility is usually neither, and is better banded than gated, because glare, blur, occlusion and awkward viewpoints are properties of deployment rather than only of poor capture.

are the rows that enrichment creates. A clip cannot be assessed for task utility until something has established what happens in it, and an annotation cannot be trusted until something has established that it means what it says.

Two consequences are worth stating plainly, because they are why a capability tour is the right way to make this argument. Clean footage is not necessarily useful: a perfectly exposed clip of an empty worktable contains no manipulation to learn from, and no signal-level measurement will say so. And a valid clip can be valuable for one objective and unsuitable for another, so there is no single admitted pool, only dispositions against specifications.

2 Locate: Where the Hands Are

For a manipulation policy, the hands are the action. If they are not visible, or not resolvable into a pose, the clip may still be pleasant video and it is not a demonstration of anything. Figure 2 shows what the hand layer emits on three industrial assembly and repair sequences: a full 21-landmark skeleton per hand, an egocentric handedness label, and a track identifier that persists while the hand stays in view. The estimator is the MediaPipe hand landmarker [30], run through the pinned `hand_landmarker` task asset.

Three things about this layer decide how it can be used, and all three are visible above.

Handedness is a viewpoint question, not a model question. The standard advice for this family of estimators is to swap the handedness label, because the usual deployment is a camera pointed at a person and the image is mirrored relative to them. A head-mounted camera looks along the wearer's own gaze, so their right hand appears on image right exactly as in a mirror and the raw label is already correct; applying the conventional correction inverts it. That is a systematic left-right relabelling of every hand in the corpus, and it is the kind of error that survives every aggregate check because the totals do not change.

The estimator degrades exactly where the interaction is. Egocentric hands are near-field, frequently motion-blurred, and often half-occluded by the object being worked on. In the first panel

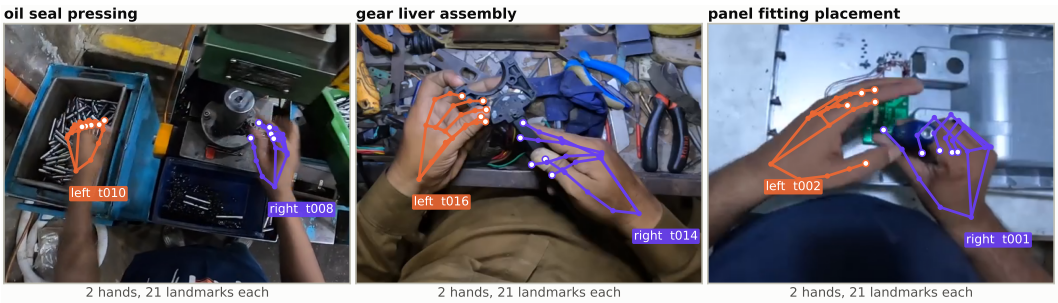

Figure 2: Hand pose from the MediaPipe hand landmarker [30] on three industrial sequences, drawn on the pixels the estimator was given. Each hand carries 21 landmarks in a fixed canonical order, wrist first, then thumb, index, middle, ring and pinky; fingertips are hollow. Colour is egocentric handedness and the tag gives the hand's track identifier. The crops are tight because that is where the argument is: a clip is useful for manipulation to the extent that this structure is recoverable, and no clip-level statistic reports it.

a hand reaches into a parts bin and the skeleton is compressed against the occluding edge; in the third a hand is behind the panel it is fitting. These are the frames a policy most needs and the ones the estimator finds hardest, so hand-pose coverage is not distributed randomly across a clip. It is anti-correlated with the moments of interest. These are established difficulties rather than a quirk of our footage: egocentric hand-pose benchmarks are built around exactly this regime [19].

Recovering a hand can take a second pass. Running the estimator on the full frame misses hands that occupy a small fraction of it, and a second pass over a crop around a detected hand region recovers a substantial share of them. Whether a hand was found directly or by a crop pass is recorded per instance, because a consumer counting hands should know the count depends on how hard the pipeline looked.

Body pose is the sibling layer, and it is a limitation rather than a capability here: a head-mounted camera sees the wearer's forearms and almost nothing else of the wearer, so a whole-body keypoint model, `vitpose-plus-huge` [34] over person boxes from `yolo11x`, returns a few visible joints and much low-confidence extrapolation. We keep it, record the per-keypoint visibility, and do not gate on it.

3 Locate: What the Objects Are Called

The other half of locating is the objects. Figure 3 runs two detectors over the same two frames: an open-vocabulary detector prompted with nouns drawn from the task description, `grounding-dino-base` at a box threshold of 0.3 [16], and a supervised detector, `yolo11x` with BoT-SORT association, whose fixed 80-class vocabulary is that of COCO [15].

The comparison is not about which detector is better. It is about what a vocabulary can express. The fixed vocabulary is confident about the wrong things for this purpose: it has a class for `person` and none for `hand`, so in egocentric manipulation footage it cannot name the single most important object in the frame. It also lacks `tap`, `cutting board` and `drawer`, so a coverage claim about those attributes cannot be computed from it at all.

The prompted vocabulary can express them and pays for it in reliability. Its boxes are roughly right and its nouns frequently are not: a plausible-looking wrong noun is the characteristic failure of this family of models, and the more domain-specific the scene the more of them appear. Measured against human labels elsewhere, the verdict we act on is split. The same output is admissible as evidence that a hand is present and where it is, and inadmissible as an authority on what the hand is holding, so the geometry is used and the class names are carried as unvalidated.

That split has a consequence for composition: objects are the attribute most often requested as a diversity axis and the one we can least support, because the labels that would define it are the ones we do not trust. Coverage constraints in our deliveries are stated over sites, scenes, environments and tasks, and object coverage is reported as provisional.

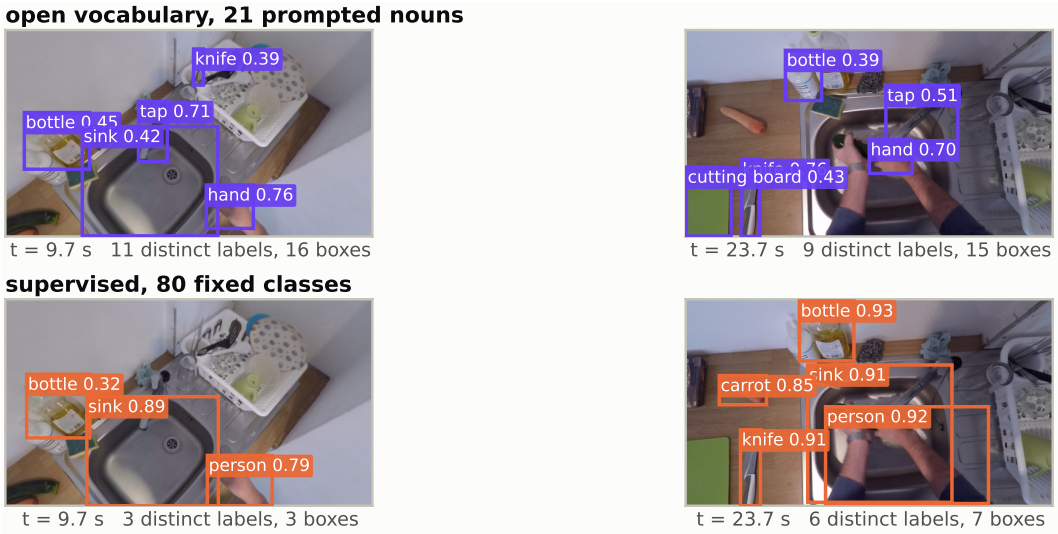

Figure 3: The same pixels under two vocabularies. Above, `grounding-dino-base` [16] prompted with nouns from the task description finds the task-relevant things, including the hands and the tap. Below, a supervised 80-class detector is confident and generic: it finds the sink and the bottle, calls the wearer a person, and has no class for a hand at all. Only the highest-scoring box per label is drawn.

