Practical Advances in Artificial Intelligence and Computing
Frontier Research Publishing • https://www.frpubs.org
Vol. 5, No. 7, 2026

Microscopic Cell Tracking through Detection Identity and Lineage

Isvenna Thornquill[1]*, Caelric Bellamere[1], Ysolde Fennwick[1]
[1]University of Maine, Orono, United States
*Corresponding author: isvennathornquill757@outlook.com

Abstract: Microscopic cell tracking requires temporal identity and lineage evidence beyond accurate regions in individual frames. This article develops a structured measurement framework informed by 2026 tracking work and real published microscopy examples. Bright-field, phase-contrast, and fluorescence displays retain their source appearance. A constructed intensity field illustrates threshold dependence, known crossing trajectories isolate identity exchange, and analytic curves expose link-error accumulation and frame-interval effects. The proposed validation separates detections, links, complete graphs, and biological endpoints across independent recordings. It retains uncertainty, physical calibration, and review effort. No new segmentation benchmark, tracking leaderboard result, or biological experiment is claimed.

Keywords: cell tracking; microscopy; lineage; identity; segmentation; biological measurement

1 A Cell Is More than a Region in One Frame

A microscopy image can support a convincing cell boundary while leaving the cell's identity across time unresolved. Tracking requires linking observations, handling disappearance and reappearance, and representing divisions. A segmentation score on isolated frames does not measure those relationships. This paper develops a measurement-oriented framework for cell tracking, grounded in real bright-field, phase-contrast, and fluorescence examples. Its central argument is that detection, identity, and lineage should be evaluated separately before they are combined into a biological conclusion.

Recent work broadens the range of models considered for these tasks. CellTrack-Bench, released in 2026, examines microscopic tracking with vision-language models [1]. A 2026 higher-order tracking transformer explicitly addresses relationships that extend beyond an isolated pair of detections [2]. Their relevance is the question they expose: what temporal and relational evidence is necessary to maintain a biologically meaningful trajectory? This manuscript does not evaluate those models or claim a new leaderboard result. It uses source images and constructed tracking examples to define a reproducible stress-test programme.

The biological object should be specified before annotation. A nucleus, a whole cell, a bacterial body, and a fluorescently marked subcellular structure have different boundaries and division conventions. A tracking record may concern a visible object rather than every biological cell in the specimen. If a cell is outside the focal plane or loses its marker, its absence from a frame need not imply death. The output schema should retain these distinctions instead of interpreting every missing detection as a biological event.

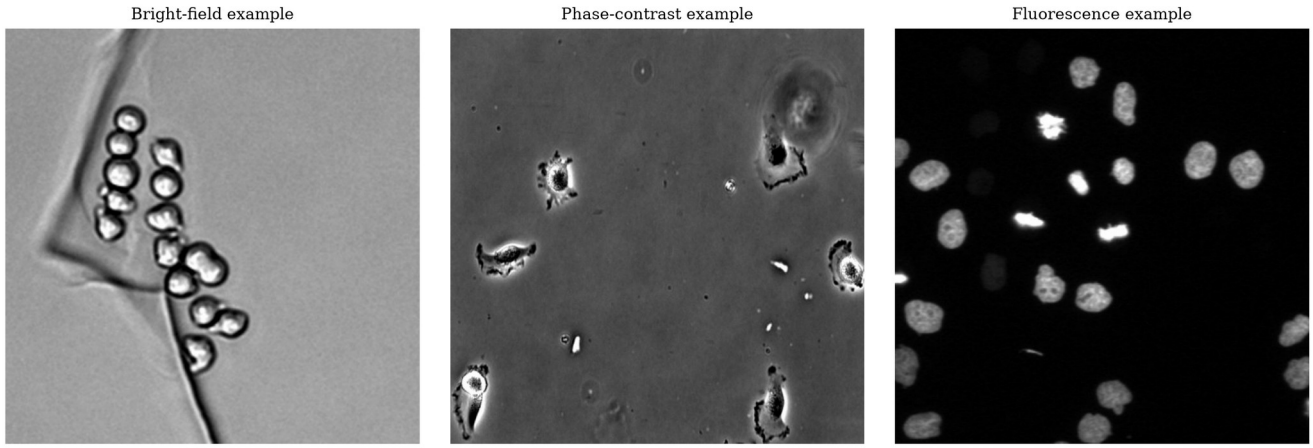


Figure 1. Unmodified microscopy examples from the source images distributed with CellTrack-Bench [1], arXiv:2610.31566v1. The panels are BF-C2DL-HSC, PhC-C2DH-U373, and Fluo-N2DL-HeLa. Images are juxtaposed with external modality labels; no new masks or predictions are overlaid. Source license: CC BY-NC-ND 4.0.

Figure 1 presents actual microscopy examples published with the 2026 benchmark. Their appearance is retained without inventing new instance labels. Figure 2 applies declared intensity thresholds to a constructed fluorescence-like field; the masks are image-processing illustrations, not reference segmentations. Figure 3 constructs two crossing trajectories with known identities, and Figure 4 calculates the accumulation of link uncertainty under explicit assumptions. These roles are kept separate throughout the paper.

2 Detection Evidence across Imaging Modalities

Cellpose demonstrated a generalist approach to cellular segmentation [3]. Cellpose 2.0 examined practical model adaptation [4], and Cellpose3 extended the problem to image restoration for improved segmentation [5]. These contributions show that image quality and local annotation can matter alongside network architecture. They do not imply that a restored boundary is an independently observed biological contour. Restoration should be assessed for its effect on the downstream measurement and checked for features introduced or removed by the processing.

