LabProc and Tacit: Quantifying the Visual–Textual Prior Gap in Autonomous Laboratory Perception

Anonymous Authors[1]

Abstract

Autonomous laboratory systems direct robotic platforms and execute multi-step procedures, but their perception layer is typically a frontier vision-language model (VLM) queried with sampled frames. Whether VLMs are the appropriate perception substrate for laboratory video is an open empirical question. We introduce **LabProc**, a benchmark for laboratory procedure understanding that organizes six tasks along a structural axis from single-frame state recognition to triplet anchor matching where all clips share the same nominal physical state, and **Tacit**, a 300M-parameter domain-adapted V-JEPA-2.1 video encoder we release as a vision-only baseline. Across this axis we observe a structural-axis gradient: Claude Opus leads Tacit by 41 points on single-frame state classification and by 6–16 points on tasks that retain language-amenable structure, but on the motion-only TED-Visual Strict Hard subset the gradient reverses and Tacit leads Claude by 9.1 points (66.7% vs. 57.6%) despite a $\sim$1000$\times$ parameter asymmetry. Tacit's continued pretraining required 28 minutes on a single H100 ($1.30 in compute) and improves over base V-JEPA-2.1 on every task with a mean uplift of $+8.2$ points. We also identify a representational tension central to laboratory self-supervised learning: state-invariance objectives (EMA target encoders combined with motion-conditioned masking) collapse within-state temporal physics, attenuating fine-grained motion-progression signals (Same-State CCR Kendall's τ drops from $+0.048$ at base to -0.062 at the released checkpoint). The released v1 dataset and Tacit checkpoint serve as a calibration target for future laboratory perception modules.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the AI for Science workshop (ICML 2026).

1. Introduction

Autonomous laboratory platforms (Coscientist, A-Lab, Kosmos) plan experiments, control instruments, and draft conclusions. Each step depends on a perception module that reads procedure state from video, and the dominant choice for that module is a frontier VLM such as Claude Opus or GPT-5.5, queried with a few sampled frames. VLMs are easy to deploy, require no domain adaptation, and read instructional captions and apparatus labels well. They also have a structural blind spot.

Consider a multi-step organic synthesis: distillation, extraction, recrystallization, then a reflux that must run to steady state before the next reagent is added. A perception module that can distinguish a Soxhlet from a fractional column, but cannot distinguish actively-boiling reflux from settled reflux, will trigger the next reagent injection prematurely and ruin a multi-day synthesis. The state labels of the two clips are identical (*reflux_running*); the equipment is identical; only the motion-state of the fluid distinguishes them. Claude Opus's pretraining gives it strong priors over the labels and the equipment but no priors at all over which fluid-state of identical equipment is depicted in five sampled frames.

We construct a six-task benchmark to surface this asymmetry, organized along a structural axis from *language-amenable* (single-frame state recognition) to *motion-only* (triplet anchor matching where the linguistic state of all candidate clips is identical). On this axis we report a structural-axis gradient: a frontier VLM dominates the language-amenable end by 41 points, the gap closes monotonically as we move toward motion-only, and on the structurally hardest motion-discrimination subset (*TED-Visual Strict Hard*, where anchor and distractor share equipment AND nominal state) the sign reverses. A 300M-parameter domain-adapted vision encoder leads Claude Opus by 9.1 points despite a $\sim$1000$\times$ parameter asymmetry. Reading the gradient: where pretraining provides the signal, pretraining wins; where motion-state is the only distinguishing feature, dense-temporal video representations win.

This paper releases two artifacts that make the gradient measurable. **LabProc** is a public dataset and benchmark for laboratory procedure understanding with six tasks: phys-

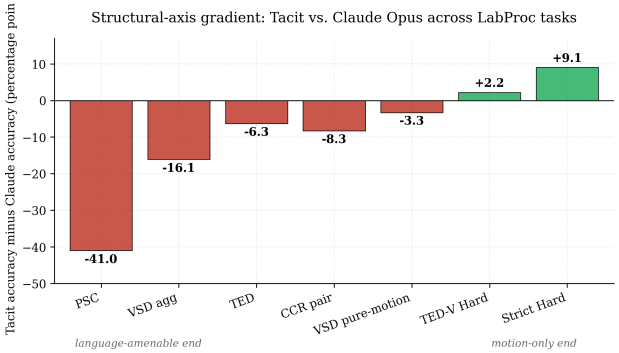

Figure 1. Structural-axis gradient. Tacit's accuracy minus Claude's, in percentage points, across LabProc tasks ordered from most language-amenable (left) to most motion-only (right). The gradient is monotonic and reverses sign on Strict Hard.

ical state classification, transition error detection, causal chain reconstruction, visual state discrimination, triplet anchor matching, and same-state ordering. **Tacit** is a domain-adapted V-JEPA-2.1 (Assran et al., 2025) encoder requiring 28 minutes on one H100 ($1.30) that serves as the v1 vision-only baseline and improves over base V-JEPA-2.1 on every LabProc task (mean uplift $+8.2$ points). Beyond these artifacts, we identify a representational tension that surfaces during Tacit's adaptation: the same EMA + motion-masking pipeline that produces the Strict Hard win attenuates within-state temporal coherence (Same-State CCR Kendall's τ drops from $+0.048$ at base to -0.062 at the released checkpoint). State-invariance objectives, in their current form, collapse the within-state physics that within-state-ordering tasks require. We document this as a fundamental constraint on self-supervised laboratory representation learning rather than a limitation specific to our run.

Scope. LabProc v1 covers 241 annotated clips from 38 organic-purification YouTube source videos. We do not claim Tacit replaces Claude Opus as a general-purpose laboratory perception layer: on aggregate accuracy, Claude leads. The claim is narrower: each of the two has a profile structured by its pretraining, the profiles are complementary, and an autonomous laboratory deployment optimizing for cost or for motion-only signal should consider a domain-adapted vision encoder rather than treating a VLM as the only option.

2. Related Work

Video understanding for science. Self-supervised video models (Assran et al., 2025; Tong et al., 2022; Radford et al., 2021) have made strong progress on general video understanding, but laboratory video has been studied primarily through small task-specific datasets focused on instrument-level outcomes (e.g., chromatography peak detection, microscopy segmentation) rather than procedure-level understanding. Boiko et al. (2023) and parallel autonomous-laboratory work treat perception as a black-box module supplied by a commercial VLM, with limited downstream evaluation of perception quality. LabProc is, to our knowledge, the first benchmark that targets procedure understanding (state recognition, ordering, transition error detection, motion discrimination) at the granularity needed by an AI scientist's perception layer.

Vision-language models as perception modules. Frontier VLMs (Claude Opus, GPT-5.5, Gemini) are competitive with specialized vision systems on many recognition tasks when the relevant signal is captured in a few sampled frames and the relevant categories appear in pretraining captions. A growing body of work (e.g., Assran et al., 2025) documents specific gaps where this strategy underperforms: dense temporal motion, fine-grained state transitions, and visually similar configurations that differ only in ongoing dynamics. Our benchmark deliberately surfaces this asymmetry by structuring tasks along a language-amenability axis and reporting per-task rather than aggregate accuracy.

Domain-adapted vision encoders. Continued self-supervised pretraining on domain video (Grill et al., 2020; Caron et al., 2021) has been shown to improve representation quality for narrow domains. The novel contribution here is not the adaptation method (we use a straightforward EMA + motion-conditioned masking recipe) but the empirical mapping of where modest-scale adaptation does and does not yield gains relative to a commercial VLM, on a benchmark designed to surface this asymmetry.

3. The LabProc Benchmark

LabProc v1 contains 241 four-second clips drawn from 38 public YouTube source videos covering organic-purification procedures (distillation, recrystallization, liquid-liquid extraction, column chromatography). Source videos pass a three-stage filtering pipeline (keyword pre-filter, three-signal automated quality control on CLIP semantic score, motion richness, and change density, then AI-assisted human curation against eight contamination categories). Annotation is single-annotator AI-assisted, producing 4-D dense labels per frame: physical state (25-class controlled vocabulary), substance properties ($\sim$40 tags), action ($\sim$25 tags), and equipment ($\sim$50 tags). The full pipeline, controlled vocabularies, and Claude annotation skill prompt are released alongside the data.

3.1. Six Tasks Along a Structural Axis

Each task isolates a different capability axis of a video model. We list them in order of *language-amenability*, i.e.,

from tasks for which Claude Opus has the most useful priors to tasks for which language priors are structurally insufficient.

PSC (Physical State Classification). 240 frame-level items, single-frame multi-class classification over 10 process-level physical states (e.g., *distillation_setup_running*, *lle_two_phase_settled*, *column_packed*). A representative still plus chemistry-textbook priors typically suffice. Random baseline: 10%.

TED (Transition Error Detection). 244 four-MCQ items asking which of four candidate physical states correctly follows a depicted transition. The visual+text variant uses a CLIP (Radford et al., 2021) text head over the four candidate state labels; the visual-only variant scores the candidate clips directly. Random baseline: 25%.

CCR (Causal Chain Reconstruction). 20 leave-one-group-out CCR groups; each group contains 3–5 clips from the same procedure family that must be ordered along the causal chain (e.g., *settled* → *draining* → *drained*). The probe is RankNet-style antisymmetric: $\text{score}(A, B) = g(A) - g(B)$ for a single shared projection head g. We report pairwise accuracy and Kendall's τ. Random baselines: 50% pairwise, $\tau = 0$.

VSD (Visual State Discrimination). 97 binary-discrimination items across 6 hardcoded pair types of visually similar states. Two of the six pairs are *pure-motion* (LLE settled vs. draining; column packed vs. equilibrated) where the equipment is identical and only the motion-state of fluid distinguishes the pair. Random baseline: 50%.

TED-Visual. 146 motion-triplet items: an anchor clip and two candidate clips, exactly one of which depicts the same physical state. Difficulty is determined by the visual relationship between the anchor and the distractor: *Easy* (different equipment), *Medium* (same equipment, different state), *Hard* (same equipment, similar state, $n = 46$), and *Strict Hard* (same equipment, same nominal state, distinguishable only by motion-state difference, $n = 33$). Strict Hard is the load-bearing structural test: any model that solves Strict Hard must encode motion-only signal. Random baseline: 50%.