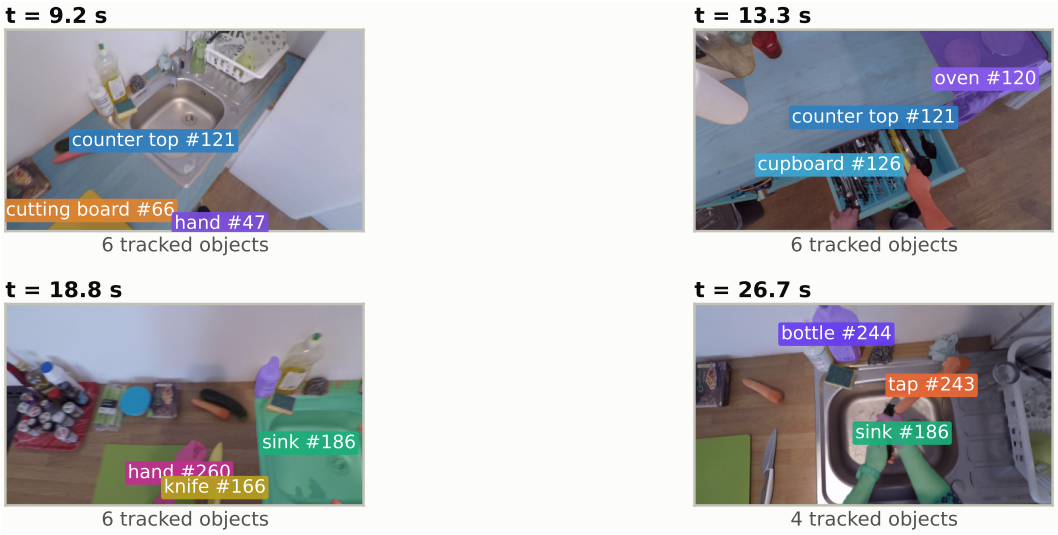

Figure 4: Tracked masks from `sam2.1-hiera-large` [23] with persistent identities, four moments of one clip. Colour is object identity, so the counter top in the first two panels is one object and the sink in the last two is another. Identities enter and leave as the wearer moves through the room, which is what the changing counts record. Masks are drawn at partial opacity so the underlying pixels stay visible.

4 Link: Identity, and the Difference Between Gone and Unsupported

A frame label says what was visible once. It says nothing about a transition, and a transition is what most robot-learning objectives consume. The first layer that says anything about time is one that carries an identity from frame to frame. Figure 4 shows tracked segmentation masks over one clip: a mask per object per frame, and a colour per identity, so the same colour in two frames is a claim that this is the same object. Masks come from `sam2.1-hiera-large` [23], prompted by the same open-vocabulary detector and reseeded every twelve frames. Temporally consistent pixel masks with active objects and hand-object relations are by now an established annotation target for egocentric video [5], which is the precedent this layer follows.

Figure 5: Hand-object contact as an overlay and as episodes. Left and centre, a hand mask and the mask of the object it is in contact with, at two moments of the same clip. Right, every episode in the clip by object, split by the two mechanisms that produce them: a mask intersection test, and a test for hand landmarks falling inside an object mask. The two are derived from different inputs and do not produce the same episodes.

This is the layer that makes object permanence checkable, and it is also the layer most often misread. Mask propagation runs only as long as the detector keeps proposing the object, so a track that ends carries two possible meanings: the object left the scene, or the detector stopped supporting it. Both are consistent with the annotation as stored, and nothing in the record distinguishes them. A criterion keyed on track length is therefore keyed on a joint property of the tracker and of the detector that prompted it, which is why we keep this layer per object and per interval rather than reducing it to a per-clip score.

5 Link: Contact Is the Unit of Interaction

Hands and objects located separately are still only a scene description. What a manipulation policy needs is the relation between them, resolved into episodes with a start and an end. Detecting hands together with the objects they are in contact with is a long-standing task in its own right [27]; what we add here is the resolution into episodes with parameters attached. Figure 5 shows both: the overlay at two moments, and every contact episode in the clip laid out per object.

The episode view is what turns contact into something a specification can ask for. “At least one sustained contact episode with a handled object” is a criterion; “contact detected” is not, because it does not say for how long, with what, or whether the same object was involved throughout. The per-object lane also exposes the shape of a clip at a glance: here most of the contact time is with a sink and a tap, and the episodes involving the knife are shorter and interleaved, which is a different kind of demonstration from a single long grasp.

Two properties of this layer matter more than its coverage.

Contact is parameterised before it is measured. The criterion is proximity, not overlap: a hand is in contact with an object when the two masks come within a stated dilation of one another. That makes the layer unable to hallucinate, since it runs no model of its own, and entirely dependent on one scalar. Raise the dilation and every frame carries contact; lower it and clips that plainly show manipulation report none. There is no observer-independent value, because the right number depends on resolution, on how far the hand is from the camera, and on how tightly the masks fit. So the record carries the dilation it was computed at, and the detector and landmark inputs it was computed from, and a consumer comparing two clips can at least tell whether they were scored under the same rule.

Two things called contact are not one field. The lane in Figure 5 is split for a reason. Mask intersection and landmark-in-mask are two derivations from two different upstream layers, and a third mechanism appears elsewhere in our stack, in which an object box is marked in contact when it intersects a hand box grown by a fixed fraction of its own size. All three answer to the word contact and none is a substitute for another. Pooling them into one boolean would produce a field whose meaning depends on which pipeline happened to write it.

