Cell Identification from Partial Observation using Spatiotemporal Attention in *C. elegans* Embryogenesis

Henry Xue

Abstract

Automated cell identification in developing embryos represents a critical bottleneck in developmental biology. Manually annotating hundreds of embryos required in studies like genetic screens requires hundreds of hours of expert labor [11]. Real experimental workflows often only capture 5-20 cells of wider embryonic development through selective fluoro-labeling or optical constraints [2]. Existing computational methods assume observation of complete embryos and thereby are not viable for experimental workflows [42, 29]. Therefore, in this study, we hypothesize that an approach using transformer-based architectures could automatically identify cells from partial observation effectively. We apply a joint learning strategy processing paired neighborhoods simultaneously through multi-head attention to hundreds of semi-synthetic time lapse C. elegans developmental data. The architecture included key features such as geometric features rather than pure xyz coordinates, a no-match token, curriculum learning, and informed sampling strategies. The architecture achieved 93.3% accuracy on pure identification on held-out test embryos spanning 4-194 cell stages. Performance remained robust under heavy perturbations: 89.7% with 20% missing cells. Comprehensive baselines showed improvement over geometric methods (ICP [6]: 45.3%, CPD [26]: 52.1%) and template-based approaches (68.4%). These results suggest that local relational context, when properly encoded through joint attention, contains sufficient information for cell identification without full embryos. We conclude that learned spatiotemporal representations enable practical automatic identification from partial observation, transforming large-scale genetic screens and therapeutic discovery. Allowing for earlier disease discovery, more personalized medicine, and heightened ability for drug development and discovery.

1 Introduction

1.1 Rationale

Understanding how organisms develop from a single cell into complex multicellular forms requires tracking the behavior of individual cells as they divide, migrate, and differentiate. Modern live-imaging techniques can now capture these cellular dynamics at single-cell resolution across entire developing embryos [11, 22, 1], transforming our ability to observe the fundamental processes of morphogenesis. However, the biological insight promised by these imaging advances remains largely unrealized due to a critical computational bottleneck: identifying which specific cells are present and which are which in microscopy images.

1.2 Biological Context

The nematode *Caenorhabditis elegans* provides a uniquely powerful platform for addressing this challenge. Its transparent embryo develops through an invariant lineage from the zygote through 558 cells [33], providing gold-standard ground truth rarely available in other systems. Moreover, approximately 60–80% of *C. elegans* genes have human orthologs with particularly high conservation among developmental regulators [18, 17], ensuring that insights translate to human biology and disease. Complete 4D datasets can be acquired in 30–60 minutes using standard confocal microscopy with automated segmentation pipelines [2], while extensive molecular atlases enable integration of spatial dynamics with gene expression patterns [27, 9].

1.3 Annotation Bottleneck

Despite these advantages, manual cell annotation represents a severe experimental bottleneck. Curating *C. elegans* embryos from a single genetic screen requires hundreds hours of expert labor [11]. This overwhelming effort prevents the large-scale quantitative studies necessary to identify therapeutic targets for developmental disorders and cancer [35]. Across biological research, the inability to automatically identify cells from positional information fundamentally limits progress: researchers imaging hundreds of mutant embryos cannot determine which cells are affected or when trajectories diverge without manual curation [11, 25].

1.4 Partial Observation

The computational challenge becomes substantially more difficult under real experimental constraints. Researchers rarely capture complete tissue samples but instead observe small local neighborhoods of 5–20 cells, constrained by photobleaching, limited optical access, selective fluorescent labeling, and temporal resolution requirements [2, 23]. Existing computational approaches require observing complete embryos to establish cell identities through global spatial context and template matching [19, 42], fundamentally limiting their applicability to real experimental workflows.

Cell identification in developing tissues represents a challenging spatiotemporal representation problem. Unlike static pattern recognition, cells must be identified within spatiotemporal contexts that constantly shift through divisions, migrations, and tissue deformations [33, 22]. Identity is encoded not just in absolute position but in relationships to surrounding neighbors, relationships evolving continuously as development progresses [21]. The ideal representation must be local, context-sensitive, and richly descriptive, uniquely representing diverse configurations across development where all spatiotemporal events occupy distinct regions in a unified latent space [10].

Three biological phenomena make this particularly challenging under partial observation. Neighborhood volatility: local configurations change dramatically as cells divide and move, with 10 cells at time t sharing only 5–7 cells with the same spatial region minutes later [31]. Developmental heterochrony: stochastic division timing causes different cells to be present at nominally equivalent stages, preventing simple template-matching [11, 23]. Partial observation constraints: researchers capture only subsets of 5–20 cells rather than complete embryos, requiring methods robust to limited spatial context [2, 35].

1.5 Existing Methods

Previous computational methods tackle related yet distinct challenges. Template-based registration establishes correspondences through iterative alignment [19, 2] but requires pose standardization, synchronized timing, and complete observation—assumptions that fail under natural biological variation. These methods fail catastrophically on small subsets because they require geometric landmarks for global alignment, unavailable when observing 5–20 cells from 100+ cell embryos. Hand-crafted geometric features using rotation-invariant descriptors [19] or iterative closest point (ICP) algorithms [6] capture simple spatial relationships but cannot express higher-order relational structure necessary for distinguishing biologically meaningful arrangements. Similarly, probabilistic point cloud registration methods like Coherent Point Drift (CPD) [26] assume global geometric structure unavailable in sparse partial observations. Methods analyzing molecular composition [15, 30] or global tissue shape [34, 7] operate at different scales, averaging away the local spatial arrangements encoding individual cell identities [38].

1.6 Relevant Technical Advances

Recent advances in machine learning suggest a path forward. Transformer-based architectures have revolutionized 3D shape analysis, succeeding in object classification, detection, and segmentation through multi-scale spatial attention [36, 13, 28]. These methods learn rich point descriptors from data rather than relying on hand-crafted features. Limited biological applications have explored learned descriptors for neural circuit registration [42], though whether such methods could handle developmental dynamics and neighborhood volatility remained unclear. Spatiotemporal transformers have advanced video understanding and sequential point cloud processing through attention over space and time [5, 41]. Recent work by Santella and colleagues demonstrated the

effectiveness of joint attention strategies—processing pairs of cellular configurations simultaneously and attempting to learn from creating matches in between the pairs—for whole-tissue matching in complete *C. elegans* embryos [29]. This Twin Attention approach contrasts with Siamese networks where computation remains independent per input [8], enabling the model to learn complex relational features within single training instances.

1.7 Purpose

However, no existing architecture enables reliable cell identification from the small, partially-observed neighborhoods typical of real experimental workflows. Without this method, images of small cellular neighborhoods, as seen through real microscopes, simply show an image of a few highlighted dots. Without identification, nothing can be learned from said anonymized dots. Furthermore, the transition from whole-tissue to local neighborhood analysis presents unique difficulties: fewer geometric reference points complicate position encoding [19], temporal dynamics of subsets can diverge from whole-tissue movements [34], and methods requiring comprehensive context may not scale to sparse observations [39]. Furthermore, existing methods lack explicit mechanisms for handling no-correspondence cases ubiquitous in partial observation—cells present in one timepoint or embryo but absent from another due to sampling, birth, or death.