The Cell Tracking Challenge established a long-running framework for objective comparison [6]. Star-convex three-dimensional representations offer a different geometric model for detection and segmentation [7]. Ultrack extends tracking across biological scales [8]. These methods have different assumptions about object shape, image modality, annotation, and temporal context. A fair comparison should state those assumptions and evaluate the intended specimen rather than treating all cellular imagery as one interchangeable domain.

Bright-field and phase-contrast images can show halos and internal texture that complicate a simple intensity boundary. Fluorescence can make a tagged structure conspicuous while leaving untagged structures unseen. Saturation, bleaching, background variation, and changes in focus affect all later measurements. The reference protocol should identify which visible structure defines the object. A model should not be penalized for missing a cell that the chosen imaging channel cannot resolve, nor credited with a complete biological count when only a labeled subset is visible.

Table 1. Observation quantities and their reference requirements.

Quantity or input	Required evidence	Interpretation
Object boundary	Imaging modality and object convention	A nucleus and a whole cell are different objects
Persistent identity	Temporal correspondences	Accurate regions can still exchange identities
Lineage division	Parent-child reference events	One-to-one linking cannot represent every division
Biological endpoint	Physical calibration and correct graph	A graph score does not validate all biological quantities

Figure 2 makes threshold dependence visible. Three masks are produced by applying three declared thresholds to a constructed sum of Gaussian intensity spots with declared amplitudes and widths. They show how a numerical choice can split, merge, or remove bright regions. The field is synthetic and has no biological specimen claim. The masks do not identify a biological reference boundary. That requires an appropriate reference and, where necessary, additional imaging information. The display also illustrates why a visually pleasing mask is insufficient evidence for measurement accuracy: different masks can look coherent while assigning different areas and object counts.

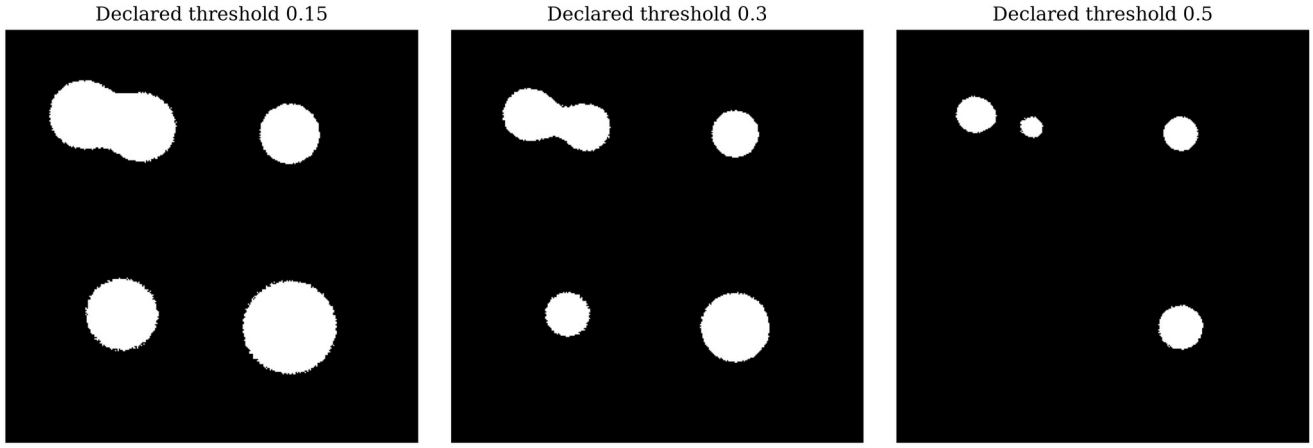


Figure 2. Thresholds of 0.15, 0.30, and 0.50 applied to a constructed Gaussian-spot intensity field with seed 32. This is an analytic image-processing example, not a measured biological specimen, a reference segmentation, or a trained-model output.

3 An Explicit Tracking Representation

Let an object detection at frame t have a location, shape, and appearance representation. A linking model evaluates whether it corresponds to a detection at a later frame. Equation (1) gives a simple weighted cost combining displacement, appearance difference, and shape difference. It is a design example. The terms require normalization and should be chosen for the imaging conditions. A large change in apparent shape can reflect a real division or a focus change; the cost alone cannot distinguish those explanations without temporal and biological context.

$$C_{ij} = \alpha\, d_{position}(i,j) + \beta\, d_{appearance}(i,j) + \gamma\, d_{shape}(i,j) \tag{1}$$

A one-to-one assignment is appropriate only while the chosen object persists without division or merging under the output convention. Division introduces one-to-many relationships. A missed frame introduces a possible gap. A segmentation merge can create a temporary many-to-one observation without implying that biological cells fused. The representation should distinguish these cases. Otherwise, a graph can be mathematically consistent while expressing the wrong biological process.

Figure 3 isolates identity from localization. Two constructed trajectories cross. The positions can be detected perfectly in every frame, yet nearest-position linking can exchange the identities. Both outputs have the same collection of observed points. A frame-wise detection metric cannot distinguish them. The tracking reference must include temporal correspondences, and the biological analysis must state whether an identity exchange affects its intended quantity, such as migration direction, division timing, or cell-specific response.