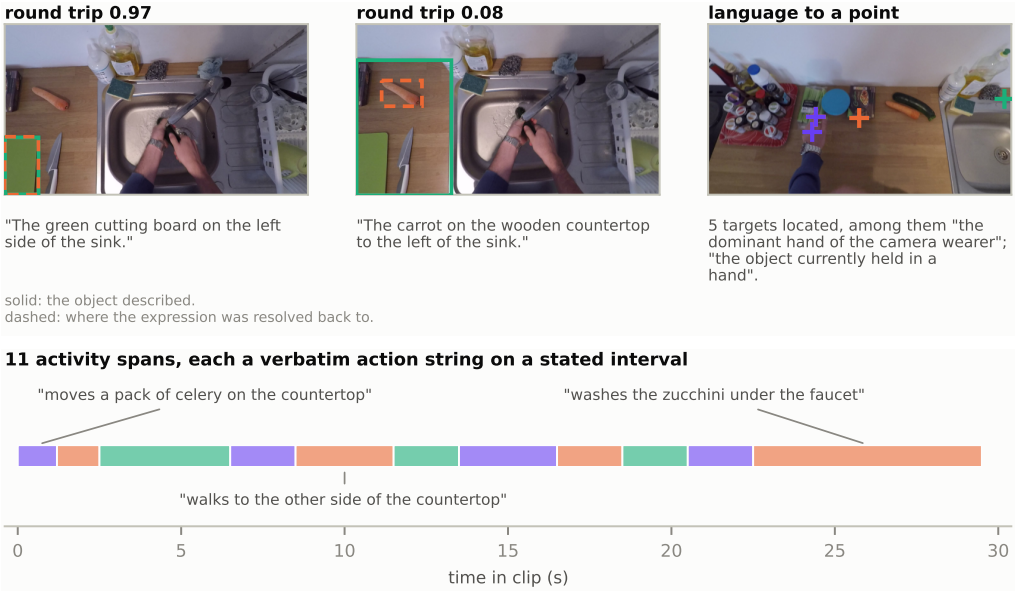

Figure 6: Language attached to a region and to an interval, produced and resolved by `Qwen3-VL-8B-Instruct`. **Above**, an expression that resolves back to the object it describes (left); an expression from the same clip that does not, where the description is accurate English and the model resolves it somewhere else (centre); and a keyframe with five language targets located as points (right). Both expressions are fluent, and fluency is uninformative about whether either is usable. **Below**, the same clip resolved into activity spans, each carrying a verbatim action string and explicit bounds, so a window can be selected by what is happening in it rather than by where it falls in the file.

6 Describe and Ground: Language That Points at Something

The layers above are geometric. The layers that describe are the ones that let a person, or a retrieval system, or a conditioned policy, address the footage in words. Their failure mode is different in kind: geometry is wrong in ways you can see, and language is wrong in ways that read perfectly well.

Figure 6 shows the mechanism we rely on most for that reason. A model is shown a frame with one tracked object outlined and asked to write an expression that picks that object out. It is then shown the clean frame and the expression alone, and asked which box the expression refers to. The overlap between the box it returns and the object it was describing is a round trip: it needs no human label, and it scores the annotation rather than the model. Coupling a speaker to a listener is not new; it has been used to *train* referring-expression models [36]. The difference here is the purpose: we run the inversion at inference only, as an output-level check on an annotation we have already produced. Both the expression and its resolution come from `Qwen3-VL-8B-Instruct`, with referring segmentation from `Florence-2-large` [33].

The two expressions are the argument. One resolves almost exactly. The other, describing a carrot by its position relative to the sink, resolves to a different region entirely, and reading the sentence gives no clue which of the two you are holding. A low round trip is a review flag rather than a verdict, because it does not distinguish an unusable expression from a good one applied to a scene containing two similar candidates; ambiguity is not unusability. But the check costs one extra inference call per unit rather than a labelling pass, and any layer whose production can be inverted admits the same treatment. That generalisation is the most useful thing in this section.

The interval half of describing is the lower panel of Figure 6: the clip resolved into activity spans, each with a verbatim action string and explicit bounds.

That is the layer that makes window-level supervision possible at all, and its properties to watch are about vocabulary and timebase rather than accuracy.

Free-text action strings drift. Over one continuous session the upper levels of an activity taxonomy stay fixed while the action level re-invents itself window by window, so the same behaviour arrives under several names and overlapping windows disagree at equal self-reported confidence. A criterion

keyed on the action string is keyed on something the labeller regenerates each time, and the fix is a controlled vocabulary rather than a better model.

Retrieving a moment in a video from a natural-language query is a task with its own benchmarks [14]; what we have here is the output of that task with nothing to score it against, which is why Section 10 keeps it labelled generated rather than validated.

And these layers do not share a clock. Locating layers run at the annotation grid rate, describing layers one to two orders of magnitude below it, and none at the rate the footage was captured at. Some spans are stored relative to the start of the annotated window rather than to the start of the source file. A record that merges these fields without carrying each one's timebase has merged claims about frames with claims about spans, and the resulting per-clip row cannot be safely filtered on any of them. Grounding, in the sense we use it, is exactly this discipline: every output travels with the frame, interval, track or region it is a claim about, in a stated timebase, along with the model that produced it and the parameters it was produced under.

6.1 Dense captioning, run for this note

Whole-clip captions summarise. Dense captioning enumerates: one short present-tense clause per short window, at a stride, so the description has a timebase, which is the shape of the dense-captioning task as it is usually posed [12]. We ran it for this note on the same kitchen clip at a 2-second window and a 1-second stride, two frames per window, a local 7-billion-parameter vision-language model, 40 output tokens each. Twenty-eight windows produced twenty-eight clauses (Figure 7). The model is `Qwen2.5-VL-7B-Instruct`, run locally. The output is fluent from beginning to end and it is not uniformly true, which is the failure mode object-hallucination work in captioning has documented: fluent systems name things that are not present, and ordinary language-quality scores do not necessarily expose it [24].

The second half is good. From about 18 seconds the clauses read `washing a cucumber under the tap`, `rinsing a knife under the tap`, `washing a knife under the tap`: specific, correctly localised, naming the right object and action. The first half is not. At 3 to 5 seconds, in a kitchen, the clause is `tightening a screw with a screwdriver`, and there is no screwdriver; the wearer's hands are in a low cupboard. What makes that instructive rather than merely a bad frame is the window overlapping it, which reads `retrieving a grater from the cabinet`, and there is a grater, plainly visible in the same cupboard. Two windows sharing a second of footage, one correct and one an invented tool action, delivered in the same register with no confidence attached to either. A third pattern sits between: at 16 to 18 seconds the clause is `cutting a cucumber with a knife` on a window with no hands in it at all. The action named does happen in the clip, just not then, and a consumer selecting windows by keyword would pull this one as a cutting demonstration.

Two structural properties matter as much as the errors. *Merging barely helps, because the wording moves*: collapsing adjacent clauses that agree reduced twenty-eight windows to twenty-four spans, because `washing a cucumber under the tap` and `rinsing a cucumber under the tap` describe one action in two ways and a string comparison keeps them apart. Twenty distinct opening phrases appear across twenty-eight windows. Density here is a property of the stride, not of how much happened. And *layers disagree about nouns*: this layer says cucumber, the activity-span layer over the same seconds says zucchini. Neither is wrong in English and they are not the same string, which is all a filter keyed on object names can see.

The conclusion is not that dense captioning is unusable. It is cheap, it is the only layer that describes every second of a clip, and its good half is genuinely good. It is that density and fluency carry no information about correctness, so the layer earns a place as retrieval and triage and as a candidate for review, and not as ground truth.

7 Stereo: Metric Structure, and Why a Robot Needs It

Everything so far is measured in pixels. A pixel offset is enough to say a hand is near an object and not enough to say how far away either is, which is the quantity a robot acts on. Where the capture is a calibrated stereo pair, that changes (Figure 8). The clip is a rectified pair from Ego-1K [13].