Same-State CCR. 46 within-state ordering groups. Each group's clips share the same nominal state but differ along a motion-progression axis (e.g., five clips of column flow at progressing fill levels). Released as a task specification in v1; the released Tacit checkpoint is not the appropriate evaluation target (Section 6).

3.2. The Structural Axis

Read as a sequence, these tasks span from *maximally language-amenable* (PSC: a single frame plus textbook priors) through *partially language-amenable* (TED, CCR: state labels plus ordering) to *maximally motion-only* (VSD pure-motion subset, TED-Visual Strict Hard: where the linguistic state of the depicted clips is identical and only motion-state distinguishes them). This is the structural axis. The empirical question we ask is: how does the gap between Claude Opus and a small domain-adapted vision encoder change as we move along this axis?

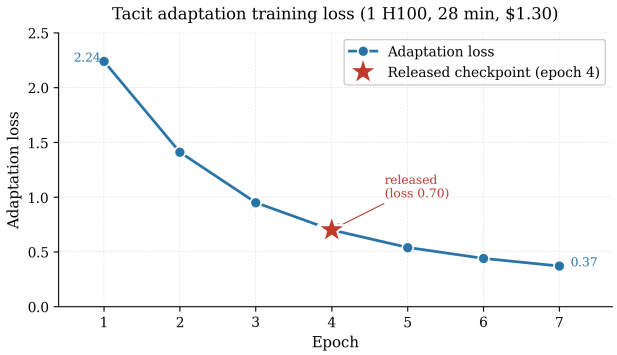

Figure 2. Tacit adaptation loss across 7 epochs. Monotonic decrease from 2.24 to 0.37 with no plateau. The released checkpoint is epoch 4 (loss 0.70).

4. Tacit: A Domain-Adapted V-JEPA-2.1 Encoder

Tacit is a continued-pretraining variant of V-JEPA-2.1 (Assran et al., 2025) ViT-L/14 (distilled from ViT-G at 384×384 resolution; 304.7M parameters). We unfreeze the last 3 of 24 transformer blocks (37.8M trainable parameters, 12.4%) and adapt for 7 epochs on a corpus of 159.8 hours of laboratory procedure video (1,037 videos spanning organic purification, polymerase chain reaction, and Western blot procedures, of which v1 LabProc annotates only the 38-video OP/YouTube subset).

Adaptation recipe. Self-supervised continued pretraining with two design choices:

- **EMA target encoder** (Grill et al., 2020; Caron et al., 2021): a copy of the student updated each step with momentum $\tau = 0.996$. The student predicts EMA-target features for masked patches.
- **Motion-conditioned masking**: 75% mask ratio with mask probability proportional to optical-flow magnitude, biasing learning toward motion-rich patches (Tong et al., 2022).

AdamW, learning rate 5×10^{-6} cosine, batch size 4, 64 frames per training clip (16 frames at inference), FP16. Total wall-clock: 28 minutes on a single H100 80GB. Total compute cost: $1.30 in rented compute.

Epoch selection. Training loss decreases monotonically across all 7 epochs (2.24 → 0.37; Appendix E, Table 4) but downstream LabProc performance is non-monotonic. We release **epoch 4** as the public Tacit checkpoint based on three criteria: (i) epoch 4 wins TED-Visual Strict Hard outright, with +3.1 points over the next-best epochs at 1 and 2 (+6.1 over base); (ii) epoch 4 ties for first place on VSD aggregate (57.8%), is within 0.5 points of the best epoch on PSC-10 (31.2% vs. 31.7% at epoch 6), and within 1.5 points on CCR pairwise (58.7% vs. 60.2% at epoch 6); (iii) epoch 4 is mid-training, providing a defensible position against either "stopped too early" or "overtrained" objections. The non-monotonicity on Strict Hard is the structurally important observation; the practical impact on language-amenable tasks is small.

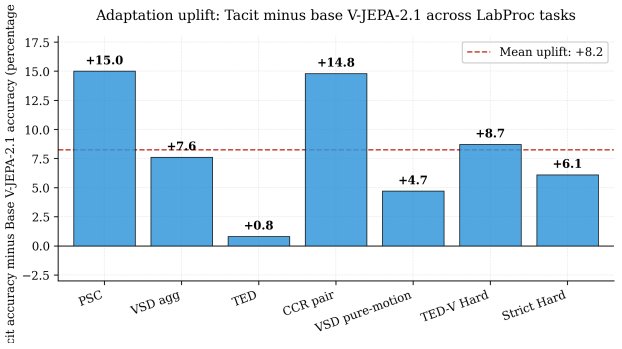

Figure 3. Tacit uplift over base V-JEPA-2.1 across LabProc tasks. Adaptation produces consistent positive uplift on every task with mean +8.2 points; largest absolute gains on PSC-10 (+15.0) and CCR pairwise (+14.8).

5. Results

We compare three conditions: (1) the unadapted V-JEPA-2.1 ViT-L/14 base, (2) Tacit (V-JEPA-2.1 ViT-L/14 + EMA + motion-mask, epoch 4), and (3) Claude Opus queried as a frontier vision-language baseline. For PSC, CCR, and VSD we train shallow probes on top of frozen features with 5-fold GroupKFold cross-validation by source video; for TED-Visual we train a 270K-parameter siamese projection head with triplet margin loss. Claude is queried via the public API with task-specific prompts; full evaluation protocol details are in the released code repository. Headline results are in Table 1.

5.1. Three Headline Findings

Finding 1: Adaptation produces consistent uplift across tasks. Tacit improves over its base V-JEPA-2.1 checkpoint on *every one* of the seven LabProc evaluation slots (six tasks; CCR is reported as both pairwise accuracy and Kendall's τ). The mean uplift across tasks is +8.2 percentage points (Figure 3). The largest absolute gains are on PSC-10 (+15.0) and CCR pairwise (+14.8); the smallest are on TED (+0.8, where the visual+text variant uses CLIP text features and is partly insulated from visual-feature changes) and VSD pure-motion (+4.7, where the absolute starting point is near chance and the adaptation budget is small relative to the task difficulty). This finding alone — that 28 minutes of single-H100 continued pretraining produces consistent positive uplift across a diverse 7-task benchmark — positions Tacit as a useful baseline for AI scientist deployments where a small domain-adapted encoder is preferable to running a 300B-parameter VLM in the loop.

Finding 2: The structural-axis gradient. Reading the rightmost column of Table 1: Tacit minus Claude monotonically increases as we move from language-amenable PSC (−41) through partially-language-amenable VSD aggregate, TED, CCR (−6 to −16) to the motion-only end of the structural axis. On VSD pure-motion the gap is −3.3, within evaluation-noise of zero on a 45-item subset. On TED-Visual Hard ($n = 46$) Tacit overtakes Claude by +2.2 points, and on TED-Visual Strict Hard ($n = 33$) Tacit leads Claude by +9.1 **points** (66.7% vs. 57.6%). Strict Hard is the structurally hardest motion-discrimination test in v1: anchor and distractor share equipment AND nominal state; only the motion-state of fluid (e.g., settled vs. draining) distinguishes them. This is where Claude's language priors stop helping. Figure 1 plots the same gradient.

Finding 3: Strict Hard is anchored by structural ground truth. A reasonable concern with finding 2 is that Strict Hard's labels might be artifacts of single-annotator bias: if our annotator agrees with Tacit and disagrees with Claude on what counts as "same state," the result reduces to annotator preference. We address this by construction: Strict Hard's ground truth is *structural*, not annotation-derived. The candidate state labels for the anchor and distractor are produced via a 4-D label intersection: anchor and distractor must share *physical state*, *equipment*, AND *substance tags*, leaving only motion-state as a distinguishing dimension. The design protects against the most concerning failure mode — that we built a benchmark on which our preferred model wins because we curated it to do so. The 9.1-point lead on Strict Hard is on items where structural ground truth, not annotator opinion, provides the labels.

6. Analysis

6.1. Why Does the Gradient Reverse?

The structural-axis gradient (Figure 1) follows from a simple decomposition of where each model's strengths come from. Frontier VLMs were pretrained on web-scale image-text pairs that include extensive chemistry-textbook and laboratory-tutorial content. When asked to recognize an

Task	Random	Base V-JEPA-2.1 ViT-L/14 (305M)	Tacit (ep4) ViT-L/14 (305M)	Claude Opus VLM (∼300B)	Tacit – Base uplift	Tacit – Claude gradient
PSC-10 (10-class)	10.0	16.2	31.2	**72.2**	+15.0	−41.0
TED visual+text (4-MCQ)	25.0	75.3	76.1	**82.4**	+0.8	−6.3
CCR pairwise	50.0	43.9	58.7	**67.0**	+14.8	−8.3
CCR Kendall's τ	0.000	−0.122	+0.175	**+0.359**	+0.297	−0.184
VSD aggregate (6 pairs)	50.0	50.2	57.8	**73.9**	+7.6	−16.1
VSD pure-motion (2 pairs)	50.0	50.0	54.7	**58.0**	+4.7	−3.3
TED-V Hard ($n = 46$)	50.0	60.9	**69.6**	67.4	+8.7	+2.2
TED-V Strict Hard ($n = 33$)	50.0	60.6	**66.7**	57.6	+6.1	**+9.1**
Mean uplift / gradient	—	—	—	—	**+8.2**	**−7.9**

Table 1. Headline LabProc v1 results. Tacit (epoch 4 of 7-epoch adaptation) improves over the V-JEPA-2.1 base on every task, with mean uplift +8.2 percentage points. Claude Opus leads Tacit on aggregate, but the gradient (Tacit minus Claude) reverses sign on the motion-only TED-Visual subsets. Bold indicates the per-task winner among Tacit and Claude.

apparatus or label a physical state from a single representative frame, Claude can match the image against linguistic priors and make a confident, often-correct call. Tacit was pretrained on web-scale general video and adapted on 159.8 hours of laboratory video; it has access to none of Claude's textbook priors but learns dense temporal representations of motion-state from its adaptation set.