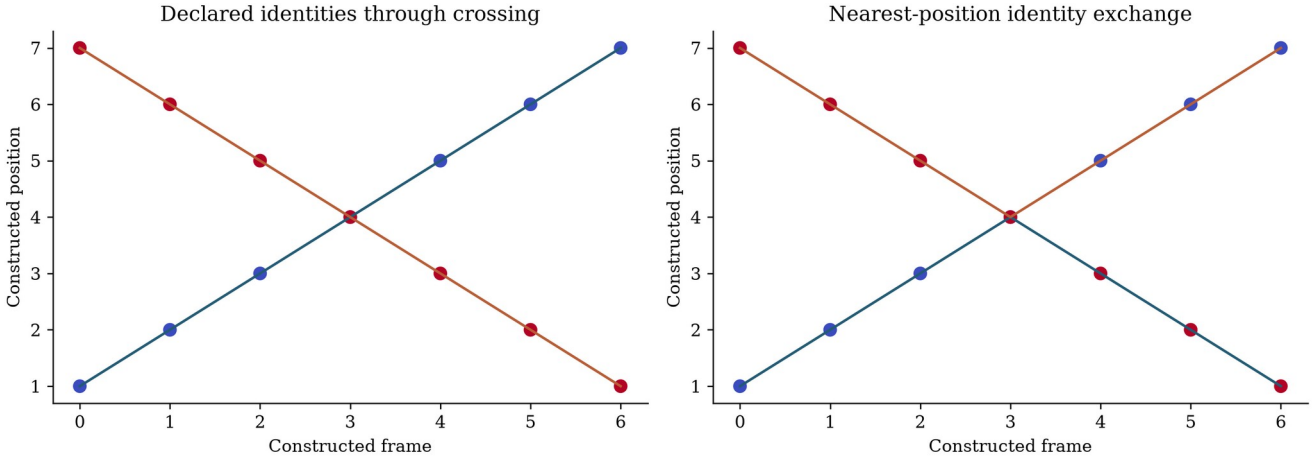


Figure 3. Constructed crossing trajectories with the same detections in both panels. Left: declared identities. Right: a position-based identity exchange. The example isolates association ambiguity and is not a benchmark tracking result.

A lineage graph should record parent-child edges, track start and end reasons, gaps, and review status. It should also retain links to source frames so that a disputed relationship can be inspected. A generated textual explanation can be useful for a reviewer, but it is not a replacement for that visual support. If a model describes a division, the graph and source interval should identify the detections that support it. The statement can then be checked rather than accepted because it is fluent.

4 Models and the Information They Receive

Random forests and support vector machines remain useful for explicit appearance or motion summaries [9], [10]. They provide baselines whose behavior can be examined directly. Residual visual features can represent richer local appearance [11]. Patch transformers and hierarchical encoders can integrate surrounding structure [12], [13]. An encoder-decoder can produce dense object evidence [14]. The comparison should control the temporal interval and image resolution supplied to each method. A model receiving future frames has a different information setting from an online tracker.

Focal loss can alter the influence of easy background pixels [15]. It does not solve the scarcity of annotated divisions or rare failure types. Language-aligned representations may permit flexible descriptions of morphology [16], but the words must be evaluated against a defined annotation convention. A phrase such as dividing cell is a relational claim about a temporal process. It should be scored through the corresponding lineage event and source interval rather than by linguistic similarity alone.

Transfer between imaging laboratories can change illumination, magnification, specimen preparation, and object appearance [17]. Boosted trees can combine explicit tracking features in a second-stage decision [18]. The stage should be trained on realistic predicted detections, including their errors. Reference detections are valuable for diagnosis, but a relation model evaluated only with perfect input detections does not establish complete tracking performance. Both settings can be reported if their inputs remain explicit.

The outer split should follow independent specimens or experiments [19]. Adjacent frames from one recording should not be randomly divided into training and testing. They share cells, imaging conditions, and often near-identical pixels. Model-comparison uncertainty should respect those experimental units [20]. A long movie supplies many temporal observations, but it does not necessarily supply many independent biological replicates. The distinction matters when an apparent algorithmic gain is used to support a biological generalization.

5 Identity Reliability over Time

Confidence should refer to a specific output: a boundary, an object detection, a link, or a division event. Calibration can be assessed for each level [21]. Proper scoring rules evaluate a probability statement [22], while graph metrics evaluate the final structured record. A well-calibrated link probability does not automatically yield a well-calibrated whole-lineage probability because errors can be dependent and graph constraints can alter decisions. The aggregation rule should therefore be stated and tested.

Equation (2) gives a deliberately simplified probability that K independent links are all correct when each has probability q. Figure 4 shows how a seemingly strong local reliability can deteriorate over a long trajectory. The assumption of independence is not asserted for actual cells. The calculation isolates accumulation as a concern. Real tracking errors may cluster around crowded or poorly focused intervals, so the observed trajectory-level reliability can differ substantially from this simple expression.