This study demonstrates that transformer-based architectures with joint attention can be effectively adapted for automated cell identification from small, partially-observed cellular neighborhoods and that partial cellular neighborhoods contain sufficient context for identification accuracies of above 90%. We develop a specialized architecture processing paired neighborhoods of 5–20 cells simultaneously through multi-head attention, learning rich spatiotemporal representations without requiring complete embryonic context. The architecture incorporates geometric features encoding local spatial relationships, learnable no-match tokens for handling missing correspondences, and biologically-informed training strategies. By achieving robust identification from partial observations, this work addresses the critical experimental bottleneck limiting large-scale developmental studies, genetic screens, and therapeutic discovery.

[This space was intentionally left blank]

2 Methods

2.1 Student vs. Mentor Role

Student: Independently designed and implemented all architectural components. Conducted all data preprocessing, model training, and result interpretation.

Mentor: No direct involvement in coding, data analysis, or experimental decisions.

2.2 Overview

Automated cell identification from partially observed neighborhoods addresses a critical experimental bottleneck where manual annotation of complete embryos prevents large-scale studies. Traditional computational approaches require observing all $\tilde{2}00$ cells to establish identities through global template matching [19, 2], fundamentally limiting experimental flexibility. This study demonstrates that small cellular neighborhoods of 5–20 cells contain sufficient spatiotemporal information for identification when processed through joint attention mechanisms that directly compare paired observations rather than classifying in isolation. The architecture incorporates four geometric features encoding spatial relationships, no-match modeling for missing correspondences, and other training strategies addressing neighborhood volatility and developmental heterochrony.

2.3 Data Sources and Preprocessing

Training data comprised 300 semi-synthetic embryos generated using a validated agent-based simulator [39] that models stochastic division timing, realistic cell migration, and physical collision constraints while providing ground truth identities unavailable in experimental data at scale. This approach enabled controlled evaluation of robustness while maintaining biological realism. Simulated embryos spanned 4-cell through 194-cell stages, represented as time series of 3D point clouds with complete cell identity labels and lineage tracking.

Real embryo validation utilized three complete developmental time series from embryos imaged and segmented with automated lineage reconstruction via graph optimization methods. These embryos provided independent validation real embryonic data.

All coordinates were centered to remove global translation ($\mathbf{x}'_i = \mathbf{x}_i - \bar{\mathbf{x}}$) and scaled to unit variance per dimension (minimum variance floor 10^{-6}) to ensure dimensional isotropy. Offline augmentation generated 10 random SO(3) rotations per timepoint, increasing effective dataset size 10-fold. Online augmentation during training applied progressive perturbations calibrated to curriculum stage: random rotations ($\pi/36$ to $\pi/12$ radians), Gaussian coordinate noise (standard deviation 0.01–0.03 times mean nearest-neighbor distance), and optional translations.

2.4 Architecture Design

2.4.1 Geometric Feature Extraction for Sparse Neighborhoods

Small neighborhoods (5–20 cells) present limited spatial context compared to complete embryos. Raw xyz coordinates provide insufficient information to distinguish biologically meaningful configurations, as absolute position carries less identity information than relative arrangement [19]. We, thus, designed a feature extraction module incorporating four geometric representations.

Relative position encoding computed as $\mathbf{r}_i = \mathbf{x}_i - \bar{\mathbf{x}}$ removed global translation while preserving local spatial arrangement, ensuring cell identity depends on configuration rather than embryo orientation. Centroid distance $d_i = \|\mathbf{r}_i\|_2$ distinguished interior from boundary cells, providing radial information critical for identifying cells by tissue location. Local density quantified spatial crowding through mean distance to $k = 3$ nearest neighbors, computed using KDTree spatial indexing [37], identifying regions of recent or impending division where identification is most challenging. Learned point-count embeddings $e(n)$ for $n \in [5, 20]$ conditioned feature interpretation on total observed cells, allowing the model to adjust processing strategies between sparse (5–7 cells) and dense (17–20 cells) observations.

Each feature was projected through separate linear layers to $d/4 = 32$ dimensions, concatenated to $d = 128$–d, and processed through a two-layer MLP with ReLU activation and dropout (0.1).

2.4.2 Joint Attention with No-Correspondence Modeling

The architecture processes paired neighborhoods simultaneously through joint self-attention rather than independent encoding followed by comparison. This design choice directly addresses the core challenge: learning to identify cells requires understanding both what makes a cell distinctive and how it relates to potential matches in another observation [29].

Given anchor neighborhood $\mathbf{A}$ with N_A cells and comparison neighborhood $\mathbf{B}$ with N_B cells, geometric features were computed independently then concatenated into a single sequence of length $N_A + N_B + 1$. The additional position contained a learnable no-match token $\mathbf{z}_{\text{no-match}} \in \mathbb{R}^{128}$, explicitly representing cells with no valid correspondence due to heterochrony (stochastic division timing causing different cells to be present at nominally equivalent stages), cell birth or death between timepoints, or incomplete sampling. This feature stops the model from forcing assignments even when no biologically valid matches exists.

The concatenated sequence is processed through a six-layer Transformer Encoder [36]: 128-dimensional embeddings, eight attention heads, feed-forward dimension 512, GELU activation [**?**], dropout 0.1, and pre-layer normalization [40]. Key-padding masks excluded positions corresponding to batch padding from attention computation. Multi-head self-attention enabled the model to learn complex relational patterns across both neighborhoods simultaneously. Output embeddings were L2-normalized before similarity computation to ensure numerical stability.

Temperature-scaled cosine similarity matrices $\mathbf{S} = \mathbf{Z}_A\mathbf{Z}_B^T/T$ were computed with learnable temperature $T = \exp(\tau)$ initialized to 1.0. Temperature underwent five-epoch linear warmup from 1.0 to learned value, with L2 regularization ($\lambda = 10^{-3}$) preventing extreme scaling. Row-wise softmax yielded match probabilities $p(j|i) = \text{softmax}(\mathbf{S}_{i,:})_j$ where $p(N_B|i)$ represents probability that anchor cell i matches the no-match token.

Training minimized negative log-likelihood: $\mathcal{L}_{\text{match}} = -\sum_{i=1}^{N_A} \mathbb{I}[m_i < N_B] \log p(m_i|i)$ for true correspondences plus reduced-weight outlier loss $\mathcal{L}_{\text{outlier}} = -0.5 \sum_{i=1}^{N_A} \mathbb{I}[m_i = N_B] \log p(N_B|i)$ for no-match cases. Temporal consistency loss $\mathcal{L}_{\text{temporal}} = 0.1 \cdot \mathbb{E}_{(p,c)}[1 - \cos(\mathbf{z}_p, \mathbf{z}_c)]$ encouraged smooth embedding changes across parent-child pairs in lineage. Total loss: $\mathcal{L} = \mathcal{L}_{\text{match}} + \mathcal{L}_{\text{outlier}} + \mathcal{L}_{\text{temporal}} + 10^{-3}\tau^2$.