The two failure modes are therefore complementary. Claude's failure mode on TED-Visual Strict Hard is structural: a 5-frame sample of two clips that are both at the "settled" state but in different motion-states cannot be distinguished by any feature linguistically tagged in pretraining. A query like "which clip matches the anchor's procedure stage?" is ill-posed for a model whose internal state is "these are both LLE settled." Tacit's failure mode on PSC is the inverse: a 16-frame clip without textbook priors is genuinely uninformative about whether the apparatus is a Soxhlet extractor versus a reflux setup if neither is dominant in the adaptation distribution. Where pretraining provides the load-bearing signal, that pretraining wins. The benchmark is structured to make this asymmetry visible.

6.2. State-Invariance Collapses Within-State Physics

The clips within a Same-State CCR group encode the temporal evolution of an underlying physical process: fluid flow through a packed bed (Darcy regime) in column chromatography, Stokes settling of a liquid-liquid emulsion in extraction, thermal relaxation toward steady-state reflux in distillation, or nucleation and crystal growth in recrystallization. Reading their order is reading that physics. A finding that surfaces during Tacit's adaptation has implications beyond our specific run. The unadapted V-JEPA-2.1 base achieves $\tau_{\text{neighbour}} = +0.048$ on Same-State CCR's antisymmetric probe; Tacit at epoch 4 produces $\tau = -0.062$. The trough deepens with continued adaptation: by epoch 7, τ recovers only to -0.038, never returning to the base rate. The released checkpoint is not the appropriate evaluation target for Same-State CCR.

The mechanism is not specific to Tacit's hyperparameters. EMA-based self-supervised adaptation drives the encoder toward features that are stable across views of the same scene, by definition. When the views are sampled within a single nominal state (a column at progressively higher fill levels, a reflux at progressively closer-to-steady), the EMA target encourages the student to map them to similar features — which is precisely the signal Same-State CCR's probe needs to read in order to order them. Motion-conditioned masking compounds the effect: by biasing learning toward patches with high optical-flow magnitude (operator hands, layer interfaces during pouring), the objective de-prioritizes patches whose temporal evolution is subtle (a column meniscus rising 2cm over 30 seconds). Both choices are deliberate and both improve cross-state recognition (PSC, TED, CCR pairwise) and coarse motion discrimination (Strict Hard). They also collapse within-state temporal physics.

We frame this as a fundamental constraint on self-supervised laboratory video representation learning, not a limitation of our adaptation budget. A perception module trained with state-invariance objectives cannot, in current form, simultaneously serve cross-state recognition and within-state ordering. AI scientist deployments that need both will require either separate encoders for each task (the design we lean toward in §7) or new objectives that preserve neighbour-similarity gradients across same-state clips. We identify this as the immediate next research target and release Same-State CCR as a v1 task specification to make the constraint measurable for future adaptation methods.

A separate operational consideration — the deployment-cost asymmetry between a commercial-VLM perception layer (polled at high frequency in a tight robotic loop) and a domain-adapted vision encoder running on a single GPU — favours hybrid architectures. We provide a representative cost decomposition in Appendix G, but emphasize that the scientific finding of this paper is the representational gap on

motion-only tasks; the operational implications follow from that gap rather than the other way around.

7. Discussion

7.1. Implications for AI Scientist Perception

The structural-axis gradient suggests that AI scientist systems should not treat "perception" as a single architectural decision. A laboratory perception layer needs to recognize equipment and procedural class (where Claude Opus excels), discriminate motion-only signal between visually similar configurations (where a domain-adapted vision encoder excels), and order temporally adjacent same-state clips (where neither system is currently sufficient). A pragmatic deployment combines:

- A commercial VLM consulted at low frequency for procedural-class identification, error detection, and language-amenable decisions.
- A domain-adapted vision encoder running at high frequency for motion-state discrimination and routine state-tracking.
- An explicit handoff protocol that escalates from the cheap perception module to the expensive one based on confidence or task type.

Treating the VLM as the only perception option, as is currently typical, both overpays for cheap decisions and underperforms on motion-only signal. LabProc's six-task profile is intended as a structural diagnostic for which slot a given model should fill in this composite design.

7.2. Limitations

Single-procedure-family v1 evaluation. LabProc v1 covers only organic purification on YouTube. PCR and Western blot annotation pipelines are constructed but not yet released; v2 will extend headline numbers to all three families.

Single-annotator AI-assisted ground truth. PSC, CCR, and VSD-aggregate labels are produced by a single human annotator with Claude Opus assistance. The headline VLM-vs-encoder claim (Tacit $+9.1$ on Strict Hard) is anchored on *structural* ground truth, not annotation-derived ground truth, which substantially reduces this concern for the load-bearing finding. We will release a multi-annotator stratified-sample re-annotation pass in v1.1.

Modest evaluation set sizes for load-bearing subsets. Strict Hard has $n = 33$, near the floor for pairwise-accuracy estimation. Reported gaps within 5–6 percentage points should be considered within evaluation noise. The 9.1-point Tacit-vs-Claude lead on Strict Hard is large relative to this floor; smaller gaps in this paper should not be over-interpreted.

No comparison to parameter-matched VLMs. We compare a 305M-parameter encoder against a $\sim$300B-parameter frontier VLM. This is the practically relevant comparison (the alternatives an AI scientist deployment is choosing between) but does not isolate the source of the gradient. Comparisons against parameter-matched open-weight VLMs (e.g., InternVL2-8B, Qwen-2.5-VL-7B) are deferred to v2.

Modest adaptation scale. Tacit's adaptation corpus, while substantial for laboratory video specifically, is small relative to general-purpose video pretraining (1M+ hours for V-JEPA-2.1). VSD pure-motion accuracy near chance (54.7% Tacit, 58.0% Claude) is consistent with the adaptation budget being insufficient for the hardest motion-discrimination cases. Scaling to thousands of hours is a v2 priority.

7.3. What This Paper Does Not Claim

We do not claim Tacit replaces frontier VLMs as a general-purpose perception layer for AI scientist systems. We do not claim the structural-axis gradient is universal across laboratory video distributions or generalizes to PCR/Western blot procedures (those are v2). We do not claim the EMA + motion-mask adaptation recipe is optimal — we use it because it works, not because we have ablated alternatives. The paper's claim is narrower and load-bearing: on the v1 LabProc benchmark, six tasks reveal a structural-axis gradient with a sign reversal on the motion-only subset, and a $1000\times$ smaller domain-adapted encoder leads a frontier VLM by 9.1 points on the structurally hardest motion-discrimination test.

8. Conclusion

LabProc and Tacit are released together as a v1 calibration target for AI scientist perception modules. The structural-axis gradient they reveal — Claude dominating language-amenable tasks, a domain-adapted vision encoder dominating motion-only tasks — is, we believe, the single most consequential finding for AI scientist deployments today: it suggests perception should be a hybrid, not a single-model decision. We hope LabProc and similar structurally-organized benchmarks become a standard step in AI-scientist-system evaluation, and we identify within-state-coherence-preserving adaptation as the immediate next research target.

Impact Statement

This work establishes a perceptual foundation for transitioning laboratory AI from passive compliance monitoring toward active autonomous execution. By quantifying the visual–textual prior gap, we show that current frontier

vision-language models, while semantically capable, lack the fine-grained temporal discrimination required to manage physical state-space transitions in a real-world lab. LabProc and Tacit provide a framework for building specialized perception layers that can verify protocol execution in real time, anticipate experimental failures before they occur, and serve as the closed-loop perceptual substrate for autonomous discovery platforms. Enabling AI to see the lab with physical rather than only linguistic fidelity is a prerequisite for reliable, scalable scientific automation.

The same automation potential carries societal considerations. Increased AI participation in laboratory workflows may displace technician roles whose value derives from procedure execution rather than experimental design; the appropriate response is a labor transition that retrains rather than discards skilled experimentalists, not an unmitigated rollout. Autonomous laboratory systems also operate in physical environments where perception failures can cause material damage, chemical exposure, or wasted reagent: the model card released alongside Tacit explicitly disclaims production deployment without additional validation, and we recommend a human-in-the-loop oversight layer at every confidence-triggered escalation point. We also recommend against repurposing the encoder for operator-behavior inference (e.g., dwell-time tracking, productivity scoring); the adaptation corpus was collected for procedure understanding, not workforce monitoring, and applying it to the latter would be both technically inappropriate and ethically problematic.

References

Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Komeili, M., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., Arnaud, S., Gejji, A., Martin, A., Robert Hogan, F., Dugas, D., Bojanowski, P., Khalidov, V., Labatut, P., Massa, F., Szafraniec, M., Krishnakumar, K., Li, Y., Ma, X., Chandar, S., Meier, F., LeCun, Y., Rabbat, M., and Ballas, N. V-JEPA-2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. URL https://arxiv.org/abs/2506.09985.

Boiko, D. A., MacKnight, R., Kline, B., and Gomes, G. Autonomous chemical research with large language models. *Nature*, 624:570–578, 2023.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z. D., Gheshlaghi Azar, M., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 220–229, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Tong, Z., Song, Y., Wang, J., and Wang, L. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

A. Datasheet for LabProc

This datasheet follows the *Datasheets for Datasets* framework (Gebru et al., 2021). A machine-readable Croissant JSON-LD file with Responsible AI metadata accompanies the dataset release.

A.1. Motivation

For what purpose was the dataset created? LabProc was created to enable evaluation of vision-only video models, vision-language models, and their composition on laboratory procedure understanding. A central design goal was to construct a benchmark whose tasks span a deliberate continuum from language-amenable static recognition to motion-only structural reasoning, so that single-task results do not confound vision-language models' strengths in textual reasoning with their weaknesses in fine-grained motion discrimination. No prior benchmark in this domain met this criterion: existing video understanding benchmarks (e.g., Kinetics, Something-Something) do not target laboratory procedures, and existing scientific video corpora (e.g., procedural cooking datasets) do not separate language-amenable and motion-only signal axes.

Who created the dataset and on behalf of which entity? The dataset was created by the paper's author as an independent research contribution. No institutional affiliation, grant, or sponsor was involved.

Who funded the creation of the dataset? No external funding supported this work. All compute, hosting, and API costs were paid by the author.