$$P(allKlinkscorrect)=q^{K} \tag{2}$$

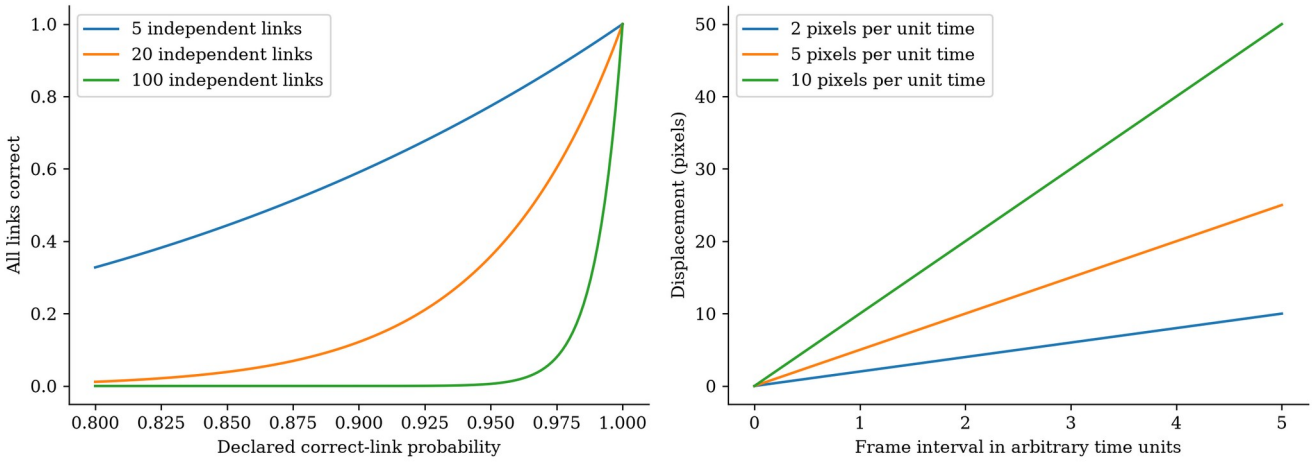


Figure 4. Declared link reliability q raised to independent-link counts, and displacement under specified speed and frame intervals. The independence and motion assumptions are illustrative, not measured biological dynamics.

Frame interval is another physical constraint. For a declared motion speed v and interval delta t, Equation (3) gives a displacement scale. When that scale becomes comparable to intercell spacing, appearance and higher-order context become increasingly important. The right panel of Figure 4 plots this relationship in arbitrary declared units. It is not a measured migration speed. A real experiment should report its physical

pixel size and acquisition interval before interpreting a tracking difficulty as a property of the algorithm alone.

$$d_{expected} = v\,\Delta t \tag{3}$$

Ensemble disagreement can help select links for review [23]. Attribution can reveal image regions affecting a detection or association [24]. These tools should be tested for whether they help a reviewer correct the graph efficiently. Shared shortcuts can make an ensemble confidently wrong [25]. An explanation focused on a bright neighbor can be a useful diagnostic, but it does not establish the correct identity. Independent reference correspondences remain necessary.

6 Stress Tests for a Tracking Study

A useful study should separate detection, linking, and lineage errors. First evaluate object observations against an agreed boundary convention. Then evaluate links using matched objects. Finally evaluate complete graphs using the actual predicted detections. This sequence identifies where errors enter and how they propagate. Reporting only the best intermediate setting can hide a weak complete pipeline, while reporting only a final graph score can hide a simple upstream cause that is easy to improve.

Hyperparameter and threshold choices should be made within the training experiments [26]. A held-out movie should not be inspected repeatedly while adjusting gap tolerance, division penalties, and object-size filters. Distribution tests can monitor changes in image summaries or representations [27], but they do not tell whether a new pattern is a biological response or an imaging artifact. A regularized linear link model provides a useful stable reference for assessing the benefit of a more complex association scheme [28].

Table 2. Prospective evaluation units and independent references.

Evidence or condition	Required comparison	Interpretation
A complete biological replicate and recording	Hold out recordings and replicates before extracting patches	Reference objects and tracks in independent recordings
A tracked trajectory under a declared acquisition cadence	Subsample complete recordings using fixed declared intervals	The original high cadence tracking reference
A division or other infrequent annotated event	Evaluate divisions separately under their observed frequency	Independently reviewed parent child relationships
A model adapted to a new recording condition	Compare frozen and adapted pipelines after the same annotation budget	Labels known before fitting and later blind reference graphs

A temporal subsampling test can increase the frame interval while retaining the known source frames. An occlusion test can remove a declared interval and evaluate gap recovery. A crowding test should use real dense regions or a clearly constructed diagnostic rather than silently pasting cells into a purported biological recording. Each test has a narrow purpose. Its outcome should not be generalized to conditions it does not reproduce, such as phototoxicity, three-dimensional motion, or a new cell type.

The system should retain ambiguous events. If two possible links are visually indistinguishable, a reference can preserve alternatives or identify the interval as unresolved. Scoring should then follow that convention. A tracker that requests review at the correct uncertain event may be more useful than one that

invents a confident identity. Review effort, coverage, and final graph quality should be reported together so that the benefit of human assistance can be evaluated rather than assumed.

7 Biological Use and Reproducibility

A biological analysis should state how tracking errors affect its endpoint. A brief identity exchange may have little effect on a total cell count but a large effect on a cell-specific trajectory. A missed division can distort lineage depth and generation time. A merged segmentation can alter apparent cell size. These consequences should be measured for the intended analysis. A single benchmark score is not a universal certificate that every downstream biological quantity is accurate.

A reproducible package should preserve source recording identifiers, imaging metadata, split assignments, annotation instructions, model versions, and graph postprocessing. Derived tracks should retain a link to every source detection. If privacy, consent, or distribution conditions limit the original images, the manifest should still make the processing path identifiable within the allowed access setting. The source examples in this article remain illustrations; they are not presented as a complete reconstruction of the benchmark archive.

Equation (4) defines one simple motion estimate from linked positions and the frame interval. It highlights why identity and physical calibration matter together. A correct numerical difference between positions is not a valid migration measurement if the points belong to different cells. Conversely, correct identity does not establish physical speed without pixel-size and time calibration. The tracking record should therefore preserve both relational and instrumental provenance.

$$v_i = ||r_{i,t+1} - r_{i,t}|| / \Delta t \tag{4}$$

Table 3. Failure mechanisms and records needed for interpretation.