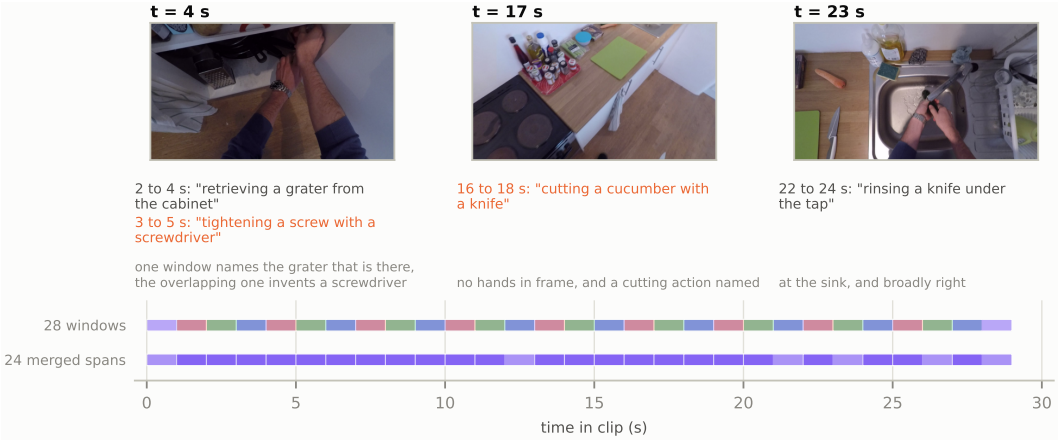

Figure 7: Dense captioning [12] run for this note with `Qwen2.5-VL-7B-Instruct`: one clause per 2-second window at a 1-second stride, twenty-eight windows over thirty seconds, quoted verbatim. **Left**, two overlapping windows, one naming the grater that is in the cupboard and one inventing a screwdriver. **Centre**, a cutting action named on a window with no hands in it. **Right**, a clause that is broadly right. **Below**, every window and the twenty-four spans they merge into; the merge is small because the wording changes even where the action does not. Orange marks the two clauses we checked against the frame and found wrong.

Disparity becomes depth by a formula with no learned scale in it: distance is the focal length times the baseline, over the disparity. On the frame shown, a learned stereo matcher, `FoundationStereo` at its `vit-large` checkpoint [32], returns valid disparity on about 96% of pixels, giving a median scene depth just under a metre and a nearest 5% at roughly 0.11 m, which is the wearer's own hands. Those are metres, recovered from geometry, and they are the same metres in every frame. Three things follow that a monocular pipeline cannot supply.

Object size becomes checkable. With metric depth, the physical size of a tracked object can be estimated and compared against what that noun should measure. On this clip that check runs over fourteen tracked objects and finds thirteen physically plausible and one not, plus two whose depth is internally incoherent across their own mask. A classical semi-global matcher [9] is run alongside as an independent second opinion. That is a validation signal for the segmentation and naming layers which simply does not exist without scale: a monocular pipeline cannot tell a mug from a mug-shaped building.

Ego-motion acquires units, and the alternative is worse than unitless. Panels (c) and (d) run three ego-motion estimators over the identical left frames. The stereo path lifts matched features to 3D using the depth above and solves for pose, returning a trajectory in metres: about 0.53 m of total camera translation over nine seconds of desk work, at a mean reprojection error under a pixel. The stereo path solves for pose by RANSAC-guarded perspective-from-three-points on the depth-lifted matches; the monocular paths are pyramidal Lucas-Kanade flow over Shi-Tomasi corners, and an essential-matrix decomposition over ORB features [26], all three from OpenCV and none of them learned. The monocular paths return no scale at all, and worse, they disagree about the rotation. One reports the camera as very nearly static across the whole clip. The other integrates to more than nine hundred degrees of cumulative rotation in nine seconds, which did not happen. All three report a tracking-loss rate of zero. The stereo path is not merely more informative here; it is the only one of the three whose output could be checked against a physical quantity, and the only one that agrees with what the footage shows.

Monocular depth borrows a scale it will not have at inference. It is tempting to substitute a monocular depth model, and the comparison on this clip shows what that costs. The model is `Depth-Anything-V2-Large` [35], and it returns relative inverse depth, which has to be fitted to something before it means anything. Fitted per frame by least squares onto the stereo disparity, its median relative depth error is about 15%, which sounds tolerable until you notice two things. The fit uses the stereo depth as the answer, so a deployment without a stereo rig cannot perform it. And the scale the fit chooses is not even stable within this one clip: the fitted scale factor varies across frames

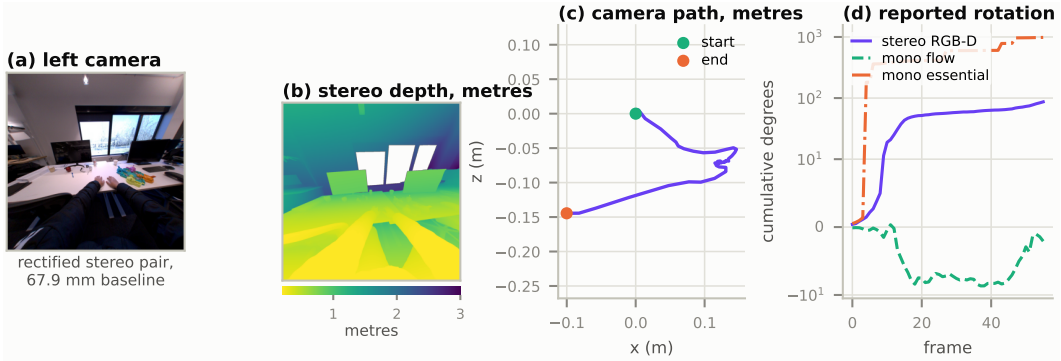

Figure 8: What a calibrated stereo pair adds, on an Ego-1K clip [13]. **(a)** The left camera of a rectified pair, 67.9 mm baseline. **(b)** Metric depth from `FoundationStereo` [32] disparity, distance in metres, with white where no valid disparity was found; median 0.85 m, nearest 5% at 0.11 m, valid on 96% of pixels. **(c)** The camera path recovered by lifting matched features to 3D with that depth and solving for pose, in metres. **(d)** Cumulative rotation reported by the same stereo path and by two monocular estimators on the identical frames, on a symmetric log scale. The monocular paths have no scale and do not agree about rotation; one reports a nearly static camera and the other more than nine hundred degrees in nine seconds. All three report zero tracking loss.

by about a third of its own value. A policy that learned distances from monocular depth would be learning distances measured with a ruler whose length changes between frames.

For robot learning this is the difference between footage that demonstrates an activity and footage that demonstrates a reachable geometry. A grasp is at a distance; a collision is at a distance; a world model that respects object permanence has to keep objects the same size as the camera moves. Stereo makes those quantities available per frame, in units, and checkable. It is why we treat calibrated stereo capture as a distinct and higher tier of the corpus rather than monocular footage with a spare camera, and why a missing per-clip calibration is recorded as a gap rather than defaulted.