A.2. Composition

What do the instances represent? Two types of instances: (i) **video clips** from laboratory procedure videos, and (ii) **benchmark items** that reference one or more video clips and pose a structured task. The underlying corpus comprises laboratory procedure videos across three procedure families: organic purification (OP), polymerase chain reaction (PCR), and Western blot (WB).

v1 release scope. The v1 release is intentionally narrow. v1 contains 4-dimensional dense annotations for 241 OP clips drawn from 38 unique YouTube videos. PCR and WB videos exist in the corpus but are not annotated in v1 and are deferred to a v2 release. Source platforms beyond YouTube (PMC/JoVE, Figshare, Zenodo, Bilibili) are also deferred to v2. Annotations for the v1 release are released under CC BY 4.0 on Hugging Face at `https://huggingface.co/datasets/Labproc/labproc`; source video files are not redistributed.

How many instances are there in total? At v1 release: 241 OP clips with 4-D dense annotations, drawn from 38 YouTube source videos. Six benchmark tasks instantiate evaluation items from this annotated set: PSC-10 (240 items), TED (244 items), CCR (288 pair-instances drawn from 20 groups), VSD (six pair types, $\sim$97 items aggregate), and TED-Visual (Hard subset 46 triplets, Strict Hard subset 33 triplets). Same-State CCR is described in the companion paper Section 6 as a representational tension between state-invariance objectives and within-state temporal coherence; the released Tacit checkpoint exhibits this tension by design and is therefore not the appropriate evaluation target for Same-State CCR. The task specification is released as a calibration target for future adaptation methods that preserve neighbour-similarity gradients.

Does the dataset contain all possible instances or is it a sample? LabProc is a curated sample. Of all videos initially considered, 70.3% passed the three-stage automated quality filter, and 5.4% of those were rejected during AI-assisted human curation. The corpus does not claim to be statistically representative of all laboratory procedure videos on the public internet; it is a quality-controlled sample suitable for benchmark evaluation rather than a population estimate. Stratification details are provided in Section 3 of the main paper.

What data does each instance consist of? Source videos are MP4/WebM files with their original audio tracks. Annotations are released as CSV files with: per-frame 4-dimensional labels (controlled vocabulary: 58 states, 40 substances, 25 actions, 50 equipment items), audit trail JSON sidecars per frame, and per-task evaluation manifests linking benchmark items to their source clip timestamps. Source video files are not redistributed; the release contains URL manifests with the original public URLs and recovery instructions for failed downloads.

Is there a label or target associated with each instance? Yes. Each video frame carries a 4-dimensional structured label. Each benchmark item has a task-specific ground truth (state class, transition error label, ordering, correct continuation, etc.) as defined per task in the main paper.

Is any information missing from individual instances? A small fraction of source videos may become unavailable over time as upstream platforms (especially YouTube, Bilibili) remove or restrict access. We mitigate this by (i) recording per-video upstream URLs in the manifest, (ii) computing and recording perceptual hashes of each clip used in benchmark evaluation so that re-acquired video can be verified against the original, and (iii) committing to add new mirror URLs as they become available on a best-effort basis.

Are relationships between individual instances made explicit? Yes. Each benchmark item carries explicit references to the source video, source timestamps, and (for multi-clip tasks) the relationship type (anchor + candidates, ordered group, paired comparison).

Are there recommended data splits? LabProc v1 is released as an evaluation-only benchmark and does not contain a training split. All released items are intended as held-out evaluation. For probe training (used in CCR), per-task scripts implement leave-one-group-out cross-validation; this is encoded in the released evaluation harness rather than as static splits.

Are there errors, sources of noise, or redundancies? Yes; we document these explicitly. Annotation was performed by a single annotator with AI assistance, which carries the risk of systematic Claude-specific biases surviving human verification. We mitigate this by (i) restricting headline VLM-vs-video-model comparisons to tasks (TED-Visual Strict Hard, VSD pure-motion subset) whose ground truth is structural rather than annotation-derived, and (ii) embedding audit information inline in the master annotation CSV via three columns: `physical_state` (initial AI prediction), `confidence` (AI confidence at prediction time: low/medium/high), and `your_label` (final human-verified state after AI-assisted curation). Sidecar audit JSON files are not produced in v1; audit information is in-row. The full annotation skill prompt is also released for reproducibility and audit. Multi-annotator inter-annotator agreement on a stratified sample is identified as a v2 priority. Same-State CCR (described in the companion paper Section 6) revealed that the released Tacit checkpoint's adaptation pipeline attenuates within-state temporal coherence; this is documented prominently in the dataset card and the companion paper.

Does the dataset rely on external resources? Yes. v1 source videos are hosted on YouTube. Future evaluation requires re-acquiring these videos via the released URL manifest. Future releases will incorporate annotations from PMC/JoVE, Figshare, Zenodo, and Bilibili sources where collection is ongoing.

Does the dataset contain confidential or sensitive information? No. All source material is publicly accessible at time of collection. The dataset contains no personally identifying information beyond what is publicly visible in the source videos themselves (e.g., visible faces of demonstrators or hands of operators). No medical, financial, or protected category data is included or annotated.

Does the dataset identify any subpopulations? The dataset does not annotate or stratify by demographic attributes of video creators or demonstrators. Country of origin is not labeled; language of audio narration is not labeled (videos span nine or more languages but specific language tags are not part of v1 annotations).

Is it possible to identify individuals? Source videos may show faces, hands, voices, or workspaces of demonstrators. We do not perform face redaction in v1. All source content was originally published by its creators on public platforms; the dataset does not introduce new identifiability risk beyond what already exists at the source. Creators retain the ability to remove their content from upstream platforms, after which it is no longer available via the URL manifest.

A.3. Collection Process

How was the data acquired? Videos were collected via keyword search on YouTube using procedure-specific query terms (e.g., "recrystallization," "column chromatography," "liquid-liquid extraction"). v1 annotated set: 241 OP clips drawn from 38 unique YouTube videos. Search queries and result handling are documented in the released collection scripts.

Future v2 releases will expand to additional source platforms (PMC/JoVE, Figshare, Zenodo, Bilibili) and to PCR and Western blot procedure types.

What was the sampling strategy? A three-stage filtering pipeline was applied: (1) keyword-based pre-filtering during search; (2) automated quality assessment using a 3-signal scoring system combining CLIP-based content scoring, optical flow magnitude, and frame change density (threshold 0.30; 70.3% pass rate); (3) AI-assisted human curation of the survivors, with 5.4% rejected at this stage. The full filtering scripts and threshold rationale are included in the release.

Who was involved in the data collection process and how were they compensated? The author performed all collection, filtering, and annotation. No external annotators or contractors were involved. No compensation was paid because no external parties participated.

Over what timeframe was the data collected? Data collection ran in early 2026.

Were any ethical review processes conducted? No formal ethics review was conducted. The work uses only publicly available video content, does not involve human subjects research as defined by typical IRB criteria, and does not collect new data from individuals.

Did you collect the data from individuals directly or via third parties? Via a third party (YouTube). Future v2 releases will expand to additional public scientific video archives.

Were the individuals notified about the data collection? No. Source content was uploaded to public platforms by its creators with the expectation of public access. Creators were not individually contacted; the dataset relies on the public licensing terms of each source platform and the original creators' upload decisions.

Did the individuals consent to the collection and use of their data? Implicit consent through public posting under YouTube's terms of service. Creators retain the right to remove content from upstream platforms, after which it is no longer accessible via the LabProc URL manifest. We provide a takedown contact (Section A.6) and commit to honoring removal requests on a best-effort basis.

Has an analysis of the potential impact of the dataset on data subjects been conducted? A formal impact analysis was not conducted. Informally, the dataset poses limited additional privacy risk because (i) it contains no PII annotations and (ii) it does not redistribute the source video files themselves but only URLs and frame-level annotations. The marginal privacy harm of LabProc relative to the existing public availability of the source videos is small.

A.4. Preprocessing, Cleaning, Labeling

Was any preprocessing or cleaning of the data done? Yes. Three-stage filtering as described above. After acceptance, videos were normalized to a consistent codec/container and frame-rate-adjusted only when source rate fell below 24 fps. No spatial cropping, color correction, or audio normalization was applied. Frames used for benchmark evaluation are sampled at task-specific timestamps documented in the manifests.

Was the raw data saved in addition to the cleaned data? The release includes the URL manifest, perceptual hashes, and filtering scores at each stage. Raw upstream files are not redistributed; users re-acquire from upstream URLs.

Is the software used to preprocess and clean the data available? Yes. The release includes all preprocessing scripts (download, hash computation, three-stage filter, AI-assisted curation harness, annotation skill prompts) under the same license as the dataset.

A.5. Uses

Has the dataset been used for any tasks already? Yes; this paper itself constitutes the first use, evaluating base V-JEPA-2.1, the Tacit adapted checkpoint, and Claude Opus across the six benchmark tasks.

Is there a repository that links to any or all papers or systems that use the dataset? At release time: no. We will maintain a list of citing works on the dataset's Hugging Face card on a best-effort basis after publication.

What other tasks could the dataset be used for? Plausible future uses include: training data for procedure-specific video models, evaluation of multimodal LLMs on procedural reasoning, transfer learning to other scientific procedure domains, vision-language alignment studies, and audit material for AI-assisted laboratory automation systems.

Is there anything about the composition of the dataset that might impact future uses? Yes. (i) Single-annotator AI-assisted ground truth means downstream tasks should not assume multi-annotator labeled gold standard; we recommend treating annotations as one labeler's output rather than consensus. (ii) v1 source distribution is YouTube-only, biased toward English-language YouTube channels and toward content that is photogenic enough to be filmed and posted publicly. (iii) v1 covers organic purification only; PCR and Western blot annotations are deferred to v2. (iv) Audit information is embedded inline in the master annotation CSV (`physical_state`/`confidence`/`your_label` columns) rather than as separate sidecar files; tooling that expects per-frame audit JSONs will need to be adapted.

Are there tasks for which the dataset should not be used? The dataset should not be used to train or evaluate systems intended to autonomously execute laboratory procedures without human supervision; it is an understanding benchmark, not a control benchmark, and contains no safety annotations. The dataset should not be used to identify, profile, or contact the individuals visible in the source videos. The dataset should not be used to train face recognition, gait recognition, or any other person-identification system.

A.6. Distribution and Licensing

Will the dataset be distributed to third parties? Yes, as a public release.