Evidence or condition	Required comparison	Interpretation
Halos occlusion and partial visibility	Boundary disagreement and unresolved object fraction	Annotation instructions and reviewer alternatives
Focus loss occlusion and marker disappearance	Gap recovery and false link rates	Gap masks and track termination reasons
Few labels and modality specific artifacts	Paired graph errors and annotation effort	Training samples cutoffs and model versions
Identity errors coupled with size and motion estimates	Endpoint bias by error type	Endpoint definitions and source supported graph provenance
New modalities cell types and preparation	Coverage errors and required review time	Instrument settings specimen metadata and complete outputs

8 Tracking Stress Studies for Independent Recordings

The following tests organize a specimen and lineage validation programme. They are designed comparisons rather than additional executed tracking experiments. They proceed from microscope calibration to object convention, independent recordings, temporal support, identity, and the biological endpoint. A laboratory should select the tests relevant to its actual specimen and imaging process. The resulting evidence

should distinguish a useful local tracking tool from a broader claim about unseen modalities or biological systems.

8.1 Imaging calibration

The first requirement in imaging calibration is a traceable observation. The experimental unit is a time lapse recording and its channels, and the quantity of interest is an observed cellular structure in calibrated space and time. The main factor to audit is magnification exposure and channel registration. This test should begin with the original measurement record, before fitted scores or attractive examples are examined. A source value should retain its units, sampling support, and availability status. If a derived quantity changes, the study should be able to identify the transformation responsible. Otherwise, a difference attributed to learning can originate in the measurement pipeline. The reference should be microscope metadata and suitable calibration observations, with its own uncertainty stated rather than assumed away.

The proposed comparison is to compare processing under documented instrument changes. This is a prospective protocol, not an additional experiment reported as completed here. It should use the same eligible units and the same target convention on both sides. Acquisition and transformation choices should be fixed before the final evaluation. The diagnostic is position size and intensity consistency. Examine its distribution across the actual independent units, retaining failures and unsupported observations in the denominator where the policy requires them. Report the raw counts or residuals alongside any normalized summary, because an identical percentage can represent very different evidential support in two collections.

The key interpretive limit is that a bright region need not be the entire biological cell. If the comparison changes both physical support and the learned representation, it cannot isolate an architectural benefit. A useful sensitivity analysis changes one documented component at a time and reports whether the conclusion persists under reasonable alternatives. Preserve pixel dimensions frame intervals and channel transforms. These records connect the calculation to the source and make an independent reconstruction possible. A technically successful result would identify the conditions under which the observation remains usable, the conditions under which it fails, and the additional measurement needed to resolve those failures. It would not infer a wider application merely from a clean numerical output.

8.2 Object convention

The reference study for object convention concerns a cell or nucleus under a defined annotation rule. It asks whether a consistent boundary and biological object identity can be supported consistently when the evidence is influenced by halos occlusion and partial visibility. Begin with a written reference convention and a sample chosen independently of model confidence. The reference is reviewed source frames with the object convention stated. Reviewers should have the information needed for that convention, but an intended independent judgment should not be anchored by the model's proposed answer. Retain disagreements and the reason for uncertainty. Consensus can be useful, yet it should not erase the fact that some observations support more than one reasonable interpretation.

To investigate this question, compare independent annotation and retained ambiguous boundaries. Score boundary disagreement and unresolved object fraction, and distinguish uncertainty in the reference from disagreement with a definite label. An ambiguous case can be evaluated through admissible alternatives, an uncertainty interval, or an explicitly excluded category, provided the choice is fixed in advance. Exclusion changes the evaluated population and must remain visible. The study should also measure the effect of the reference policy on the substantive conclusion. A small average disagreement can matter if it concentrates in

rare or consequential cases. Conversely, tiny discrepancies may be scientifically immaterial under the intended measurement tolerance.

Interpretation should respect the fact that nuclear and whole cell labels are different measurements. The purpose of review is to establish what the observation supports, not to make every case fit a binary success story. Compare the result under reasonable reference conventions without selecting the most favorable one after seeing the model. Preserve annotation instructions and reviewer alternatives, including the unresolved cases and the final adjudication history. These materials allow later researchers to distinguish a model improvement from a change in annotation practice. They also identify where a better instrument, another observation, or domain expertise would provide more value than further optimization against an uncertain label.

8.3 Independent specimens

The generalization test for independent specimens uses a complete biological replicate and recording as the unit that must remain intact. Its target is transfer to a new specimen, while the dependence of concern is adjacent frames sharing cells and acquisition conditions. A random split of convenient numerical rows can conceal this dependence. The partition should therefore be created at the source-unit level before overlapping crops, windows, or augmented records are derived. Each child record should inherit its parent's assignment. The independent reference is reference objects and tracks in independent recordings. This establishes an evaluation population that corresponds to the claim rather than to the easiest available data arrangement.

The proposed experiment is to hold out recordings and replicates before extracting patches. Evaluate recording conditional detection and graph errors for the held-out units, showing their range as well as a pooled value. The split should also state which kinds of novelty remain shared across the boundary. Holding out one axis does not automatically hold out every other axis. If the intended use requires several changes at once, assess that combined condition explicitly and use additional comparisons to diagnose the source of any deterioration. Development decisions should be made without inspecting the locked outer reference labels.

The main limit is that many frames do not establish many biological replicates. An uncertainty calculation should resample or otherwise model the units that are independent for the scientific question. Treating correlated children as independent trials can make a small difference appear decisive. Preserve specimen identifiers and patch ancestry, including the construction of every derived record. A successful test would support a statement about the chosen transfer axis and the represented population. It would also identify whether errors concentrate in particular groups. This is more informative than a single favorable split that cannot be related to a concrete future observation setting.