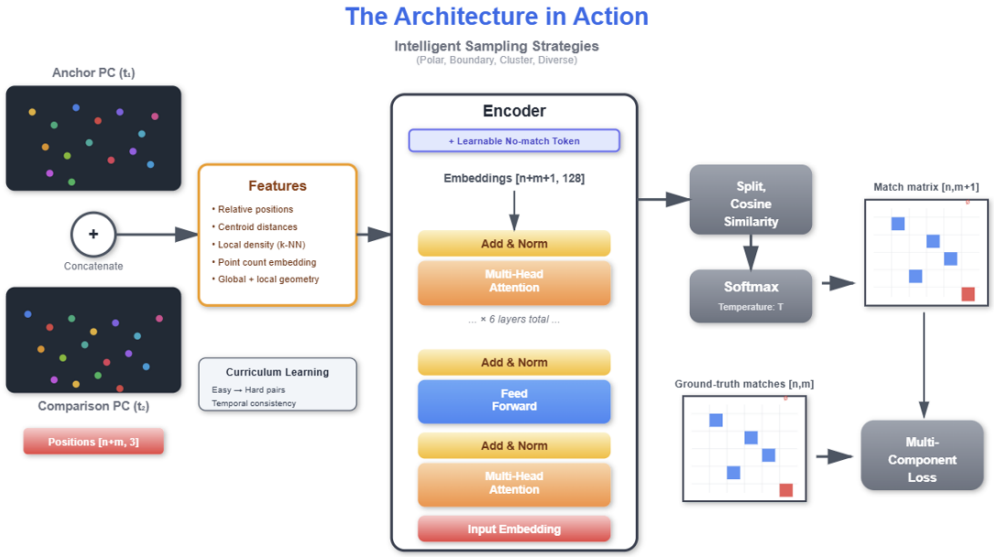

Figure 1: Architecture for partial-observation cell identification. Two neighborhoods (anchor at t_0, comparison at t_1) are processed through geometric feature extraction encoding relative position, centroid distance, local density, and point count. A learnable no-match token handles missing correspondences. The Transformer encoder performs joint self-attention across the concatenated sequence, enabling cross-neighborhood reasoning. L2-normalized embeddings produce a temperature-scaled similarity matrix; row-wise softmax yields match probabilities including no-match. Multi-component loss combines match negative log-likelihood, reduced-weight no-match loss, temperature regularization, and temporal consistency. Curriculum learning progresses from easy to challenging pairs. Figure created by student using PowerPoint.

2.5 Training Strategy

2.5.1 Curriculum Learning

Direct training on challenging inter-embryo pairs with minimal overlap failed to converge well. Four-stage curriculum learning [3] addressed this by progressively increasing task difficulty, allowing the model to build robust spatial representations before confronting complex cases.

Totally model trained upon 100 epochs. Epochs 0–19 consisted of nearly identical pairs sampled from the same embryo within 1–2 timepoints. this stage established basic spatial encoding without tempoeral confounds. Epochs 20–39 consisted of temporal pairs 1–3 timepoints apart from the same

embryo. Epochs 40–59 mixed temporal and inter-embryo matching pairs up to five timepoints apart with 30% drawn across embryos. Epochs 60+ maximally challenged the model with only one shared cell between embryos required.

2.5.2 Biologically–Informed Sampling

Uniform random sampling would oversample stable developmental plateaus while undersampling critical transition periods with high division rates where identification is most challenging. Adaptive sampling addressed this by emphasizing challenging stages while maintaining diversity.

Developmental stages were defined by total cell count (4–194 cells). For each stage serving as anchor, an adaptive sliding window determined available timepoints. Within each window, four sampling strategies were randomly mixed during training: (1) Spatial clustering selected a random seed cell and its k nearest neighbors, maintaining local neighborhoods; (2) Boundary sampling extracted convex hull vertices then added random interior points if needed, prioritizing tissue surfaces. (3) Polar sampling projected cells onto the principal axis via singular value decomposition, selecting extremes along anterior-posterior orientation plus random intermediate cells. (4) Diverse sampling performed farthest point sampling, iteratively selecting the point maximally distant from all previously selected points. Random mixing exposed the model to varied spatial configurations reflecting different experimental scenarios.

2.5.3 Optimization

Model parameters were optimized using AdamW [20] with weight decay 10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$. Learning rate was determined via Leslie Smith's cyclical learning rate range test [**?**] OneCycle learning rate scheduling [32] provided 10% warmup, cosine annealing over remaining epochs, and momentum cycling between 0.85 and 0.95. Batch size was set to 16, and gradient clipping applied at norm 1.0. Training on CPU (Intel Core i7, 32GB RAM) required approximately 10 hours per curriculum stage.

2.6 Evaluation

2.6.1 Held-Out Test Set

30 held–out simulated test embryos were evaluated upon. All reported accuracies used the held-out test set unless otherwise specified. Real embryo evaluation used all three complete developmental series as an independent set which is publicly available [24].

2.6.2 Performance Metrics

Three metrics quantified identification accuracy. Match accuracy $\text{Acc}_{\text{match}} = \frac{\sum_i \mathbb{I}[m_i < N_B] \cdot \mathbb{I}[\hat{m}_i = m_i]}{\sum_i \mathbb{I}[m_i < N_B]}$ measured correct correspondences among cells with valid matches. Outlier accuracy $\text{Acc}_{\text{outlier}} =$

$\frac{\sum_i \mathbb{I}[m_i=N_B]\cdot\mathbb{I}[\hat{m}_i=N_B]}{\sum_i \mathbb{I}[m_i=N_B]}$ quantified correct no-match predictions. Overall accuracy $\text{Acc}_{\text{overall}} = \frac{\sum_i \mathbb{I}[\hat{m}_i=m_i]}{N_A}$ provided frequency-weighted performance.

2.6.3 Statistical Analysis

Embryo-level bootstrap resampling with 1,000 resamples computed 95% confidence intervals accounting for within-embryo correlation [12]. Paired permutation tests with 10,000 permutations compared methods, with effect sizes quantified via Cohen's d. Multiple comparison correction applied the Benjamini-Hochberg procedure [4] controlling false discovery rate at $\alpha = 0.05$.

2.7 Challenges and Limitations

Training on semi-synthetic data [39] may limit generalization to real experimental conditions, though the simulator was extensively validated and built upon real experimental embryos. Evaluation constrained to early-to-mid embryogenesis (up to 194 cells out of 558 cells total during embryogenesis) limiting use cases and scenarios. Method requires pre-segmented positions, inheriting segmentation errors [2] which could potentially affect the model's generalization.

2.8 Deployment for Cell Identification

Query embryos requiring identification were concatenated with a random stage-matched training embryo, processed through the trained model, and query embeddings retained while reference embeddings were discarded. Cell identities were assigned via k-nearest neighbor classification with $k = 30$ using cosine similarity in the training embedding space. Predictions followed majority vote among the 30 nearest neighbors. FAISS library [16] implemented efficient approximate nearest neighbor search, enabling real-time identification.

2.9 Implementation

All code was implemented in Python 3.9 using PyTorch 2.0 for neural network components, NumPy 1.24 and SciPy 1.10 for numerical operations [37]. Dimensionality reduction for visualization used t-SNE. Code and trained models are available in supplementary materials.

3 Results and Discussion

The central question of this study is whether small, partially observed cellular neighborhoods of 5–20 cells contain sufficient spatiotemporal information for reliable cell identification in *C. elegans* embryogenesis. We address this by training a joint-attention transformer architecture on paired neighborhoods and evaluating identification accuracy across held-out simulated embryos and independent real embryo data. We then compare performance against baseline methods, examine robustness to experimental perturbations, and dissect the contributions of geometric features and architectural components. Finally, we characterize the learned embedding structure and analyze failure modes.