How will the dataset be distributed? Two-part distribution: (i) Hugging Face dataset repository for working access via the `datasets` library and routine version updates; (ii) Zenodo archival snapshot with a permanent DOI for long-term reference. The Croissant JSON-LD metadata file with Responsible AI fields is included in both. The Tacit model checkpoint is released separately on Hugging Face with its own model card (Section B) and is also archived on Zenodo.

When will the dataset be distributed? At paper publication time (camera-ready release). A pre-release version is provided to workshop reviewers via anonymous links during the review period.

Will the dataset be distributed under a copyright license? The LabProc *annotations*, *evaluation manifests*, *filtering scripts*, and *evaluation harness* are released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. Source video files are *not* redistributed; their use is governed by YouTube's Terms of Service and the original creators' chosen licenses. The URL manifest is structured so that users can verify the upstream license of each video before use.

Have any third parties imposed IP-based or other restrictions? None beyond the upstream platform terms of service for the source video URLs.

Do any export controls or other regulatory restrictions apply? None known.

A.7. Maintenance

Who will be supporting / hosting / maintaining the dataset? The author. As an independent researcher, support is provided on a best-effort basis without a formal commitment of duration or staffing level. The dataset is hosted on Hugging Face at `https://huggingface.co/datasets/Labproc/labproc` and the released Tacit checkpoint at `https://huggingface.co/Labproc/tacit`.

How can the owner / curator / manager of the dataset be contacted? Via the email address listed on the dataset card. Issues, questions, and removal requests can be filed on the GitHub repository associated with the dataset (URL on the dataset card).

Is there an erratum? Errata will be tracked on the GitHub repository's issues page. Substantive errors will trigger a versioned re-release on Hugging Face with a corresponding new Zenodo DOI; the previous version remains accessible.

Will the dataset be updated? Yes, as a versioned series. v1 is the release accompanying this paper. Planned v2 directions include: PCR and Western blot annotated subsets, expansion to additional source platforms (PMC/JoVE, Figshare, Zenodo, Bilibili), full Same-State CCR evaluation under improved adaptation strategies (the v1 release does not include Same-State CCR as a benchmark task because the released Tacit checkpoint's adaptation pipeline attenuates within-state temporal coherence; this is documented in the companion paper Section 6), and a multi-annotator inter-annotator agreement study on a stratified sample. Maintenance commitment is best-effort; release of v2 is not guaranteed on a fixed timeline.

If the dataset relates to people, are there limits on the retention of data? Source video URLs may become unavailable as upstream creators remove their content; the manifest is updated to reflect such removals. The dataset itself does not store any user-specific or behavioral data beyond what is publicly visible in the source videos.

Will older versions of the dataset continue to be supported / hosted? Older versions are preserved on Zenodo with their original DOIs, ensuring scientific reproducibility of any work citing a specific version. Active issue triage and patch releases are focused on the current version.

If others want to extend / augment / build on / contribute to the dataset, is there a mechanism for them to do so? Yes. Contributions can be proposed via pull requests to the GitHub repository. Substantive proposed changes (new tasks, new annotations, new source platforms) will be evaluated on the merits and integrated into the next versioned release if accepted. The CC BY 4.0 license also permits independent forks and derivative works, provided attribution is preserved.

B. Model Card for Tacit

This model card follows the recommendations of Mitchell et al. (2019).

B.1. Model Details

- **Model name.** Tacit
- **Model type.** Self-supervised video encoder, domain-adapted from V-JEPA-2.1 ViT-L/14 (distilled from ViT-G at 384×384 resolution).
- **Architecture.** Vision transformer with 24 layers, 16 attention heads, hidden size 1024, MLP ratio 4. Patch size 16, image size 384×384, 64 frames per training clip with tubelet size 2. Uses RoPE positional encoding with interpolation, supporting variable frame counts at inference (we use 16 frames per clip at evaluation time).
- **Total parameters.** ∼300M (full encoder); 37.8M trainable during adaptation (12.4%, last 3 of 24 transformer blocks).
- **Output.** 1024-dimensional clip-level features after mean-pooling across spatiotemporal patch tokens.
- **Base model.** V-JEPA-2.1 ViT-L distilled from ViT-G at 384×384, released by Meta FAIR (Assran et al., 2025).
- **Adaptation method.** EMA target encoder (Grill et al., 2020; Caron et al., 2021) ($\tau = 0.996$) combined with motion-conditioned masking (Tong et al., 2022) (mask ratio 0.75).
- **Released checkpoint.** Epoch 4 of a 7-epoch adaptation run; training loss 0.70 (selection rationale in Section 4).
- **Release date.** v1, accompanying this paper.
- **License.** CC BY 4.0 (model weights and accompanying code).
- **Repository.** `https://huggingface.co/Labproc/tacit`.
- **Authors.** Anonymized for review.

B.2. Intended Use

Primary intended uses. Tacit is intended as a calibration target for laboratory video understanding research. Specifically:

- Producing frozen visual features for laboratory procedure clips, to be used as input to downstream linear or shallow probes.
- Comparison against future video encoders, larger adaptation runs, parameter-matched open-weight VLMs, and v2 PCR/Western blot benchmark instantiations.
- Reproducing the LabProc v1 benchmark numbers reported in this paper.

Primary intended users. Researchers in video representation learning, AI for laboratory science, and benchmark methodology.

Out-of-scope uses. Tacit is *not* intended for, and we explicitly do *not* recommend it for:

- Production deployment in laboratory safety, quality assurance, or regulatory compliance settings without substantial additional validation. The Strict Hard accuracy of 66.7% is a research-grade signal, not a deployment-grade reliability target.
- Use on non-laboratory video content. Adaptation on laboratory video specifically reshapes the representation toward this domain; performance on general video is not characterized and likely to be worse than the unadapted V-JEPA-2.1 base.
- Use as a foundation for behavioral or biometric inference from laboratory operator footage. The adaptation corpus deliberately downweights operator-specific signal, but operator presence is incidentally retained in the representation.
- Same-State CCR evaluation (within-state temporal ordering). The released checkpoint's adaptation pipeline attenuates within-state temporal coherence; users investigating this task should re-adapt with strategies that preserve neighbour-similarity gradients (Section 6).

B.3. Training Details

Training data. Adaptation corpus: laboratory procedure videos collected via the three-stage filtering pipeline described in Appendix C, spanning organic purification, polymerase chain reaction, and Western blot procedures. v1 LabProc benchmark evaluates only on the organic purification subset; the adaptation set spans all three branches.

Training procedure. Single-node, single-GPU continued pretraining with AdamW (learning rate 5×10^{-6}, cosine schedule, weight decay 0.01), batch size 4, 64 frames per training clip, 7 epochs. EMA target encoder updated each step with $\tau = 0.996$. Motion-conditioned masking with mask ratio 0.75 and motion-importance scores computed from optical flow magnitude. Mixed-precision (FP16) training. The last 3 of 24 transformer blocks are unfrozen; earlier blocks remain frozen.

Training hyperparameters.

- Optimizer: AdamW
- Learning rate: 5×10^{-6}, cosine schedule
- Weight decay: 0.01
- Batch size: 4
- Frames per clip (training): 64
- Frames per clip (inference): 16
- Mask ratio: 0.75
- EMA momentum: 0.996
- Epochs: 7 (release: epoch 4)
- Precision: FP16 mixed-precision
- Trainable parameters: 37.8M (last 3 of 24 transformer blocks)

Compute and environmental impact. Total adaptation: 28 minutes of single-H100-80GB wall-clock. Total compute cost: $1.30 (rented H100 at $2.79/hour). No multi-node training or hyperparameter sweep at this stage; the released configuration was selected from preliminary ablations rather than from a costly grid search. Adaptation was performed in a commercial cloud datacenter; using regional grid mix estimates ($\sim$0.4 kg CO_2e/kWh) and an H100 TDP of 700W, total carbon impact is approximately 0.13 kg CO_2e.

B.4. Evaluation

Benchmarks. LabProc v1 (six tasks; Section 3). Headline results in Table 1.

Decision threshold and probe configuration. Probes (PSC, CCR, VSD) are trained with GroupKFold cross-validation by source video to prevent leakage. PSC uses an MLP probe (LayerNorm + 1024$\to$256 + GELU + 256$\to n_{\text{classes}}$) with cross-entropy loss; CCR uses a pairwise MLP (concat of two clip features $\to$ 512 $\to$ 128 $\to$ 1) with BCE loss; VSD uses a binary classifier with the same shape as the PSC probe. TED-Visual uses a 270K-parameter siamese projection head with triplet margin loss.

Metrics and confidence intervals. Top-1 accuracy for PSC, TED, VSD; pairwise accuracy and Kendall's τ for CCR; 2-MCQ accuracy for TED-Visual. Per-fold standard deviations are reported in the per-epoch ablation tables (Appendix E). For load-bearing subsets where item counts are small (Strict Hard $n = 33$), reported gaps within 5–6 percentage points should be considered within noise.

B.5. Bias, Risks, and Limitations

Domain bias. The adaptation corpus is heavily skewed toward English-language YouTube laboratory procedure content. Common sub-genres (university teaching demonstrations, channel-specific recurring presenters, channel-specific recurring lab spaces) likely contribute systematic features that may not transfer to industrial laboratories, foreign-language workflows, or atypical equipment configurations.

Operator bias. Although adaptation downweights operator-specific signal via motion masking, operators (their hands, gloves, body positions, lab coats) are visible in nearly every frame. Tacit should not be used as input to systems that infer operator identity, behaviour, or performance.

Procedural-class bias. The 25-class OP physical state vocabulary represents one annotator's discretization of laboratory state space, biased by the channel mix in our training corpus. Some real-world states (hazardous, unconventional, or rare-substrate procedures) are underrepresented; some textbook states are oversampled. Downstream evaluations should not rely on Tacit's representation generalizing uniformly across the full chemistry-state distribution.

Adaptation-induced trade-off. As documented in Section 6, Tacit's adaptation attenuates within-state temporal coherence by an estimated $\sim$0.14 in $\tau_{\text{neighbour}}$ versus the V-JEPA-2.1 base. Users who require both within-state ordering and cross-state recognition will need a different adaptation strategy.