8.4 Frame intervals

Timing in frame intervals should be defined for a tracked trajectory under a declared acquisition cadence. The intended statement is identity persistence at the intended temporal resolution, and the potential mismatch is motion between observations and omitted frames. Record when the physical observation occurred, when the processed input became available, and when the model or decision was issued. These timestamps can differ substantially. A retrospective array can contain information unavailable at the nominal issue time even when its elements carry earlier labels. The timing reference is the original high cadence tracking reference. The test should reconstruct the allowed information state rather than assuming that a correctly sorted table is necessarily causal.

The proposed comparison is to subsample complete recordings using fixed declared intervals. Measure identity switches and track survival by interval under the actual buffering and processing policy. A result

should retain the exposure and issue schedule, including intervals that produce no usable output. Delayed reference information must not enter an earlier model update. If the target interval overlaps successive issues, the dependence should remain part of verification. A useful ablation changes the permitted history or waiting rule while holding other conditions fixed. This separates an improvement due to additional observation from an improvement due to a better representation of the same observation.

The interpretation is constrained because coarser time sampling changes association difficulty. Preserve retained frame lists and physical time calibration, so that another analysis can replay the issue sequence and verify what the system could have known. Report the distribution of delays, not merely the average forward-pass time of one model call. Long tails and missed deadlines can determine usefulness for a time-sensitive action. A defensible result identifies the time at which information becomes actionable and the conditions under which the intended interval is no longer supported. It does not infer timely operation solely from accurate retrospective localization or classification.

8.5 Temporary disappearance

The incomplete-observation study for temporary disappearance concerns a trajectory with an explicit observation gap. It tests gap recovery without inventing biological death when observations are affected by focus loss occlusion and marker disappearance. Missingness should have an explicit representation and provenance. A numeric placeholder is not automatically a physical zero, and an absent output is not automatically a negative observation. Define the eligibility rule and fallback before evaluating results. The reference is complete source trajectories where the gap is constructed, which should distinguish unavailable measurement from a genuine observed absence. This separation is necessary to keep both coverage and accuracy interpretable.

The experiment should remove declared observations and compare gap handling policies. Evaluate gap recovery and false link rates jointly, because a method can improve accepted accuracy simply by declining difficult inputs. Record the reasons for rejection and examine whether they are associated with the target or with a particular population. A constructed deletion test is useful for isolating an algorithmic dependency, but it should remain identified as constructed. Naturally occurring interruption may have a different mechanism and may correlate with difficult environmental or operational conditions. Both forms of evidence can be useful if their roles are kept separate.

The limit is that a missing detection does not establish cell death. A robust system should make its unsupported interval visible and retain enough context to choose a more informative next observation. Preserve gap masks and track termination reasons. Report the number and distribution of eligible, accepted, rejected, and recovered units under the same accounting rule. This prevents a clean-looking map or sequence from implying coverage it does not have. The substantive conclusion should concern reliability under the tested missingness conditions, with a clear account of how the fallback changes the scientific or operational target. An untested imputation strategy should not silently turn absent evidence into confident fact.

8.6 Rare lineage events

The prevalence test for rare lineage events takes a division or other infrequent annotated event as its exposure population. The intended outcome is reliable event relationships rather than pooled frame accuracy. The difficulty is many easy nondivision links, which can make the evaluation mixture unlike the mixture encountered in use. State how the source sample was selected and which conditional rates are being transported to a new scenario. The independent reference is independently reviewed parent child

relationships. A balanced development set can be appropriate for fitting, but its class proportions should not be mistaken for a measurement of the future population.

The proposed analysis is to evaluate divisions separately under their observed frequency. Measure division precision recall and review count, retaining the eligible exposure denominator and the decision aggregation rule. A probability sample can estimate operating performance directly within its supported population. Scenario reweighting can expose consequences under alternative assumed proportions, but the assumed stability of conditional errors must be stated. It is often useful to report both the empirical source result and the scenario calculation. They answer different questions. Do not select only periods with known target events when the claimed workload concerns a continuous stream that also contains long target-free intervals.

Interpret the result with the limit that high link accuracy can hide missed divisions. Preserve event sampling and lineage matching rules, including the source sampling and any weights. A useful conclusion connects false detections and misses to a defined amount of observation, review, or action. It should show where uncertainty in prevalence materially changes the decision. Strong performance on an enriched collection can justify further study while remaining insufficient for a rare-event deployment claim. Making the population and exposure explicit helps readers distinguish those levels of evidence without dismissing the value of a carefully controlled benchmark.

8.7 Identity crossing

The controlled perturbation in identity crossing is applied to a crowded interval containing close trajectories. The scientific target remains correct identity through ambiguous spatial encounters; the factor varied is nearest neighbor identity exchange. Begin by explaining why the intervention represents a meaningful change in the observation process. A numerical alteration can diagnose dependence without reproducing a real physical condition. Its value is greatest when the preserved information and the changed information are both explicit. The reference is reference identities through the complete interval, and any constructed component should retain a known design rather than being described as a new empirical observation.

The proposed experiment is to compare position only and appearance informed linking on the same detections. Assess identity switches and endpoint specific trajectory error using the complete processing chain that the claim requires. If an intervention changes raw observations, recompute the actual derived inputs rather than changing a convenient proxy and assuming equivalence. Include a control that leaves the relevant physical quantity unchanged where that is possible. Compare errors across the declared perturbation range, avoiding a single unusually favorable setting. The test can reveal a dependence, but identifying its mechanism may require additional observations that vary the suspected factors independently.