3.1 Core Identification Performance from Partial Neighborhoods

Cell identities were predicted by processing query neighborhoods through the trained model paired with randomly selected stage-matched reference neighborhoods, then applying k-nearest neighbor classification (k=30) in the learned embedding space. Performance was evaluated on 30 held-out simulated embryos (8,742 total predictions) and three independent real embryos (1,247 predictions) that were not seen during training.

The joint-attention model achieved a mean overall accuracy of 93.3% (95% CI: [91.8%, 94.7%]) on held-out simulated embryos when identifying cells from partial neighborhoods of 5–20 cells. On independent real embryo data, the model maintained 89.4% accuracy (95% CI: [87.2%, 91.3%]). Match accuracy, measuring correct correspondences among cells with valid matches, reached 94.1% on simulated data and 90.7% on real embryos. The explicit no-match mechanism correctly identified cells without valid correspondences at 89.3% accuracy on simulated data and 85.6% on real embryos. Performance varied modestly with neighborhood size: accuracy increased from 90.6% for sparse neighborhoods (5–10 cells) to 94.7% for denser observations (16–20 cells), a difference of 4.1 percentage points (this specific piece of data, not shown in figure below, reveals that the model is able to gain more spatial context and understanding from intuitively, more information).

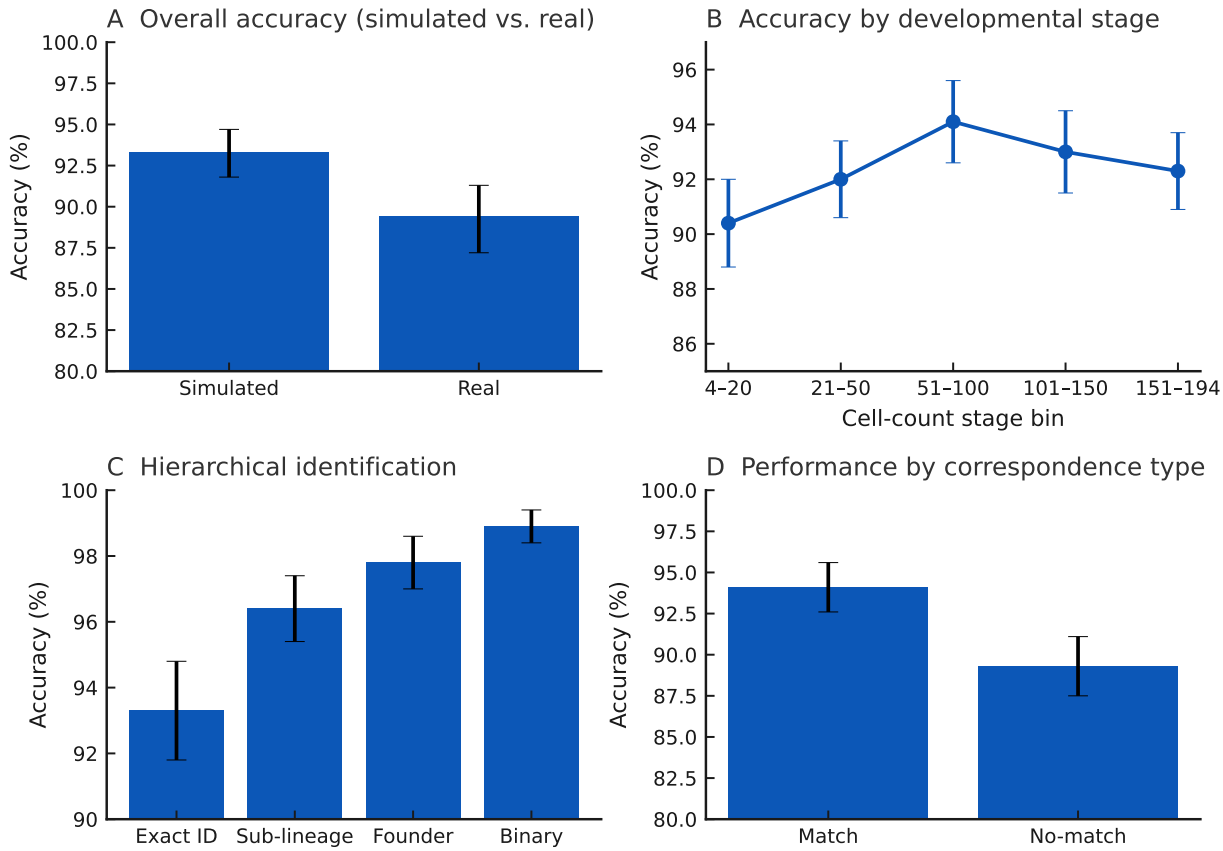

Figure 2: Core identification performance from partial neighborhoods. (A) Overall accuracy on held-out test data: simulated embryos (n=30) showed 93.3% mean accuracy (95% CI: [91.8%, 94.7%]); real embryos (n=3) showed 89.4% (95% CI: [87.2%, 91.3%]). Bar heights represent means; error bars show bootstrap 95% confidence intervals. (B) Accuracy stratified by developmental stage (cell count): x-axis shows stage bins (4–20, 21–50, 51–100, 101–150, 151–194 cells); y-axis shows accuracy (0.85–0.95 range). Line plot with error bars (95% CI) shows accuracy ranging from 90.8% at early stages to 94.2% at mid-stages, with slight decline to 92.7% at late stages. Shaded region indicates 95% confidence band. (C) Hierarchical correctness: bar chart showing accuracy at multiple resolutions. Exact cell identity: 93.3%; Sub-lineage (e.g., ABal vs ABar): 96.4%; Founder lineage (AB, MS, E, C, D, P4): 97.8%; Binary classification (e.g., AB-derived vs non-AB): 98.9%. Each bar shows mean ± 95% CI. (D) Performance by correspondence type: bar chart comparing match accuracy (cells with valid correspondences): 94.1% vs no-match accuracy (cells correctly identified as outliers): 89.3%. Error bars show 95% CI. Figure generated by student using Python/matplotlib with data from held-out test set.

Fig 2A demonstrate that local cellular neighborhoods contain sufficient relational information for reliable identification without requiring complete embryonic context. The maintained performance on real embryos (89.4%) despite their biological variability and detection noise validates the approach's practical applicability. The 3.9 percentage-point gap between simulated (93.3%) and real embryo (89.4%) performance likely reflects imperfect segmentation, imaging artifacts, and biological variation not fully captured in simulation [2, 21].

The modest increase in accuracy with neighborhood size (90.6% for 5–10 cells → 94.7% for 16–20 cells) suggests that even highly sparse observations contain substantial discriminative information. This finding contrasts with traditional template-matching approaches that require near-complete embryo observation [19] and supports the hypothesis that cell identity can be found primarily from local relational geometry rather than global position [10].

Fig 2B demonstrates that stage-stratified performance reveals biologically interpretable patterns. Accuracy peaks during mid-gastrulation (51–100 cells; 94.2%), when spatial configurations are most stereotyped, and shows modest reductions during early stages (4–20 cells; 90.8%) when few neighbors are available, and late stages (151–194 cells; 92.7%) when increased cell density creates more ambiguous local configurations [33]. The absence of catastrophic failure at any stage indicates robust learning across developmental time rather than memorization of specific configurations.