Single-annotator ground truth. LabProc v1 evaluation labels were produced by a single annotator with Claude-Opus assistance. While benchmarks deliberately favour structural ground truth (Strict Hard, VSD pure-motion) over annotation-derived ground truth for headline claims, downstream users should not interpret PSC, CCR, or VSD-aggregate Tacit accuracy as multi-annotator consensus.

Modest adaptation scale. Tacit's adaptation corpus, while substantial for laboratory video specifically, is small relative to general-purpose video pretraining (1M+ hours for V-JEPA-2.1). Pure-motion subset accuracy near chance (54.7% for Tacit, 58.0% for Claude) is consistent with the adaptation budget being insufficient for the hardest motion discrimination. Scaling to thousands of hours is identified as future work.

B.6. How to Use

The Tacit checkpoint and evaluation harness are released at:

- Model: https://huggingface.co/Labproc/tacit
- Dataset: https://huggingface.co/datasets/Labproc/labproc
- Code: https://github.com/tacit-anon/labproc

The encoder produces 1024-dimensional clip features. The accompanying evaluation harness includes user-facing scripts for each LabProc task. See the repository `README.md` for installation and quick-start instructions.

B.7. Citation

```
@inproceedings{labproc2026,
  title  = {LabProc and Tacit: Quantifying
            the Visual-Textual Prior Gap in
            Autonomous Laboratory Perception},
  author = {Anonymous},
  booktitle = {ICML 2026 Workshop on
               AI for Science},
  year   = {2026}
}
```

C. Corpus Filtering Pipeline

The LabProc source corpus is constructed from public laboratory procedure videos through a three-stage filtering pipeline. The pipeline is designed to remove off-task content (lectures, animations, vlogs), low-quality footage (talking-head channels, heavily-edited promotional material, near-static webcam shots), and domain contaminants (clinical analyzers, controlled-substance synthesis, school chemistry demonstrations) at three increasing levels of cost and specificity. We describe each stage below.

The pipeline was developed against candidates from five public source platforms (YouTube, PubMed Central Open Access including JoVE supplements, Figshare, Zenodo, and Bilibili) and three procedure branches (organic purification, polymerase chain reaction, Western blot). The v1 LabProc release annotates only the YouTube subset for organic purification (241 clips from 38 source videos); the underlying source-video set is broader. The pipeline scripts, keyword lists, CLIP prompts, threshold values, and contamination-detection prompt are released with the dataset to enable extension to additional procedure families and source platforms.

C.1. Stage 1: Keyword Pre-filter on Title and Description

Before any video is downloaded, candidate metadata (title, description, duration) is filtered against branch-specific keyword lists. The pre-filter removes the vast majority of unsuitable candidates without spending download bandwidth on them. Three filters are applied:

- **Duration bounds**, typically 60s–3600s with per-branch tuning. Western blot legitimately exceeds an hour for full procedure runs; PCR is typically shorter; OP varies. Bounds are tuned per branch to avoid clipping legitimate long-form content while excluding clearly-too-short shorts and clearly-too-long recorded lectures.
- **Reject keywords** (e.g., *animation*, *whiteboard*, *lecture*, *vlog*, *kids*, *home experiment*) that flag non-procedural content. The reject list was iteratively expanded during pilot annotation as new categories of false positives surfaced.
- **Quality keyword threshold.** A minimum number of branch-specific quality keywords (*protocol*, *tutorial*, *step-by-step*, *lab demonstration*, *procedure*, *technique*, etc.) must match. Bilibili candidates are filtered with a parallel Chinese-language keyword list constructed in collaboration with a native-Chinese-speaking colleague.

C.2. Stage 2: Three-Signal Automated QC on Downloaded Video

Each candidate that passes Stage 1 is downloaded at SD quality and scored on three orthogonal signals computed from the video pixels:

- **CLIP semantic score** (0–1): cosine similarity of mid-clip frames against a contrast set of laboratory-content prompts

(*laboratory bench*, *glassware on stand*, *test tube rack*) versus non-laboratory-content prompts (*office desk*, *kitchen counter*, *classroom whiteboard*), computed using OpenAI CLIP ViT-L/14 (Radford et al., 2021). We compute the score on three uniformly-sampled frames per video and average. A threshold of 0.45 was selected by visual inspection of pilot candidates near the boundary.

- **Motion richness**: mean optical flow magnitude (in pixels per frame) across uniformly-sampled 5-second segments. Indicates the presence of procedural activity rather than near-static talking-head footage. Computed using Lucas-Kanade flow on grayscale frames at 320×240 resolution.
- **Change density**: count of significant frame-to-frame structural changes per minute, computed as the number of frames whose perceptual hash differs from the previous frame's hash by more than a threshold. Indicates procedural progression rather than long static shots of equipment.

The three signals are combined via a learned threshold: a candidate is retained if all three are above per-branch minima determined from a 200-video pilot calibration set. The thresholds and calibration set are released with the pipeline scripts.

C.3. Stage 3: AI-Assisted Human Curation on Titles and Descriptions

After Stages 1–2, the surviving corpus contained domain contaminants that pass both pixel-level and keyword filters: clinical analyzers misclassified as PCR, controlled-substance synthesis content, basic-school chemistry demonstrations, equipment unboxing videos, and laboratory-themed marketing material. We resolved these by running an AI-assisted classification pass on the title and description of each surviving candidate, using Claude Opus with a contamination-detection prompt that flags eight failure modes (clinical analyzer, controlled-substance content, school-level demonstration, marketing/unboxing, simulation/animation, vlog/lifestyle, recorded lecture, off-topic). Each flagged candidate was manually reviewed for retention or rejection.

The AI-assisted pass and the contamination-detection prompt used by Claude are released alongside the dataset.

Final corpus. The full pipeline yields a 1,037-video corpus across all three branches and five source platforms. v1 LabProc benchmark annotates 241 clips from 38 YouTube videos in the organic purification branch (a subset of the broader corpus).

D. Annotation Methodology

AI-in-the-loop framing. The annotation methodology described below is itself an example of an AI Scientist component operating at the tool end of the spectrum: Claude Opus, prompted with a structured 8-section annotation skill, proposes per-frame labels that a human annotator accepts, modifies, or rejects. The setup is in some sense the inverse of the rest of this paper, which evaluates Claude Opus as an autonomous perception module; here Claude Opus is a labeling co-author whose suggestions are reviewed before adoption. We document the methodology in full so that future AI4Science work studying AI-assisted scientific labeling, attribution of dataset provenance, or co-authorship in dataset construction has a concrete reference point.

This appendix documents the full annotation methodology used to produce LabProc v1's 4-D dense annotations. The methodology was developed for the broader corpus (three procedure branches across five source platforms); v1 evaluation labels are the subset corresponding to organic purification on YouTube.

D.1. Single-Annotator AI-Assisted Protocol

LabProc v1 annotations were produced by a single annotator with Claude-Opus assistance. We adopt this protocol over a multi-annotator crowd-sourcing approach because the per-frame decision involves substantial domain knowledge (chemistry-specific physical states, equipment recognition, substance-property inference) that crowdsourced annotators typically lack, and because the structural-axis benchmark design intentionally restricts headline VLM-vs.-video-model claims to tasks (TED-Visual Strict Hard, VSD pure-motion subset) whose ground truth is structural rather than annotation-derived. We discuss residual risks of single-annotator ground truth in Section 7 and identify multi-annotator inter-annotator agreement on a stratified sample as a v2 priority.

The AI-assisted component uses Claude Opus to propose initial labels, which the annotator either accepts, modifies, or skips. The annotation skill prompt used by Claude — a structured 8-section prompt that defines the controlled vocabularies, the equipment-scan-first decision sequence, and the skip rules — is released alongside the dataset.

D.2. Six-Stage Annotation Pipeline

Stage 1 — Source curation. Videos that pass the corpus filtering pipeline (Appendix C) are pre-classified at annotation ingestion into three quality tiers (*good_yt*, *ok_yt*, *bad_yt*) based on a manual viewing of a 30-second sample from each video. Only *good_yt* enters annotation. This is an empirically-motivated cutoff: lower tiers yielded fewer than 5% labelable frames in pilot annotation, and the per-frame review labor on those tiers was not justified by the marginal data gained.

Stage 2 — Dense triage. For each candidate video, ten evenly-spaced frames are extracted at $i \cdot \text{duration}/11$ for $i \in \{1, \ldots, 10\}$, covering 9–91% of the timeline. Each frame is reviewed against two questions:

1. **Equipment scan**: does the apparatus visible in the frame match a procedure family in our taxonomy?
2. **Substance-state scan**: is procedural content visible (substance present in vessel, state distinguishable), or is the frame setup-only / talking-head / off-procedure?

Each frame is assigned one of three triage outcomes: *MATCH* (proceeds to dense extraction), *TAXONOMY-GAP* (apparatus visible but state not in current taxonomy — proposes a new label for review), or *NON-CATEGORY-SKIP* (rejected from this branch).

The dense-triage finding. In pilot work, we measured a *single-frame triage false-skip rate of approximately 75%*: videos whose first reviewed frame was setup-only or talking-head were rejected even when later frames contained substantial procedural content. Dense triage — ten frames covering nearly the full timeline — reduced this false-skip rate to under 5%. This is the most consequential single decision in our annotation methodology: dense triage is what makes YouTube a viable source of training-grade procedural data, rather than a sea of false negatives.

Stage 3 — Dense frame extraction. Videos that pass triage are densely sampled at 30-second intervals for full annotation. We extract frames using `ffmpeg` with the `-ss` seek flag rather than the `fps=` filter, which we observed drifts and corrupts timestamps over long videos. Native resolution is preserved; no upscaling is applied, because interpolation artifacts mislead downstream physical-state discrimination.

Stage 4 — Hierarchical 4-D labeling. Each extracted frame receives four orthogonal labels drawn from controlled vocabularies (Table 2). The primary state label ($\mathcal{S}$) is what becomes the PSC class label for benchmark items; the other three dimensions are released as auxiliary metadata to support future task design.

Dimension	Vocabulary size	Captures
$\mathcal{S}$ `physical_state`	25 canonical states	What physical state the substance is in
$\mathcal{B}$ `substance_tags`	∼40 tags	Phase, color, opacity, composition
$\mathcal{A}$ `action_tags`	∼25 tags	The operator's current physical action
$\mathcal{E}$ `equipment_tags`	∼50 tags	Apparatus visible in the frame

Table 2. The four label dimensions. The full controlled vocabulary lists are released with the dataset.