The interpretive boundary is that perfect detection can coexist with an incorrect lineage. Preserve candidate links costs and source frame references, and keep the transformed observations distinguishable from the originals. A successful diagnostic explains which part of the model or measurement becomes unstable and at what support scale. It should not infer broad robustness from an algebraic property of one feature or an attractive transformed display. The next step can be a physical replication under the corresponding condition, with independent reference measurements. This progression connects a reproducible analytic test to the real observation without claiming that the former has already substituted for the latter.

8.8 Local model updates

The adaptation study for local model updates evaluates a model adapted to a new recording condition. Its desired outcome is improvement on subsequent independent movies, and the main risk to interpretation is few labels and modality specific artifacts. A change detector and a model-quality comparison answer different questions. The former identifies a departure in a selected representation; the latter asks whether a candidate improves the intended output. Define the training cutoff, permitted labels, update budget, and promotion rule before the later evaluation. The independent reference is labels known before fitting and later blind reference graphs. Both an incumbent and a challenger should be assessed on the same eligible future observations.

The proposed comparison is to compare frozen and adapted pipelines after the same annotation budget. Evaluate paired graph errors and annotation effort with paired records wherever possible. Labels acquired after an update cannot be treated as if they had been available for fitting it. Pseudo-labels should retain their generated provenance and should not silently become independent reference truth. An annotation budget should include the actual selection and review process, because targeted easy examples can make adaptation look cheaper than it is. A frozen reference is valuable even when it is ultimately inferior, because it exposes whether change was needed and what the update actually contributed.

The interpretation must recognize that better segmentation alone may not improve tracking. Preserve training samples cutoffs and model versions, including rejected candidates and the conditions that triggered review. An adaptation result should report stability, coverage, and relevant errors rather than only the best post-update score. If the update is harmful for a rare subgroup, a pooled improvement can hide that consequence. A defensible promotion decision rests on independent later evidence and a declared operating objective. It should not be justified merely by a distribution alarm, a newer architecture, or a visually persuasive explanation of an individual prediction.

8.9 Biological endpoints

The decision-level test for biological endpoints uses a graph used for a declared biological measurement. It asks whether a valid trajectory or lineage derived quantity remains supported after the model output enters the intended analysis or action. The factor of concern is identity errors coupled with size and motion estimates. Define the downstream quantity before examining model results, including any aggregation, threshold, matching, or cost assumptions. The reference is reviewed identities plus physical calibration. This creates a link between an intermediate score and the consequence that motivates the work. Without that link, a favorable technical metric can be assigned practical meaning that the experiment never measured.

The proposed comparison is to compare endpoints from reference and predicted complete graphs. Report endpoint bias by error type alongside the intermediate errors, so that their relationship can be inspected rather than presumed. Different error types can have unequal effects on the endpoint. A small number of consequential errors may dominate a summary that looks stable on average. Conversely, some image or sample discrepancies may have little effect on the intended quantity. A sensitivity analysis should vary defensible decision assumptions without choosing the one that makes the model look best. Retain the underlying counts and source-supported measurements needed to assess alternatives.

The key limit is that one graph score does not validate every biological analysis. Preserve endpoint definitions and source supported graph provenance, including unresolved or declined decisions. A useful result identifies the conditions under which the model changes an endpoint, the uncertainty of that change, and the additional evidence required for a stronger interpretation. This is a substantive validation question,

not an administrative final step. It allows a complex learning system to be judged on the scientific or operational contribution it actually makes while keeping broader claims available for separate, appropriately designed studies.

8.10 Laboratory transfer

The external study for laboratory transfer concerns an independent microscope specimen and preparation. Its claim is the evaluated biological and imaging envelope, under changes in new modalities cell types and preparation. Replication should begin from a locked feature, label, and decision specification, while documenting any necessary local mapping. The independent reference is local object conventions and independent graph review. A new environment may use the same field names or category words while measuring something different. Check those meanings before comparing numerical scores. Unsupported mappings should remain visible so that a transfer result is not created by silently redefining the task.

The proposed experiment is to test the locked pipeline on separately acquired recordings. Measure coverage errors and required review time, including the availability of valid inputs and the effort needed to obtain the reference. Report local conditions alongside the aggregate result. A model may be useful within a constrained envelope even when it fails elsewhere. The aim is to identify that envelope precisely, not to turn every local success into a universal claim. A reproducibility package should retain the configuration needed to reconstruct preprocessing, fitting, issue, and scoring, while preserving the permitted provenance of any source material that cannot be redistributed freely.

The limiting explanation is that success on one modality does not establish universal cell tracking. Preserve instrument settings specimen metadata and complete outputs. The final account should distinguish what was observed, what was fitted, what was constructed for diagnosis, and what remains a proposed next test. Numerical reproducibility is necessary for the executed calculations, but it does not by itself establish external scientific validity. An informative replication can therefore confirm a useful result, reveal a boundary, or identify a measurement mismatch. Each outcome contributes knowledge when the source records and target definitions make the reason assessable.

9 Conclusions

Microscopic tracking is a structured measurement problem involving visible objects, persistent identity, and biological lineage. Real modality examples show why object evidence varies with the imaging process. The threshold and crossing demonstrations isolate segmentation and identity errors, while the analytic curves explain why long trajectories and frame intervals require explicit evaluation. No new tracking accuracy is inferred from these displays.