7.1 The environment, reconstructed

Depth on one frame is a measurement. Depth on every frame plus the poses between them is a scene. Figure 9 fuses them: each frame's depth back-projected into the world using the estimated pose, voxelised at a little over a centimetre, and rendered from two viewpoints that no camera ever occupied.

We show it for one reason. Panels (b) and (c) are not a depth image redrawn; they are a reconstruction that can be looked at from the side and from above, which is only possible if the geometry and the poses are consistent with each other. Reading the monitors, the desk edge and the cluster of coloured blocks in the same places in all three panels is a check on the whole stereo chain at once, and a cheap one: it takes nine seconds of footage and no ground truth.

Three honest limits sit alongside it. This is frame-to-frame odometry with **no loop closure and no bundle adjustment**, so error accumulates along the sequence and a longer clip would smear. The render keeps only points within a couple of metres of the camera path, because beyond that the disparity is a pixel or two and the depth error grows with the square of distance. And the wearer's own torso and arms are cut from the near field, or they fill the foreground of every view of the room. What survives all three cuts is about a sixth of the fused points, and that is the honest yield of nine seconds of head-mounted stereo.

8 What the Layers Make Possible

Building all of this changes what a specification can say. Without enrichment a corpus can be filtered on encoding, resolution, exposure, focus and duration, which is a filter on how the footage was recorded. With enrichment it can be filtered on what happens in it, and the criteria below are the ones our specifications use, each traceable to a layer above.

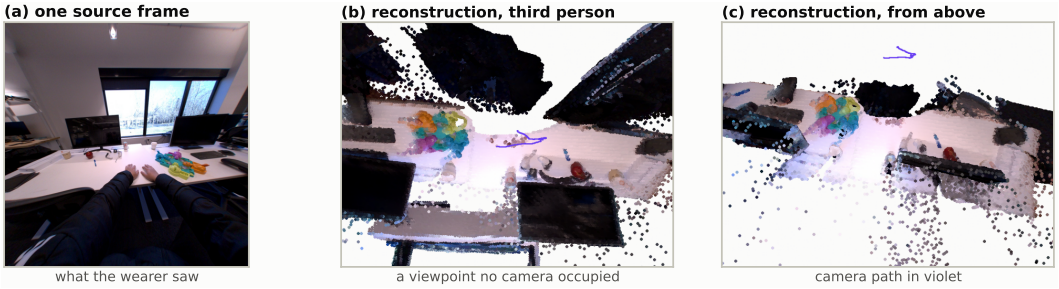

Figure 9: The room, reconstructed from stereo depth and estimated poses. **(a)** One source frame. **(b)** The fused point cloud from behind and above the wearer, a viewpoint no camera occupied. **(c)** The same cloud from overhead, with the recovered camera path in violet. Rendered with a z-buffered splat rasteriser, so the figure needs no graphics stack. Frame-to-frame odometry only: no loop closure, no bundle adjustment, points beyond a couple of metres from the path discarded, and the wearer's own body cut from the near field.

- *Is the action legible?* Hands present, with a resolvable pose, for a sustained fraction of the window rather than in one frame of five.
- *Is the wearer working on something?* At least one contact episode of a stated minimum duration, with the parameters it was derived under recorded.
- *Does the object persist?* A track that spans the window, distinguished from a track that ends because detector support lapsed.
- *Is the label coupled to the frames?* An activity span whose interval overlaps the window, rather than a clip-level task name inherited through segmentation.
- *Can this window be addressed in language?* A referring expression or a span that resolves back to what it describes.
- *How far away is it?* Where the capture is calibrated stereo, a distance in metres to the handled object, a metric object size that can be sanity-checked, and a camera path in real units.
- *Which consumer is it for?* A window with legible grasps and a coupled label suits a manipulation policy; a window with object permanence, camera motion and a non-trivial transition suits a video world model. The same footage routes differently.

None of these is available from a manifest, or from signal quality alone. They are the return on the enrichment step, and they are why we treat description as a precondition for selection.

9 Curation Is a Disposition, and Enrichment Decides What Is Possible

Given those criteria, a curation pass does not sort footage into good and bad. It assigns each unit a disposition, and there are more useful dispositions than two: *accept*; *repair* where the defect is in our handling, not the capture; *gate* for a hard requirement or invalid media; *band* where the axis is a difficulty axis; *route* where a unit suits one consumer and not another; *enrich* where a criterion needs a layer that has not been run; *relabel* where the annotation is wrong but the footage is fine; *reweight*; *quarantine* where a person must look first; and *reject*.

Two of those exist only because enrichment exists. *Enrich* covers a unit that is neither in nor out: the specification asks about contact, the layer has not run, and the honest state is `unmeasured` rather than failure. *Relabel* covers a unit whose footage is fine and whose annotation is not, a distinction you cannot draw until the annotation is inspectable. A pipeline whose only outputs are pass and fail collapses both into rejection and loses the footage with the bad label.

The distinction that does the most work is between data that is invalid and data that is merely hard. A truncated file or a viewpoint that was not ordered is a gate, because no consumer wants it. Blur, glare, occlusion, low motion and awkward viewpoints are not gates. They are the conditions a deployed system will meet, and the examples here are full of them: the occluded hand in the parts bin, the knife half in the sink, the hand behind the panel. A set restricted to the cleanest band has removed exactly the examples that would have prepared a model for deployment. Banding keeps them in stated proportion, with the fitted band edges travelling alongside any selection drawn from them.

Finally, filtering concentrates. Every criterion above is evaluated on a unit without reference to what else is in the set, so one correlated with a site, a rig or a task removes that site, rig or task faster than it removes anything else. Coverage is therefore a separate stage applied after per-unit selection, with caps on site and scene type, and it is reported as an effective count rather than a label count.

10 What These Layers Cannot Tell You

A tour of capabilities invites over-reading, so this section is the counterweight. Every item below is visible in the figures already shown.

Trust attaches to an output, not to a model. The clearest case is the detector in Figure 3, whose boxes are usable and whose class names are not. One artifact, two verdicts. A single quality score for a model cannot express that, and a pipeline that admits or rejects models wholesale will either lose the useful geometry or import the unreliable vocabulary.

A derived layer inherits every dependency of its inputs. Contact runs no model, and it depends on masks, on hand landmarks and on a dilation. It is exactly as good as the segmentation underneath it. When a mask drifts off its object, the contact episode computed from it is confidently wrong, and nothing in the contact record says so.

Absence has several meanings and they are not interchangeable. No detected contact is not contact having gone unmeasured, and a track that ends is not an object that left. A layer never run is `unmeasured`, one that does not apply to the modality is `unsupported`, a probe that errored is `failed`, a property measured and found missing is `absent`, and a numeric zero means something was measured to be zero. Tabular output flattens all six into a blank, and the flattened version misleads in one direction: it makes gaps look like findings.

Fluency is not grounding. The two expressions in Figure 6 read equally well and one is unusable; two overlapping windows in Figure 7 read equally well and one invents a tool. Neither difference shows up in any measure of how much text was produced, so provenance has to travel with the annotation for it to be recoverable at all.