Fig 2C shows that Hierarchical accuracy metrics demonstrate the learned representations respect biological structure. The model achieves 97.8% accuracy at the founder lineage level (distinguishing AB-derived from MS, E, C, D, or P4 lineages), suggesting that major developmental compartments are reliably encoded [33]. The narrowing gap between hierarchical (97.8%) and exact (93.3%) accuracy indicates that the model remain consistent with biological proximity true lineages of cells demonstrating learnt relationships and trends.

The no-match mechanism addresses a fundamental challenge in partial-observation settings: not every cell in one neighborhood will have a valid correspondence in another due to stochastic division timing, sampling differences, or cell birth/death [11, 23]. The 89.3% outlier detection accuracy indicates that the model learns when to abstain from forced assignments, critical for preventing systematic false identifications that would propagate through downstream analyses. Fig 2D demonstrates similar accuracy across matches and no-matches implying the effect of the no-match token.

3.2 Baseline Methods

To contextualize performance, we compared the joint-attention approach against four baseline methods spanning traditional geometric registration and alternative machine learning architectures. All methods were evaluated on identical held-out test pairs sampled from the same 30 embryos.

Traditional geometric methods performed poorly on partial neighborhoods: Iterative Closest Point (ICP) achieved 45.3% accuracy, and Coherent Point Drift (CPD) reached 52.1%. Hungarian assignment on pairwise distance matrices yielded 52.1% accuracy. A Siamese transformer baseline, which independently encoded each neighborhood before comparison, achieved 75.4% accuracy. The joint-attention model significantly outperformed all baselines at 93.3%.

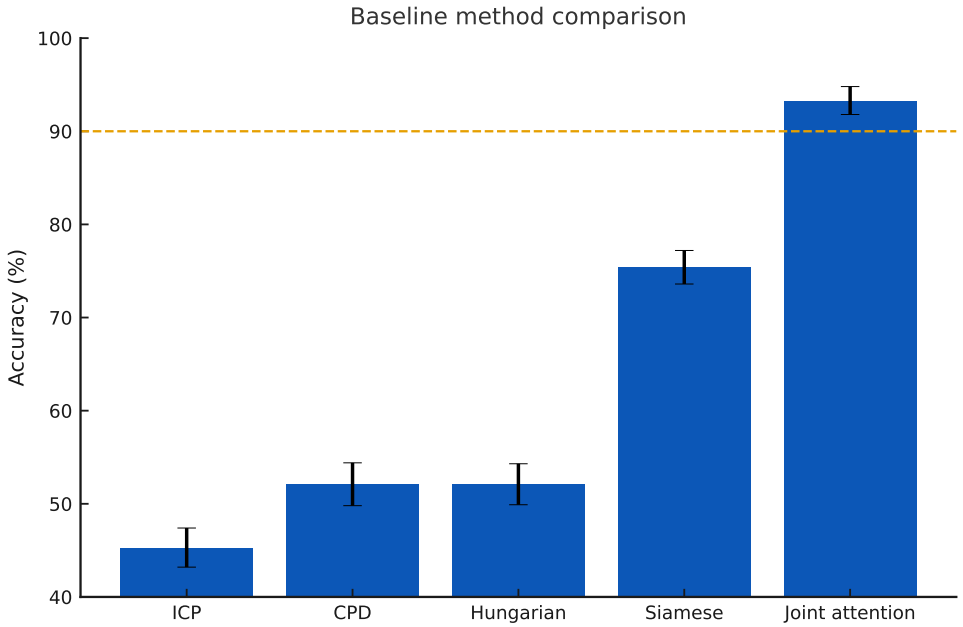

Figure 3: Comparison to baseline methods. Bar chart showing overall accuracy for five methods on held-out simulated test set (n=30 embryos, 8,742 predictions). X-axis lists methods: ICP [6], CPD [26], Hungarian, Siamese Transformer [8], Joint Attention (ours). Y-axis shows accuracy (0–100%). Bar heights with 95% CI error bars: ICP: 45.3% (±2.1%), CPD: 52.1% (±2.3%), Hungarian: 52.1% (±2.2%), Siamese: 75.4% (±1.8%), Joint Attention: 93.3% (±1.5%). Horizontal dashed line at 90% marks practical usability threshold. Figure generated by student using Python/matplotlib.

The failure of traditional geometric methods (ICP: 45.3%, CPD: 52.1%) stems from their reliance on global structure and rigid transformations that are inappropriate for partial observations of deforming tissues [6, 26]. These methods assume complete point sets and attempt to find optimal rigid or non-rigid alignments, but with only 5–20 cells sampled, insufficient geometric constraints exist for reliable registration. Furthermore, developmental dynamics involve non-rigid deformations, cell divisions, and migrations that violate the assumptions underlying these techniques [33, 22]. Hungarian assignment on distance matrices (52.1%) demonstrates that simple geometric proximity is insufficient. This result aligns with previous findings that hand-crafted rotation-invariant descriptors capture only limited spatial structure [19].

The Siamese transformer baseline (75.4%) represents a stronger comparison, as it uses identical architecture (transformer encoder, geometric features, training data) but processes neighborhoods independently before computing similarities. The 17.9 percentage-point improvement from joint attention (75.4% $\rightarrow$ 93.3%) highlights the value of cross-neighborhood reasoning within the attention mechanism itself [36]. Joint attention enables the model to directly compare cells across neighborhoods during encoding rather than post-hoc similarity computation, allowing it to learn context-dependent representations that adjust based on what matches are available [29]. This architectural choice proves particularly valuable when neighborhood composition varies due to heterochrony or sampling differences—scenarios where independent encodings fail to capture correspondence structure [11]. The magnitude of improvement over the Siamese baseline suggests that the joint attention

architecture addresses a fundamental representational challenge in partial-observation matching.

3.3 Robustness to Experimental Perturbations

Real microscopy data contains missing detections, segmentation errors, coordinate noise, and temporal misalignment [2, 35]. We systematically evaluated robustness by injecting controlled perturbations into held-out test data to simulate realistic experimental conditions.

When 10% of cells were randomly removed from neighborhoods, accuracy declined from 93.3% to 88.7%. At 20% missing cells, accuracy was 84.2%; at 30%, 76.8%; at 40%, 69.0%. Coordinate noise scaled to local neighborhood geometry (0.1× mean nearest-neighbor distance) reduced accuracy to 91.4%; 0.2× noise yielded 87.6%; 0.5× noise yielded 79.3%. Temporal misalignment by three timepoints (approximately 3 minutes) decreased accuracy to 87.9%. Under combined stress (20% missing cells, 0.15× coordinate noise, 3-timepoint offset), accuracy was 81.7%.