The strongest next study would compare complete pipelines on independent recordings, retain uncertain events, and measure the effect of errors on a declared biological endpoint. It would distinguish perfect-reference diagnostics from autonomous operation and report review effort alongside coverage. Under that design, increasingly capable visual and language models can contribute useful evidence without obscuring the biological meaning of the tracks they produce.

References

[1] Theodoro, Ilan, Bragantini, Jordão, Royer, Loïc A., "CellTrack-Bench: Microscopic Cell Tracking Benchmark for Vision-Language Models", arXiv preprint 2610.31566, 2026, [Online]. Available: https://arxiv.org/abs/2610.31566.

[2] Bragantini, Jordão, Theodoro, Ilan, Royer, Loïc A., "Higher-Order Cell Tracking Transformer", arXiv preprint 2607.11754, 2026, [Online]. Available: https://arxiv.org/abs/2607.11754.

[3] C. Stringer, T. Wang, M. Michaelos, M. Pachitariu, "Cellpose: a generalist algorithm for cellular segmentation", Nature Methods, vol. 18, no. 1, pp. 100-106, 2021, doi: 10.1038/s41592-020-01018-x.

[4] M. Pachitariu, C. Stringer, "Cellpose 2.0: how to train your own model", Nature Methods, vol. 19, no. 12, pp. 1634-1641, 2022, doi: 10.1038/s41592-022-01663-4.

[5] C. Stringer, M. Pachitariu, "Cellpose3: one-click image restoration for improved cellular segmentation", Nature Methods, vol. 22, no. 3, pp. 592-599, 2025, doi: 10.1038/s41592-025-02595-5.

[6] M. Maška et al., "The Cell Tracking Challenge: 10 years of objective benchmarking", Nature Methods, vol. 20, no. 7, pp. 1010-1020, 2023, doi: 10.1038/s41592-023-01879-y.

[7] M. Weigert, U. Schmidt, R. Haase, K. Sugawara, G. Myers, "Star-convex Polyhedra for 3D Object Detection and Segmentation in Microscopy", 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 3655-3662, 2020, doi: 10.1109/wacv45572.2020.9093435.

[8] J. Bragantini et al., "Ultrack: pushing the limits of cell tracking across biological scales", Nature Methods, vol. 22, no. 11, pp. 2423-2436, 2025, doi: 10.1038/s41592-025-02778-0.

[9] L. Breiman, "Random Forests", Machine Learning, vol. 45, no. 1, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.

[10] C. Cortes, V. Vapnik, "Support-vector networks", Machine Learning, vol. 20, no. 3, pp. 273-297, 1995, doi: 10.1007/BF00994018.

[11] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition", 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016, doi: 10.1109/CVPR.2016.90.

[12] A. Dosovitskiy et al., "An Image is Worth 16x16 Words Transformers for Image Recognition at Scale", International Conference on Learning Representations, 2021, [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy.

[13] Z. Liu et al., "Swin Transformer Hierarchical Vision Transformer Using Shifted Windows", Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012-10022, 2021, [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper.html.

[14] O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", Lecture Notes in Computer Science, pp. 234-241, 2015, doi: 10.1007/978-3-319-24574-4_28.

[15] T. Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, "Focal Loss for Dense Object Detection", 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2999-3007, 2017, doi: 10.1109/ICCV.2017.324.

[16] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision", Proceedings of the 38th International Conference on Machine Learning, vol. 139, pp. 8748-8763, 2021, [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html.

[17] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, J. W. Vaughan, "A theory of learning from different domains", Machine Learning, vol. 79, no. 1-2, pp. 151-175, 2010, doi: 10.1007/s10994-009-5152-4.

[18] T. Chen, C. Guestrin, "XGBoost: A Scalable Tree Boosting System", Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016, doi: 10.1145/2939672.2939785.

[19] D. R. Roberts et al., "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure", Ecography, vol. 40, no. 8, pp. 913-929, 2017, doi: 10.1111/ecog.02881.

[20] C. Nadeau, Y. Bengio, "Inference for the Generalization Error", Machine Learning, vol. 52, no. 3, pp. 239-281, 2003, doi: 10.1023/a:1024068626366.

[21] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, "On Calibration of Modern Neural Networks", Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 1321-1330, 2017, [Online]. Available: https://proceedings.mlr.press/v70/guo17a.html.

[22] T. Gneiting, A. E. Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation", Journal of the American Statistical Association, vol. 102, no. 477, pp. 359-378, 2007, doi: 10.1198/016214506000001437.

[23] B. Lakshminarayanan, A. Pritzel, C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles", Advances in Neural Information Processing Systems, vol. 30, pp. 6402-6413, 2017, [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

[24] S. M. Lundberg, S. I. Lee, "A Unified Approach to Interpreting Model Predictions", Advances in Neural Information Processing Systems, vol. 30, pp. 4765-4774, 2017, [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[25] R. Geirhos et al., "Shortcut learning in deep neural networks", Nature Machine Intelligence, vol. 2, no. 11, pp. 665-673, 2020, doi: 10.1038/s42256-020-00257-z.

[26] J. Bergstra, Y. Bengio, "Random Search for Hyper-Parameter Optimization", Journal of Machine Learning Research, vol. 13, pp. 281-305, 2012, [Online]. Available: https://jmlr.org/papers/v13/bergstra12a.html.

[27] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, A. Smola, "A Kernel Two-Sample Test", Journal of Machine Learning Research, vol. 13, pp. 723-773, 2012, [Online]. Available: https://jmlr.org/papers/v13/gretton12a.html.

[28] A. E. Hoerl, R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems", Technometrics, vol. 12, no. 1, pp. 55-67, 1970, doi: 10.1080/00401706.1970.10488634.