Generated is not validated. Some layers produce plausible output with nothing to check it against. Temporal grounding is the example in our stack: it answers queries with intervals, and on footage that ships no timed narration there is no ground truth to score those intervals. The output is real and it is unvalidated, and it stays labelled that way wherever it appears, including in Table 2.

11 The Boundary of What Enrichment Can Reach

Two boundaries deserve stating explicitly, because a capability tour is exactly the kind of document that blurs them.

Human video does not contain robot actions. Egocentric footage holds no robot-native control actions, no torques or forces, no proprioceptive state, no robot embodiment, and no outcome of an intervention that was not performed. No annotation produces these, because they were never recorded, and a grasp seen from a head-mounted camera is not a gripper trajectory. What the layers here expose is narrower and still substantial: visual affordances, hands and manipulated objects, contact episodes with their start and end, object persistence, task and step structure, and the spatial and temporal relations between them. Those are the observable correlates of manipulation, which is what makes the footage worth curating carefully rather than a solution to the human-to-robot embodiment gap. Our system also does not answer whether a clip suits a specified task; that inference belongs to whoever writes the specification.

Detecting a face is not a governance policy. The industrial frames in Figure 2 were selected from a face-screened subset, and the crops are tight partly for legibility and partly because a wider crop on one of them included legible text on the wearer's clothing. That screen is an identifiability measurement, one input to a governance process rather than the process. Whether material may be

Capability	Native unit	What checks it	Standing
Hand landmarks	Frame, track	Agreement with human labels on a labelled probe; FreiHAND [37] for pose error	Validated for presence
Object detection, geometry	Frame	Agreement with human labels on a labelled probe	Validated for presence
Object detection, class names	Frame	The same probe, which refutes it	Refuted for this output
Mask tracking and identity	Track	Nothing yet; coverage is confounded with detector support	Experimental
Contact episodes	Event	A parameter sweep only, with no ground truth	Experimental
Activity spans	Span	Verbatim inspection; vocabulary is uncontrolled	Pilot
Dense captioning	Window	Verbatim inspection against the frames; two of three checked clauses wrong	Pilot, triage only
Referring expressions	Region	A self-inverting round trip, needing no human label	Validated by self-check
Temporal grounding	Span	Nothing on this footage	Generated, unvalidated
Body pose (ViTPose [34])	Frame	Per-keypoint visibility, which is low from this viewpoint	Recorded, not gated
Stereo metric depth	Frame	Geometry, not learning, for the scale; a second matcher agrees within 2 px on three-fifths of pixels; metric object sizes plausible on 13 of 14	Validated where calibrated
Stereo ego-motion	Transition	Sub-pixel reprojection error, and the only path here with units	Validated on one clip

Table 2: What stands behind each layer shown in this note. The third column decides the fourth. Nothing here is labelled production, because nothing in our sources establishes that status, and where the reference labels are themselves model output rather than human adjudication we say so rather than calling the result an accuracy.

used depends on consent, licensing, capture-region restrictions, access controls, retention and the treatment of bystanders, screens and personal information in frame, and a detector decides none of it. It measures faces, so it says nothing about a visible document, a badge or a plate, and a text search over a built document cannot see a mark burned into a raster figure. The kitchen and stereo examples come from openly licensed research datasets, EPIC-KITCHENS-100 [4] and Ego-1K [13], chosen for that reason. What our system records is each annotation's producer, prompt hash, timestamp and licence class, so a delivery can be filtered by licence afterwards.

12 Measuring Diversity Without Labels

Coverage in Section 9 is reported over labels: sites, scenes, environments, tasks. That is only as good as the vocabulary, and this note has already shown two ways the vocabulary fails, in detector class names and in free-text action strings. So it is worth asking whether the set can be measured for diversity without asking anything to name what is in it.

Embeddings offer that. Encode sampled frames from every clip, pool to a vector per clip, and the geometry of the resulting cloud is a description of how varied the set is that no label touched. Figure 10 does it over the twenty-three clips we hold locally, using the `clip-vit-base-patch32` image-text encoder [22] at ten frames per clip.

The useful part is the disagreement between ways of counting. The set is twenty-three clips. Its industrial half carries sixteen distinct task names. Grouped coarsely it is four families. And the effective dimension of the embedding cloud, the exponential of the entropy of its eigenvalue spectrum, is about fourteen; that is the effective rank of the cloud in the established sense [25]. Four numbers,

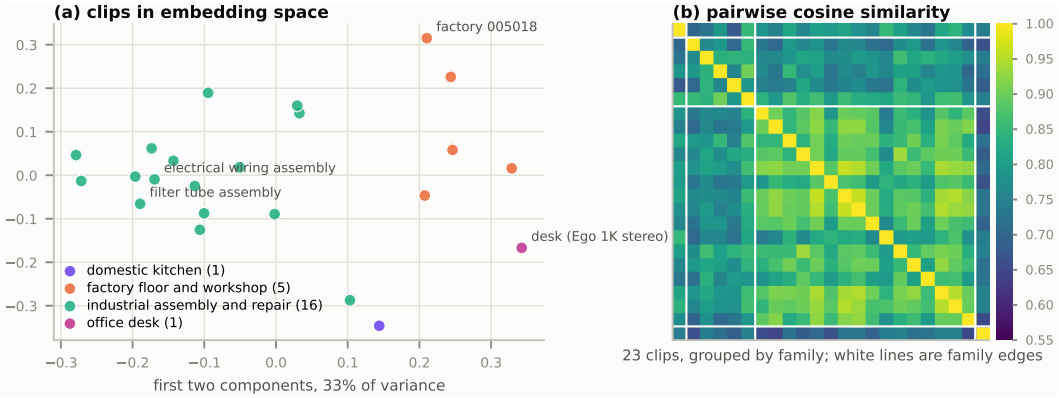

Figure 10: Diversity measured without labels, over the twenty-three clips held locally, ten frames each, mean-pooled per clip. **(a)** The first two components of the clip cloud, coloured by a coarse family tag, carrying a third of the variance; the two clips the embedding calls closest are named, as are the two furthest out. **(b)** Pairwise cosine similarity, clips grouped by family with white lines at the family edges. The large bright block is the industrial family, sixteen differently named tasks that the encoder finds largely interchangeable. Effective dimension of the whole set is about fourteen, against twenty-three clips and sixteen task names.

all defensible, all answering "how diverse is this set" differently, and a delivery that quotes one without saying which has told the buyer very little.

The pairwise view is sharper still. Mean cosine similarity between clips is about 0.82, and the closest pair sits at 0.96: two sequences named `electrical wiring assembly` and `filter tube assembly`. By name they are separate tasks and a label-based coverage count credits the set with two. By appearance they are nearly the same clip. Several other pairs in the same family behave the same way. This is the label-count problem measured from the other side, and it is the first thing we would use an embedding for in production, because near-duplicate detection has a checkable answer in a way that a diversity score does not. It also complements the content-hash deduplication of a delivery pipeline, which catches identical bytes and nothing else; deduplicating on embedding proximity instead is by now a standard move at web scale [1].

12.1 Clusters, and getting back to the footage

A diversity number is only auditable if you can go from a point in the space back to the frames it came from. Figure 11 does that: the clip cloud in three components, with four clips tied to the frames they were computed from.