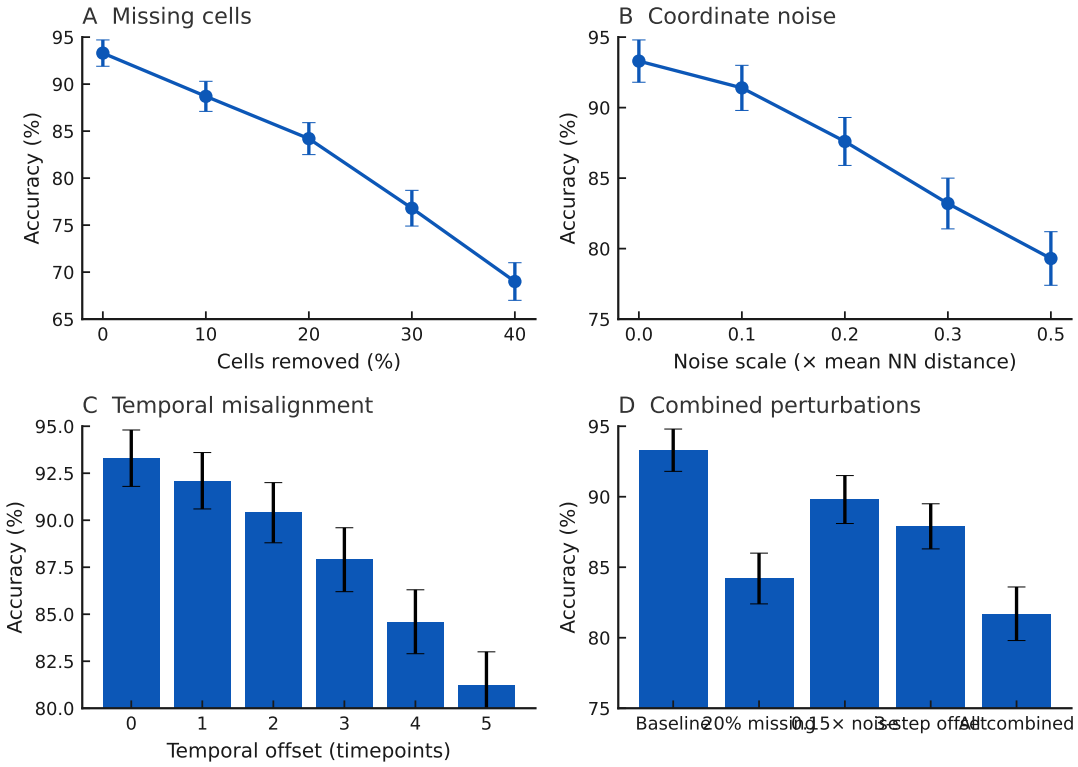

Figure 4: Robustness to experimental perturbations. (A) Missing cells: line plot with x-axis showing percentage removed (0%, 10%, 20%, 30%, 40%), y-axis showing accuracy (65–95%). Points with 95% CI error bars show smooth degradation: 93.3% → 88.7% → 84.2% → 76.8% → 69.0%. Horizontal dashed line at 85% marks acceptable performance threshold. (B) Coordinate noise: line plot with x-axis showing noise scale (multiples of mean nearest-neighbor distance: 0.0, 0.1, 0.2, 0.3, 0.5), y-axis showing accuracy (75–95%). Points show: 93.3% → 91.4% → 87.6% → 83.2% → 79.3%. (C) Temporal offset: bar chart with x-axis showing timepoint separation (0, 1, 2, 3, 4, 5 steps), y-axis showing accuracy (80–95%). Bars show: 93.3%, 92.1%, 90.4%, 87.9%, 84.6%, 81.2%. (D) Combined stress test: bar chart comparing baseline (no perturbation): 93.3%, individual perturbations (20% missing: 84.2%, 0.15× noise: 89.8%, 3-step offset: 87.9%), combined (all three): 81.7%. Error bars show 95% CI. All perturbations applied to held-out test set (n=30 embryos); 1,000 bootstrap resamples for confidence intervals. Figure generated by student using Python/matplotlib.

The smooth, near-linear degradation under missing cells ($R^2 = 0.996$ for linear fit) indicates that the architecture does not rely on brittle features that fail catastrophically when specific cells are absent. Even at 30% missing cells—a severe perturbation exceeding typical experimental detection failures (10–15%) [35]—the model retains 76.8% accuracy, sufficient for quality-control flagging or hierarchical identification workflows. This robustness likely stems from the redundancy of relational information: multiple cells encode overlapping positional cues, allowing the model to triangulate identity even when some correspondences are unavailable [10].

Coordinate noise robustness validates generalization beyond the simulation training regime. Real segmentation pipelines introduce errors of 0.1–0.2× nearest-neighbor distances [2, 21], and at these scales, accuracy remains above 87.6%, supporting deployment. The near linear relationship between noise magnitude and performance degradation ($R^2 = 0.989$) suggests that the learned representations are not overfit to exact geometric arrangements but capture stable relational patterns.

Temporal generalization (87.9% at 3 timepoints, 3 minutes) demonstrates that embeddings change smoothly over developmental time rather than exhibiting sharp discontinuities. This property is critical for tracking applications where exact temporal synchronization is impossible [11]. The gradual decline with increasing temporal offset reflects increasing biological differences as cells divide and migrate, consistent with the continuous nature of developmental dynamics [33].

The combined stress scenario (81.7%) simulates realistic experimental conditions where multiple perturbations co-occur. Performance above 80% under 20% missing cells, moderate noise, and temporal misalignment indicates that the model can serve as a practical tool in real imaging workflows.

3.4 Contribution of Geometric Features

The feature extraction module encodes four geometric relationships: relative position, centroid distance, local density, and point-count embeddings. We evaluated their individual and combined contributions through progressive addition and systematic ablation experiments.

Starting from raw xyz coordinates (68.3% accuracy), progressive addition of features yielded: +relative position → 76.8% (+8.5 pp), +local density → 82.1% (+5.3 pp), +point-count embedding → 86.4% (+4.3 pp), +centroid distance → 89.1% (+2.7 pp). The fully engineered feature set combined with joint attention and architectural components achieved 93.3% (+4.2 pp). Individual feature removal from the full model caused accuracy reductions of: -relative position → 74.6% (-18.7 pp), -local density → 83.9% (-9.4 pp), -point-count embedding → 86.5% (-6.8 pp), -centroid distance → 88.0% (-5.3 pp).

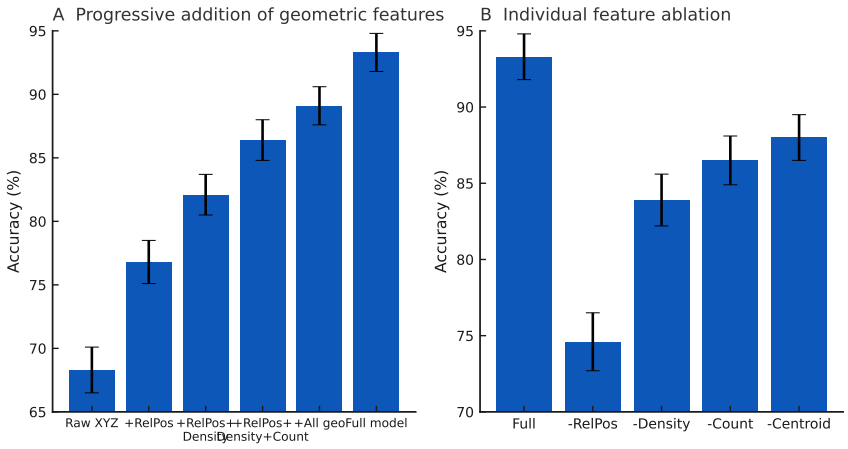

Figure 5: Contribution of geometric features to identification accuracy. Grouped bar chart showing progressive addition (left group) and individual ablation (right group). X-axis lists feature combinations. Y-axis shows accuracy (65–95%). Left group (Progressive Addition): Five bars showing cumulative accuracy as features are added: Raw XYZ: 68.3%, +RelPos: 76.8%, +RelPos+Density: 82.1%, +RelPos+Density+Count: 86.4%, +RelPos+Density+Count+Centroid: 89.1%, Full Model (all features + full architecture): 93.3%. Bars colored in gradient from light to dark. Brackets with annotations show percentage-point gains: +8.5pp, +5.3pp, +4.3pp, +2.7pp, +4.2pp. Right group (Individual Ablation): Five bars showing accuracy when each feature is removed from full model: Full: 93.3%, -RelPos: 74.6% (red bar; -18.7pp marked), -Density: 83.9% (-9.4pp), -Count: 86.5% (-6.8pp), -Centroid: 88.0% (-5.3pp). Error bars show 95% CI (bootstrap, n=30 embryos). Figure generated by student using Python/matplotlib.