Anti-anchoring discipline. A naive single-pass labeler tends to anchor on the first plausible label that comes to mind. To counter this, we enforce an explicit two-step decision sequence per frame: (i) the equipment-scan eliminates label groups by *absence* of apparatus before any state label is considered; (ii) the substance-state scan picks among the surviving candidates. This protocol turns equipment recognition into a hard filter rather than a confirmatory cue, and is the structural reason our 20–40% intentional skip rate is signal-of-honest-labeling rather than evidence of laziness: a frame that contains a Soxhlet extractor and an ambiguous liquid state is correctly labeled with the state, but a frame containing a Soxhlet extractor and an indistinguishable substance is correctly skipped.

Stage 5 — Inline audit trail. For each annotated frame, three columns in the master annotation CSV record the audit trail: `physical_state` (the AI-proposed initial label), `confidence` (the AI's confidence at proposal time, in {*low, medium, high*}), and `your_label` (the final human-verified label after AI-assisted curation). When the AI proposal and human label differ, the row preserves both for downstream verification. This inline-column scheme replaces the sidecar JSON design from earlier annotation sessions and is the v1 release format.

Stage 6 — Bundle compilation. Annotations are emitted under a strict filename contract and a fixed CSV schema. Deduplication runs against the (branch, video_file, timestamp) triple. The cross-session deduplication invariant has preserved corpus integrity across multiple annotation sessions with no observed data-loss incidents.

D.3. Empirical Yield Calibration

The pipeline is calibrated against a per-archetype yield prior, measured during pilot annotation. These priors set realistic batch-size expectations and double as anomaly detectors: a video whose yield falls substantially outside the prior triggers review of the labeler session and the taxonomy.

Video archetype	Labeled-frame yield	Rows / video (30s sampling, 5–10 min)
Bench procedure (full workflow)	30–45%	5–8
RT-PCR / sample prep	50–55%	8–10
Educational / promo (talking-head)	$\sim$10%	1–2
Out-of-category (mis-categorized)	0% (skip)	0

Table 3. Per-archetype yield priors. A bench-procedure video yielding $<$15% triggers session-level review; deviation is more likely to indicate labeler fatigue or a taxonomy gap than genuine archetype change.

D.4. From 4-D Labels to PSC Items

The PSC task labels (Section 3) are the `your_label` primary dimension of the 4-D annotation; the remaining three dimensions are released as auxiliary metadata. The OP v1 PSC release contains 240 frame-level items spanning 38 unique source videos, after dropping items whose physical state is not in the 10-class process-level taxonomy.

We do not yet construct dedicated task data from `substance_tags`, `action_tags`, or `equipment_tags`, but we release them in full so that future LabProc versions can support tasks framed around those dimensions (e.g., substance-conditioned action prediction, equipment-conditioned state recognition) without re-annotation cost.

D.5. Reporting Agreement

Because a single annotator was used, we do not report Cohen's κ or analogous multi-annotator statistics. Instead, the inline-audit-trail design enables an unusually direct form of post-hoc verification: a second annotator can replay the per-row `physical_state` / `confidence` / `your_label` columns and verify, in approximately 15 seconds per item, whether the AI's initial proposal was reasonable and whether the final human label is justified. We commit to producing this independent re-annotation pass on a stratified sample of OP items as a v2 release.

Methodological gaps acknowledged. We flag three open methodological gaps in this v1 release:

1. Inter-annotator agreement is not yet measured.
2. Confidence-tier calibration is heuristic and has not been validated against downstream model error patterns.
3. Skip decisions are not labeled, so the $\sim$25–40% of frames marked as skipped are released as unlabeled negatives without a typed rejection reason.

We treat each as a v2 priority. Section 7 discusses the residual risk that systematic labeler-specific biases survived single-annotator review; we mitigate this by (i) releasing the full annotation skill prompt, (ii) releasing the controlled vocabularies in full, and (iii) restricting the headline VLM-vs.-video-model comparisons to tasks (VSD pure-motion subset, TED-Visual Strict Hard) whose ground truth is structural rather than annotation-derived.

E. Per-Epoch Ablation

Tacit's adaptation runs for 7 epochs on a single H100. The released checkpoint is epoch 4. This appendix reports per-epoch downstream sensitivity across all six LabProc tasks plus the adaptation loss curve. All numbers are extracted directly from the per-epoch JSON results files released in the supplementary materials.

E.1. Adaptation Loss Trajectory

Training loss decreases monotonically across all 7 epochs with no plateau (Table 4, Figure 2). The absence of a plateau on the adaptation loss does not, however, translate to monotonic gains on downstream LabProc tasks; we report the per-epoch downstream sensitivity in the next subsection.

Epoch	1	2	3	**4 (released)**	5	6	7
Adaptation loss	2.24	1.41	0.95	**0.70**	0.54	0.44	0.37

Table 4. Tacit per-epoch adaptation loss (mean over training batches in the final 10% of each epoch). The released checkpoint corresponds to epoch 4.

E.2. Per-Epoch Downstream Accuracy

Table 5 reports accuracy across all 7 epochs plus the unadapted base, on each LabProc task. Bold values mark the released-checkpoint column (epoch 4); per-task best epochs are summarized in the per-task subsections that follow.

Task	**Random**	**Base**	**ep1**	**ep2**	**ep3**	**ep4**	**ep5**	**ep6**	**ep7**
PSC-10	10.0	16.2	29.6	30.8	30.4	**31.2**	30.8	31.7	30.8
PSC-20	5.0	8.0	24.4	24.8	24.8	**25.6**	25.6	26.5	25.6
PSC-23	4.3	7.6	22.3	20.6	23.5	**20.6**	21.0	23.1	19.7
TED visual+text	25.0	75.3	73.3	74.5	77.0	**76.1**	77.0	75.3	77.4
TED visual-only	25.0	35.4	35.0	33.7	34.2	**34.6**	35.8	33.3	36.2
CCR pairwise	50.0	43.9	59.3	59.9	60.0	**58.7**	59.7	60.2	60.1
CCR Kendall's τ	0.000	-0.122	$+0.185$	$+0.198$	$+0.200$	$+$**0.175**	$+0.194$	$+0.203$	$+0.202$
VSD aggregate	50.0	50.2	57.7	57.8	56.9	**57.8**	56.6	56.9	56.4
TED-V Hard	50.0	60.9	63.0	65.2	60.9	**69.6**	63.0	63.0	60.9
TED-V Strict Hard	50.0	60.6	63.6	63.6	60.6	**66.7**	60.6	60.6	57.6
Same-State CCR τ	0.000	$+0.048$	-0.067	-0.058	-0.062	$-$**0.062**	-0.046	-0.039	-0.038
Same-State CCR pair	50.0	52.4	46.7	47.1	46.9	**46.9**	47.7	48.0	48.1

Table 5. Per-epoch accuracy (%) and rank correlation per LabProc task. *Base* = unadapted V-JEPA-2.1 ViT-L/14 at 384×384. The released Tacit checkpoint corresponds to **epoch 4**; bolded values are at the released checkpoint, not the per-task best epoch. CCR uses the antisymmetric RankNet-style probe over 20 leave-one-group-out folds (Section 3); 3 seeds per fold. Same-State CCR uses the antisymmetric probe over 46 leave-one-group-out folds, 3 seeds per fold.

E.3. Headline Per-Epoch Findings

The non-monotonicity finding. Adaptation produces a large jump on PSC-10 at epoch 1 ($+13.4$ points over base), then plateaus across epochs 1–7 within a 2.1-point spread on a 240-item evaluation. CCR pairwise accuracy follows a similar pattern: a large jump from base to epoch 1 ($+15.4$ points), then a tight 1.5-point spread across all subsequent epochs (58.7% to 60.2%). VSD aggregate plateaus from epoch 1 onward in a 1.4-point window. TED-Visual Strict Hard, in contrast, peaks specifically at epoch 4 ($+3.1$ over the next-best epochs at 1 and 2; $+6.1$ over base; $+9.1$ over the worst epoch at 7). This non-monotonicity on the load-bearing motion-discrimination test is the structurally important per-epoch observation in this paper.

The Same-State CCR adaptation trade-off. Same-State CCR Kendall's τ peaks at **base** ($+0.048$) and turns negative at every adapted epoch. The trough is at epochs 3–4 (-0.062) and shows a slow recovery toward zero by epoch 7 (-0.038) without ever returning to positive. Same-State CCR pairwise accuracy follows the same pattern: 52.4% at base, dropping to 46.7–48.1% across all adapted epochs. This is the cleanest single piece of evidence in our results that EMA + motion-conditioned-masking adaptation, as configured for Tacit, reduces within-state temporal coherence in the representation. The released Tacit checkpoint is *not* an appropriate evaluation target for Same-State CCR; we identify within-state-preserving adaptation as a v2 priority (Section 6).

Why epoch 4 is the released checkpoint. We select epoch 4 as the released Tacit checkpoint based on the following reasoning, repeated from Section 4 with per-epoch data context:

1. Epoch 4 wins TED-Visual Hard outright by $+4.4$ points over the next-best epoch (epoch 2 at 65.2%) and TED-Visual Strict Hard outright by $+3.1$ points over the next-best epochs at 1 and 2 (both at 63.6%). Strict Hard is the load-bearing structural-axis test; this is the only LabProc task for which epoch selection meaningfully changes the headline number.
2. Epoch 4 ties for first place on VSD aggregate (57.8%, tied with epoch 2) and is within 0.5 points of the best epoch on PSC-10 (31.2% vs. 31.7% at epoch 6) and within 1.5 points on CCR pairwise (58.7% vs. 60.2% at epoch 6). On all language-amenable tasks, the choice between epoch 4 and the per-task best epoch is within probe-noise variance.
3. Epoch 4 trails best on TED visual+text by 1.3 points (76.1% vs. 77.4% at epoch 7), within probe-noise variance for the 244-item evaluation.
4. Epoch 4 is mid-training rather than at either extreme, providing a defensible position against either "stopped too early" or "overtrained" objections.

The non-monotonicity on Strict Hard is the structurally important observation; the practical impact on language-amenable tasks is small.