Panel (a) colours the points by clusters found in the embedding itself, with nothing named and no label consulted. They are lopsided: one cluster holds twelve clips, another seven, and two are singletons. The callouts show why the singletons are singletons. One is a close top-down view of a press, the other a close view of a fan housing, both visually unlike the wide bench shots that dominate the rest of the set. That is a useful thing for a curator to be told, and no label in the delivery says it.

Panel (b) is the same points and the same projection, coloured instead by the environment recorded in the delivery. The two panels do not agree, and they are not supposed to. The delivered labels group by where the footage was shot; the embedding groups by what the frames look like, which is driven as much by camera distance and framing as by place. Sixteen clips share one delivered environment and the embedding splits them across several clusters. Neither grouping is wrong. They are different questions, and a delivery that reports one diversity figure has silently chosen which question it answered.

One negative result belongs here. Because this encoder has a text tower, we also tried assigning each clip an environment zero-shot, by embedding a short list of place phrases and taking the nearest. Fifteen of the twenty-three clips landed on the same phrase. We do not use that assignment for anything, and it is recorded as a caution: a zero-shot label from a retrieval encoder is not a classifier, and the fact that it produces a plausible-looking answer for every clip is exactly what makes it dangerous.

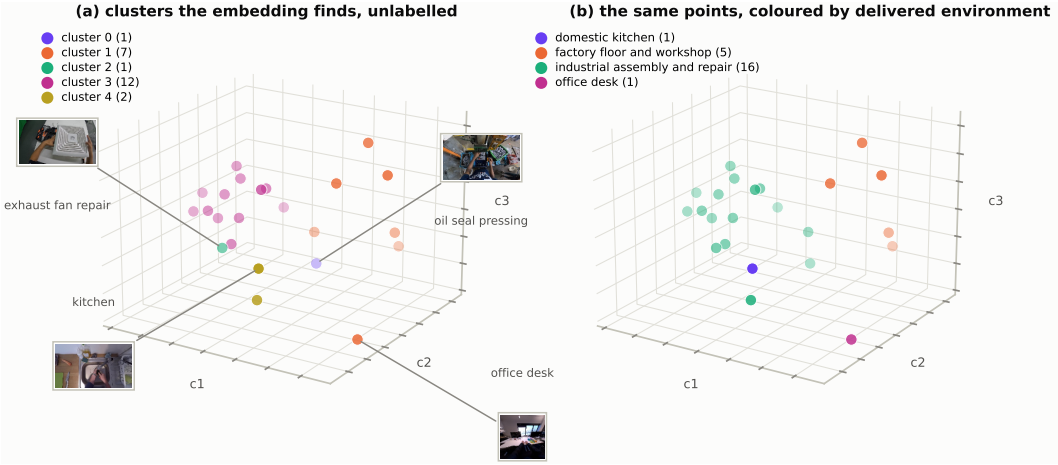

Figure 11: The clip cloud in three components, carrying 44% of the variance, with four clips tied to a representative frame each. **(a)** Clusters found in the embedding with no label consulted; the two singletons are close-range views unlike the wide bench shots that dominate. **(b)** The same points and projection, coloured by the environment recorded in the delivery. The two groupings cut across each other, because one asks where the footage was shot and the other asks what the frames look like.

What this measure is, and is not. Four limits, all visible in what was run. The encoder embeds *appearance*, so two clips in the same room doing different jobs sit close together while the same job in two rooms sits apart; it is not a measure of what is happening. Mean-pooling frames discards temporal structure entirely, so a static shot and a busy one can land in the same place, and a video encoder is the right instrument for any temporal notion of diversity. The two-dimensional projection in panel (a) carries only a third of the variance, so it is a picture and not the geometry. And the number depends on the encoder: a different backbone gives a different diversity, which means an embedding diversity figure is only meaningful with the encoder and its weights pinned, exactly like any other validator in Section 10.

Encoders worth trying, and what we have not run. We ran one image encoder because it was local and cheap, and it is the weakest reasonable choice. Self-supervised image features [20] would remove the language-supervision bias. Video encoders [29, 31] would restore the temporal axis that pooling threw away, and a predictive video model [2] would come closer still, since its latent is optimised for anticipating the next state rather than for matching text. The most interesting option for our purposes is a latent from a video world foundation model [18], because that latent is trained to *predict* rather than to retrieve, so distance in it is closer to the question a curator actually has, which is whether a clip would teach a model something it does not already know. Commercial video-embedding services exist and we have used one for retrieval over a curated set rather than for diversity scoring. None of these four has been run at scale here, and the comparison between them, scored against a downstream measure rather than against each other, is on the list in Section 14.

13 The Sim-to-Real Gap, and What Real Footage Is For

Everything above is a reason to collect and describe real video, so it is worth saying plainly what simulation already does well and where it does not reach. This section is a position rather than a measurement, and it is marked as one.

Simulators are strong where the physics is rigid and the geometry is authored. Kinematics, gross collision, articulated linkages and repeatable resets are cheap in simulation and expensive in the world, and domain randomisation has narrowed the purely visual half of the gap considerably [28]. The gap that has not closed is not mainly about how the pixels look.

Where the gap actually sits. Four things are hard to author and easy to record. *Contact and deformation*: friction, compliance, cloth, cable, food, granular material and the exact moment a grasp slips are the behaviours simulators approximate most crudely, and they are precisely the content of the footage in this note. *Clutter as it really occurs*: the tail of a real worktable, half-open drawers,

tools resting on other tools, is not a distribution anyone writes down; it is sampled. *Task structure*: the order in which a person actually does a job, the corrections and re-grasps, the dead time, is a prior about behaviour rather than about physics. And *sensing as it really degrades*: motion blur from a head turn, blown highlights from a window, a hand filling the frame. A simulator can be told to add noise; it cannot be told the joint distribution of these things without being shown.

Why this matters for learned simulators. The interest in video world models and world action models is that a learned simulator might capture what an authored one cannot, and inherit real contact statistics rather than a contact model somebody chose. But a learned simulator inherits the distribution it was trained on, exactly and without warning. Trained on synthetic rollouts it learns the synthetic dynamics, including their errors, and the sim-to-real gap moves inside the model where it is harder to see. Trained on real video it can learn real transitions, and then the binding constraint becomes whether the data carries the structure the objective needs, which is the whole subject of this note. A next-frame objective needs transitions that are real and non-trivial, object permanence across the window, recoverable camera geometry, and continuous timing. Those are the properties the layers above measure, and a corpus that has not been described cannot be selected for any of them.

What real egocentric footage supplies, and what it does not. It supplies the observation dynamics: how scenes actually evolve under human manipulation, what occlusion and disocclusion look like at close range, how objects persist and change state, and, where the capture is calibrated stereo, all of that in metres rather than in arbitrary units (Section 7). Those are the ingredients of a world model, and the evidence that they transfer is direct: representations pretrained on egocentric human video improve downstream robot manipulation [17], and action-free video pretraining followed by a comparatively small amount of robot trajectory data is the division of labour a recent video world model adopts [2]. What it does not supply is the action side. As Section 11 says, there are no robot-native actions, torques, forces or proprioception in this footage, so it cannot on its own train an action-conditioned simulator: it can teach what happens, and not what the robot did to make it happen. Closing that half needs data on the target embodiment, and the honest division of labour is that human egocentric video supplies the priors, the affordances and the dynamics of the world, while robot data supplies the mapping from command to consequence. Recent work takes exactly that shape, pairing egocentric video and 3D hand tracking with a smaller set of robot demonstrations rather than treating the human footage as though it contained robot actions [10].