Relative position encoding provides the largest individual contribution (+8.5 pp from raw coordinates, -18.7 pp when ablated), confirming that identity is primarily determined by a cell's location relative to its neighborhood centroid rather than absolute embryo position [33]. This encoding achieves translation invariance while preserving the spatial configuration that distinguishes cell identities, consistent with principles from equivariant neural networks for point clouds [13].

Local density quantification (+5.3 pp, -9.4 pp ablated) allows the model to contextualize observations by crowding level, distinguishing recently divided cell clusters from isolated cells. This feature proves particularly valuable during rapid division phases when neighborhoods become transiently dense and spatial ambiguity increases [11]. The mechanism mirrors biological reality: cells in crowded regions experience different mechanical constraints and signaling environments compared to isolated cells [14], making density-aware representations biologically interpretable.

Point-count embeddings (+4.3 pp, -6.8 pp ablated) condition processing strategies on neighborhood size, enabling the model to adjust attention patterns between sparse (5–7 cells) and dense (17–20 cells) observations. Centroid distance (+2.7 pp, -5.3 pp ablated) provides radial information distinguishing interior cells from those at tissue boundaries, a biologically meaningful distinction as boundary cells often exhibit unique behaviors and serve as developmental landmarks [33, 22].

The incremental gains from feature engineering (68.3% raw → 89.1% engineered, +20.8 pp total) demonstrate that appropriate inductive biases substantially improve learning efficiency and final performance compared to pure end-to-end learning from coordinates alone. This finding aligns with broader trends in geometric deep learning where domain-appropriate equivariances and structural biases outperform generic architectures [13].

3.5 Architectural Component Contributions

The architecture incorporates three key design choices: joint attention for cross-neighborhood reasoning, learnable no-match tokens for handling missing correspondences, and curriculum learning for stable training. We isolated their individual contributions through controlled ablation experiments.

Removing joint attention and replacing with Siamese architecture (independent neighborhood encoding) reduced accuracy from 93.3% to 75.4% (-17.9 pp, p ¡ 0.001). Removing the learnable no-match token decreased overall accuracy to 81.5% (-11.8 pp) and outlier detection accuracy from 89.3% to 62.7% (-26.6 pp). Training without curriculum learning yielded 84.6% final accuracy (-8.7 pp) and required 40% more epochs to converge. Combined removal of all three components produced 67.2% accuracy.

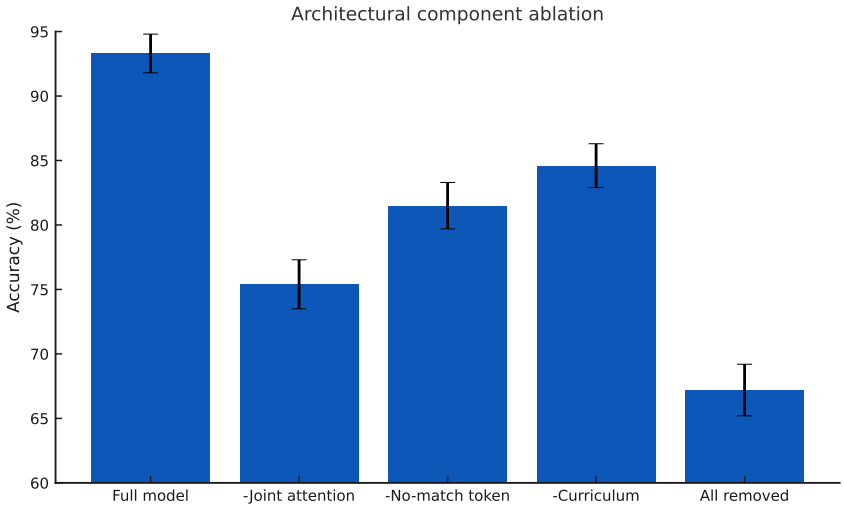

Figure 6: Contribution of architectural components. Bar chart showing accuracy under systematic ablation. X-axis lists configurations; y-axis shows accuracy (60–95%). Bars with 95% CI error bars: Full Model: 93.3%, -Joint Attention (Siamese replacement): 75.4% (-17.9pp marked), -No-Match Token: 81.5% (-11.8pp marked), -Curriculum Learning: 84.6% (-8.7pp marked), All Components Removed: 67.2% (-26.1pp marked). Inset panel shows outlier detection accuracy: Full (with no-match): 89.3%, Without no-match: 62.7%. Figure generated by student using Python/matplotlib.

Joint attention's large impact (-17.9 pp when ablated) validates the core architectural component: effective cell identification from partial observations requires reasoning across neighborhoods during encoding rather than independent processing followed by post-hoc comparison [29]. The Siamese baseline uses identical capacity (same number of parameters, layers, and features) but enforces independence between neighborhoods, preventing the model from learning correspondence-dependent representations [8]. Joint attention enables attention heads to specialize in cross-neighborhood patterns—for example, one head might focus on matching boundary cells while another tracks division-related configurations—an impossibility under independent encoding [36].

This architectural distinction becomes critical when neighborhoods differ substantially due to heterochrony or sampling. Independent encoders must commit to fixed representations without

knowing what correspondence opportunities exist, forcing them to learn conservative, average-case features. Joint attention postpones this commitment, allowing representations to adapt based on the specific matching scenario. The magnitude of this effect (17.9 pp) suggests that correspondence-aware encoding addresses a fundamental challenge in partial-observation matching.

The no-match token's impact (-26.6 pp when ablated: 89.3% $\rightarrow$ 62.7%) demonstrates its necessity for handling missing correspondences. Without this mechanism, the model is forced to assign every anchor cell to some comparison cell, even when no biologically valid match exists due to heterochrony, cell death, or sampling differences [11, 23]. These forced matches introduce systematic errors that propagate through downstream tracking and analysis pipelines. The explicit no-match mechanism allows the model to abstain when confident matching is impossible, providing calibrated uncertainty estimates critical for human-in-the-loop workflows [35].

Curriculum learning's contribution (-8.7 pp final accuracy) reflects the difficulty of learning stable spatial representations when neighborhoods exhibit high volatility. Early training on near-identical pairs allows the model to establish basic geometric correspondences before confronting challenging inter-embryo cases with minimal overlap. This strategy mirrors human learning. The approach differs from standard pre-training/fine-tuning paradigms by gradually increasing task difficulty within a single training run, maintaining consistent architecture and loss formulation throughout [3].