E.4. Per-Pair VSD Sensitivity

The 6 VSD pairs differ markedly in difficulty. Table 6 reports per-pair accuracy across epochs. The two pure-motion pairs (LLE drainage and column flow) exhibit the smallest accuracy across all conditions, consistent with these pairs being the structural test of motion-only signal in VSD.

Pair	n	Base	ep1	ep2	ep3	ep4	ep5	ep6	ep7
LLE settled vs. draining†	31	48.4	54.8	54.8	54.8	54.8	54.8	51.6	51.6
Column packed vs. equilibrated†	14	64.3	42.9	42.9	42.9	42.9	42.9	42.9	42.9
Mixture dissolved vs. reflux	32	46.9	62.5	62.5	62.5	68.8	65.6	65.6	62.5
Fractions collecting vs. analysis	18	55.6	72.2	77.8	72.2	66.7	66.7	66.7	66.7
Crystals forming vs. complete	23	60.9	73.9	73.9	73.9	73.9	69.6	69.6	69.6
Distillation vs. reflux	20	25.0	40.0	35.0	35.0	40.0	40.0	45.0	45.0
Aggregate (mean over 6 pairs)	—	50.2	57.7	57.8	56.9	57.8	56.6	56.9	56.4

Table 6. VSD per-pair accuracy across all epochs. † marks the two pure-motion pairs (where equipment recognition is structurally insufficient and only motion-state distinguishes the pair). Aggregate is the unweighted mean across the 6 pairs.

F. Evaluation Protocol

This appendix documents the evaluation protocols used for each LabProc task and for the Claude Opus baseline.

F.1. Probe Architectures and Cross-Validation

For tasks evaluated by linear or shallow probing on top of frozen video features (PSC, CCR, VSD, TED-Visual), we train task-specific probes with GroupKFold cross-validation. The source video is the group: all clips originating from the same source video are placed in the same fold, ensuring probe evaluation tests generalization to new videos rather than memorization of within-video cues (camera angle, bench layout, lighting, operator). On preliminary experiments, naive K-fold splitting (clip-level random) inflated PSC accuracy by 10–15 percentage points relative to GroupKFold; we report only GroupKFold numbers throughout the paper.

The same probe architecture is used across all encoder conditions compared (V-JEPA-2.1 base and Tacit), so probe-architecture choice cannot account for cross-encoder differences.

PSC probe. A 2-layer MLP (LayerNorm + 1024→256 + GELU + Dropout(0.1) + 256→ N_{classes}) trained with cross-entropy. AdamW, learning rate 1×10^{-3}, weight decay 0.01, 100 epochs, batch full-fold. 5-fold GroupKFold (reduced to fewer folds when the number of source videos is below 5).

CCR probe. A pairwise concat MLP (LayerNorm(2×1024) + Linear($2048 \to 512$) + GELU + Dropout + Linear($512 \to 128$) + GELU + Dropout + Linear($128 \to 1$)) trained with binary cross-entropy on randomly-oriented within-group pairs. AdamW, learning rate 1×10^{-3}, weight decay 0.01, 100 epochs. Leave-one-video-out across the 20 CCR groups; each

held-out group's $N(N-1)/2$ pairs are evaluated. Reports Kendall's τ over predicted ordering and pairwise accuracy on held-out pairs.

VSD binary probe. A 2-layer MLP with the same shape as the PSC probe, trained per pair-type with cross-entropy on the binary discrimination. 3-fold GroupKFold by source video. Reports per-pair-type accuracy and aggregate mean.

TED-Visual siamese probe. A single shared projection head (LayerNorm + 1024→256 + GELU + Dropout(0.1) + 256→128, output L2-normalized) applied independently to anchor, correct, and distractor. Trained with triplet margin loss (margin 0.2) for 200 epochs. AdamW, learning rate 1×10^{-3}. 5-fold GroupKFold by anchor-video. The siamese architecture is symmetric by construction, eliminating positional shortcuts that concat-based pairwise classifiers can exploit when the per-fold training set is small.

TED. Two probe variants: a visual-only PSC-style classifier scored on the 4 candidate option labels at evaluation time, and a visual+text classifier that combines V-JEPA visual features with CLIP text features for the question and 4 options through a cross-attended scorer. Reports both numbers separately. The visual+text variant requires CLIP ViT-L/14.

F.2. Claude Opus Baseline

The Claude Opus baseline is queried via the Anthropic API with task-specific prompts that present representative video frames as images. Specifically:

- **PSC, TED, VSD, TED-Visual.** Five evenly-spaced frames from the 4-second clip are sampled and presented as a 5-image input. The prompt asks Claude to choose from the candidate states/options/clips. Each query is run once with default sampling and a constrained-output instruction; we do not aggregate over multiple samples.
- **CCR.** The N clips' representative frames (one per clip) are presented as a multi-image input. The prompt asks Claude to predict the temporal order; the output is parsed into a permutation and scored against ground truth.

The full Claude prompts are released alongside the evaluation harness. The Claude API runs we report were performed in the period preceding the v1 release; running the same prompts at a later date may produce different numbers as the model is updated.

F.3. Random Baselines and Reporting Conventions

Random baselines per task: PSC-10 (10%), PSC-25 (4%), TED 4-MCQ (25%), CCR pairwise (50%), CCR Kendall's τ (0), VSD binary (50%), TED-Visual 2-MCQ (50%).

For tasks with item counts below 50 (TED-Visual Hard, Strict Hard; VSD pure-motion subset), we treat reported gaps within 5–6 percentage points as within evaluation noise. The 9.1-point Tacit lead on Strict Hard ($n = 33$) is large relative to this noise; smaller gaps in the paper should not be over-interpreted (Section 7).

G. Operational Analysis

This appendix expands the deployment-cost calculation summarized in §6. The numbers below are illustrative rather than universal: actual costs depend on the specific VLM provider, the prompt structure, the polling frequency, and the GPU class. We report a representative case to make the order-of-magnitude asymmetry concrete.

G.1. Cost Model

Consider an autonomous liquid-liquid extraction (LLE) platform that polls a perception module every 5 seconds during an 8-hour operating shift to determine whether the current physical state has progressed (e.g., from *settled* to *draining*, or from *draining* to *drained*). The polling frequency is set by the temporal resolution at which downstream actions need to be triggered: a 5-second poll produces a ∼5-second worst-case action latency, which is appropriate for procedures whose state transitions take 30 seconds to several minutes.

Commercial VLM perception layer. At 5-second polling intervals over an 8-hour shift: $8\ \text{hr} \times 3600\ \text{s/hr} / 5\ \text{s} = 5{,}760$ API calls per shift. Each call submits 5 sampled frames plus a structured prompt that asks the VLM to classify the current state

and any observed transition. Image-conditioned long-context queries through current Claude Opus or GPT-5.5 endpoints typically run $0.50–$2.50 per call depending on prompt size and image resolution; we use $1.25 as a representative midpoint. Daily cost: $5{,}760 \times \$1.25 = \$7{,}200$ per platform per shift.

Domain-adapted vision encoder layer. Tacit's inference footprint is small: ViT-L/14 at 384×384 resolution, 16 frames per clip, runs at ∼15–20 ms per clip on an A10G GPU and ∼8–12 ms on a T4 GPU. A polling rate of 5 seconds is well below the encoder's serving rate; the GPU sits idle most of the time. Reserved A10G capacity on AWS, Azure, or GCP is approximately $0.50–$0.80 per hour; over an 8-hour shift, ∼$4–$6 of GPU time. Even amortized including a probe head, an inference server, and overhead, the daily cost is ∼$20 per platform per shift.

Asymmetry. $7,200 vs $20 per platform per shift is a ∼360× deployment-cost ratio. Per perception decision: $1.25 vs ∼ $0.0035, a ∼360× ratio at the per-call level as well. Scaling to a 24-hour autonomous platform (3 shifts) the ratio is preserved: ∼$21,600/day for the commercial-VLM layer vs ∼$60/day for the domain-adapted encoder.

G.2. Hybrid Deployment Pattern

The asymmetry above does not translate to "replace the VLM with the encoder." On the language-amenable end of the structural axis (PSC, TED, CCR pairwise; §5), Claude Opus leads Tacit by 6–41 points, and that gap is meaningful for procedural-class identification, error detection, and decisions that require reading instructional captions or reasoning about apparatus identity. Replacing the VLM entirely would forfeit those gains.

The asymmetry instead motivates a *hybrid* deployment pattern. A domain-adapted vision encoder runs in the high-frequency polling loop (every 5 seconds at $0.0035/call), handling routine state-tracking and motion-discrimination. A commercial VLM is consulted at low frequency — on confidence-triggered escalation (the encoder's predicted state probability falls below a threshold), at procedural transition points (e.g., when the platform schedules a new step), or for explicit error-detection queries (TED-style "did the last transition succeed?") — at perhaps 10–50 calls per shift rather than 5,760. At 50 calls per shift the VLM cost falls to ∼$60/shift while preserving its language-amenable advantages.

Combined cost. Hybrid (encoder high-frequency + VLM low-frequency at 50 calls/shift): ∼$80/shift, vs ∼$7,200/shift for VLM-only. A two-order-of-magnitude reduction at minimal capability cost.

G.3. Caveats

Per-call pricing varies. Actual VLM API pricing depends on prompt size, image resolution, image count per call, and provider tier. The $1.25/call midpoint we use is representative for image-conditioned long-context queries with 5 frames at moderate resolution; production deployments should cost-model their specific prompt structure rather than relying on this number directly.

Polling frequency varies by procedure. Some procedures (e.g., a slow recrystallization where state transitions take hours) tolerate 60-second polling; the asymmetry then narrows to ∼30×. Faster procedures (e.g., a real-time mixing operation where transitions happen in seconds) may need 1-second polling; the asymmetry widens to ∼1,800×.

Encoder operating cost is GPU-class-dependent. A T4 instance ($0.35/hr on AWS) is sufficient for Tacit at its current inference rate; an A10G ($0.50–$0.80/hr) provides headroom for batched serving across multiple platforms.

The scientific finding precedes the operational case. The structural-axis gradient (§5) and the within-state-physics finding (§6) stand independently of the cost model. The cost model converts those findings into deployment guidance but does not justify them.