That division is why curation matters more here than in a purely visual setting. If a learned simulator is to stand in for physics, the footage it learns from has to contain the physics of interest, at the unit the model consumes, with the geometry recoverable. Selecting that footage is not a preprocessing step; it is most of the work.

14 What Remains to Prove

This note shows what the layers emit and argues from mechanism about why each matters. It does not show that any curation decision built on them improves any model, and that gap is a research programme rather than a caveat. The studies that would close it, in order of cost: relate a pool's per-dimension composition to action-prediction accuracy for vision-language-action policies [11] and rollout error for video world models [3]; inject quantified defects at known severities, which is the design that can support a causal claim; ablate one criterion at a time, including the band-versus-gate choice; and test this footage against teleoperated demonstrations on the target embodiment [21]. Public egocentric corpora [7, 4, 8] are the natural setting for the labelled subsets several of these need, since the layers we cannot validate are those with no ground truth on our own footage.

15 Conclusion

The video is not the dataset. Between them sits a corpus described well enough to select from: hands resolved into a pose, objects located and named with the naming held at arm's length, identities carried across frames, contact resolved into episodes, and language attached to regions and intervals with its timebase and producer intact. Every one of those descriptions is also a claim that can be wrong, and the pictures here show both halves at once, which is the reason to look at output rather than at summary statistics.

So: describe before selecting, because the criteria worth using do not exist until something produces them. Trust per output, not per model. Say `unmeasured` when you mean it. Band the difficult footage instead of gating it away, because the occluded hand and the half-hidden tool are the job rather than the defect. And record the disposition and its reason at the unit where the decision was made, so the next version can disagree with this one on purpose.

References

[1] A. Abbas, K. Tirumala, D. Simig, S. Ganguli, and A. S. Morcos, "SemDeDup: Data-Efficient Learning at Web Scale Through Semantic Deduplication," *arXiv:2303.09540*, 2023.

[2] M. Assran *et al.*, "V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning," *arXiv:2506.09985*, 2025.

[3] J. Bruce *et al.*, "Genie: Generative Interactive Environments," in *ICML*, 2024.

[4] D. Damen *et al.*, "Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100," *IJCV*, vol. 130, pp. 33–55, 2022.

[5] A. Darkhalil, D. Shan, B. Zhu, J. Ma, A. Kar, R. Higgins, S. Fidler, D. Fouhey, and D. Damen, "EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations," in *NeurIPS*, 2022.

[6] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, "Datasheets for Datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.

[7] K. Grauman *et al.*, "Ego4D: Around the World in 3,000 Hours of Egocentric Video," in *CVPR*, 2022.

[8] K. Grauman *et al.*, "Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives," in *CVPR*, pp. 19383–19400, 2024.

[9] H. Hirschmüller, "Stereo Processing by Semiglobal Matching and Mutual Information," *IEEE TPAMI*, vol. 30, no. 2, pp. 328–341, 2008.

[10] S. Kareer, D. Patel, R. Punamiya, P. Mathur, S. Cheng, C. Wang, J. Hoffman, and D. Xu, "EgoMimic: Scaling Imitation Learning via Egocentric Video," *arXiv:2410.24221*, 2024.

[11] M. J. Kim *et al.*, "OpenVLA: An Open-Source Vision-Language-Action Model," in *Proc. 8th Conference on Robot Learning (CoRL)*, pp. 2679–2713, 2024.

[12] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-Captioning Events in Videos," in *ICCV*, pp. 706–715, 2017.

[13] J. Y. Lee *et al.*, "Ego-1K: A Large-Scale Multiview Video Dataset for Egocentric Vision," in *CVPR*, pp. 19854–19863, 2026.

[14] J. Lei, T. L. Berg, and M. Bansal, "Detecting Moments and Highlights in Videos via Natural Language Queries," in *NeurIPS*, 2021.

[15] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," in *ECCV*, 2014.

[16] S. Liu *et al.*, "Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection," in *ECCV*, 2024.

[17] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3M: A Universal Visual Representation for Robot Manipulation," in *Proc. Conference on Robot Learning*, PMLR vol. 205, pp. 892–909, 2023.

[18] NVIDIA, "Cosmos World Foundation Model Platform for Physical AI," *arXiv:2501.03575*, 2025.

[19] T. Ohkawa, K. He, F. Sener, T. Hodan, L. Tran, and C. Keskin, "AssemblyHands: Towards Egocentric Activity Understanding via 3D Hand Pose Estimation," in *CVPR*, pp. 12999–13008, 2023.

[20] M. Oquab *et al.*, "DINOv2: Learning Robust Visual Features without Supervision," *TMLR*, 2024.

[21] A. Padalkar *et al.*, "Open X-Embodiment: Robotic Learning Datasets and RT-X Models," in *ICRA*, 2024.

[22] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *ICML*, 2021.

[23] N. Ravi *et al.*, "SAM 2: Segment Anything in Images and Videos," 2024.

[24] A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko, "Object Hallucination in Image Captioning," in *EMNLP*, pp. 4035–4045, 2018.

[25] O. Roy and M. Vetterli, "The Effective Rank: A Measure of Effective Dimensionality," in *EUSIPCO*, 2007.

[26] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An Efficient Alternative to SIFT or SURF," in *ICCV*, 2011.

[27] D. Shan, J. Geng, M. Shu, and D. F. Fouhey, "Understanding Human Hands in Contact at Internet Scale," in *CVPR*, pp. 9869–9878, 2020.

[28] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World," in *IROS*, pp. 23–30, 2017.

[29] Z. Tong *et al.*, "VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training," in *NeurIPS*, 2022.

[30] A. Vakunov *et al.*, "MediaPipe Hands: On-device Real-time Hand Tracking," 2020.

[31] Y. Wang *et al.*, "InternVideo2: Scaling Foundation Models for Multimodal Video Understanding," in *ECCV*, 2024.

[32] B. Wen *et al.*, "FoundationStereo: Zero-Shot Stereo Matching," in *CVPR*, 2025.

[33] B. Xiao *et al.*, "Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks," in *CVPR*, 2024.

[34] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation," in *NeurIPS*, 2022.

[35] L. Yang *et al.*, "Depth Anything V2," in *NeurIPS*, 2024.
[36] L. Yu, H. Tan, M. Bansal, and T. L. Berg, "A Joint Speaker-Listener-Reinforcer Model for Referring Expressions," in *CVPR*, pp. 7282–7290, 2017.
[37] C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. Argus, and T. Brox, "FreiHAND: A Dataset for Markerless Capture of Hand Pose and Shape from Single RGB Images," in *ICCV*, 2019.