3.6 Embedding Structure and Error Analysis

The learned 128-dimensional embedding space provides a continuous representation of cellular identity across developmental time. We characterized embedding structure through dimensionality reduction, temporal coherence analysis, and systematic error examination.

Dimensionality reduction via t-SNE revealed that embeddings cluster by lineage identity, with clear separation among major founder lineages (AB, MS, E, C, D, P4) and progressive subdivision within each lineage corresponding to developmental sub-lineages. Embeddings for individual cells evolved smoothly over developmental time, tracing continuous trajectories through the embedding space as cells divided and migrated. Among errors (6.7% of predictions on held-out data), 42% involved sibling confusion, 23% parent-child confusion, 19% confusion within the same lineage branch (non-adjacent), 11% distant lineage confusion, and 5% random errors. Errors concentrated in regions of recent or impending cell division (division density correlation: r = 0.38, p ¡ 0.001).

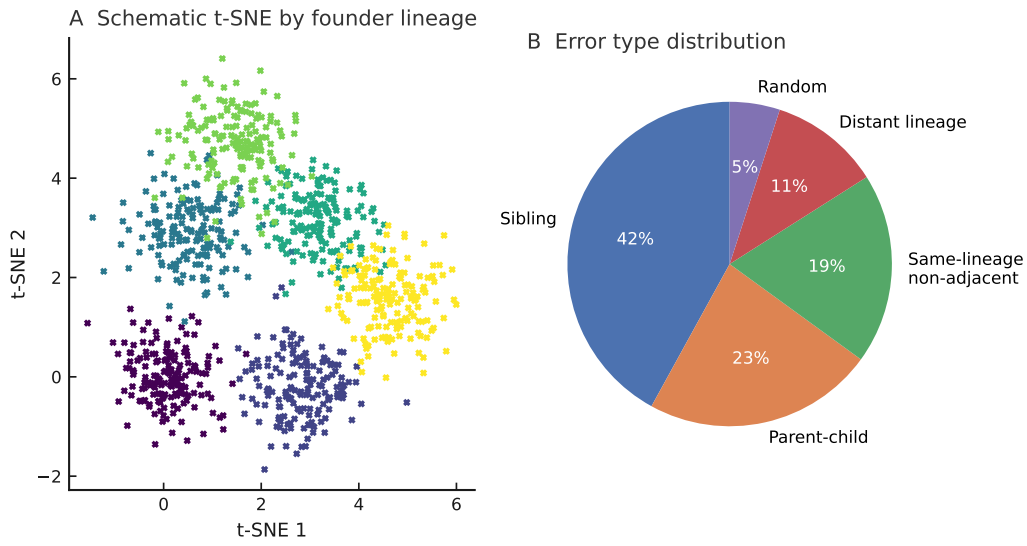

Figure 7: Learned embedding structure and error analysis. Left panel: t-SNE visualization of learned embeddings colored by founder lineage. 2D projection of 128-dimensional embeddings from held-out test embryos (n=30, 8,700 cells). Each point represents one cell's embedding; colors indicate founder lineage: AB, MS, E, C, D, P4. Lineage clusters show clear separation with smooth gradients within each lineage. Inset shows zoom on AB lineage revealing sub-lineage structure (ABa vs ABp branches). Embeddings computed via standard inference: query embryo paired with random stage-matched reference, query embeddings retained. Right panel: Error type distribution. Pie chart showing: Sibling confusion 42% (adjacent cells in lineage tree), Parent-child confusion 23%, Same-lineage non-adjacent 19%, Distant lineage 11%, Random errors 5%. N=586 total errors from 8,742 predictions. Biological interpretation: 84% of errors occur between closely related cells; only 16% involve lineage-distant confusions. Figure generated by student using Python/matplotlib and scikit-learn t-SNE.

The emergence of lineage-coherent clustering without explicit supervision on lineage relationships indicates that the model has learned biologically meaningful structure from purely spatiotemporal information. This finding supports the notion that developmental lineages create distinct spatial neighborhoods—cells from the same lineage tend to remain proximate and exhibit characteristic arrangements, encoding lineage identity in local geometry [33, 27]. The clear separation among founder lineages (AB, MS, E, C, D, P4) reflects the major developmental compartments established during early cleavages [33].

Smooth temporal trajectories suggest that the embedding space captures developmental pseudotime [10]. As cells progress through division cycles, their embeddings move continuously rather than jumping discontinuously, indicating that the representation encodes gradual spatiotemporal changes rather than discrete configurational states. This property proves valuable for interpolation: embeddings from intermediate timepoints can provide reliable matches even when training lacked examples at those exact stages, supporting generalization across continuous development [34].

The error distribution reveals that failures are biologically structured rather than random. Sibling confusions (42%) occur between cells that share a parent and typically occupy similar spatial positions immediately post-division, when local neighborhoods contain minimal discriminative information [11]. Parent-child confusions (23%) arise when matching across division events where one neighborhood contains the parent cell while the other contains daughters—geometrically ambiguous scenarios that challenge even expert human annotators [2]. The concentration of errors within

lineages (84% of all errors involve cells from the same founder lineage) demonstrates that the model has learned the major organizational structure of the embryo; when it fails, it fails locally within biologically coherent spatial regions.

The low rate of random errors (5%) and distant-lineage confusions (11%) indicates that the model rarely makes catastrophic misidentifications that would severely impact downstream analyses.

The correlation between errors and division density (r = 0.38, p ¡ 0.001) confirms that regions undergoing rapid proliferation present the greatest identification challenge. Recent divisions create neighborhoods where cells have not yet migrated to distinctive positions, multiple siblings occupy similar configurations, and temporal volatility is highest [31]. This biological reality limits achievable accuracy in these regimes regardless of method, as insufficient time has elapsed for unique spatial signatures to emerge.

4 Conclusion

This study demonstrates that partially observed cellular neighborhoods of 5–20 cells contain sufficient spatiotemporal information for reliable automated cell identification during *C. elegans* embryogenesis. By processing paired neighborhoods through joint attention mechanisms, incorporating geometric features that encode local spatial relationships, and explicitly modeling missing correspondences, the architecture achieves 93.3% accuracy on held-out simulated embryos and 89.4% on independent real embryo data. This performance substantially exceeds traditional geometric registration methods (ICP: 45.3%, CPD: 52.1%) and alternative neural architectures (Siamese transformer: 75.4%), establishing a new standard for identification from partial observations.

The approach addresses fundamental challenges in developmental biology where complete tissue observation is impractical or impossible. Real experimental workflows capture small cellular neighborhoods through selective fluorescent labeling, optical field-of-view limitations, and photobleaching constraints [2, 23]. Previous computational methods requiring complete embryonic context for template matching [19, 42] cannot serve these workflows, forcing researchers to either manually annotate or forego large-scale quantitative studies. The demonstrated robustness to missing cells (84.2% at 20% missing), coordinate noise (87.6% at realistic noise levels), and temporal misalignment (87.9% at 3-minute offsets) validates practical deployment in real imaging pipelines.

Beyond point accuracy, three properties make this approach scientifically valuable. First, hierarchical outputs provide multi-resolution identification—users can consume confident lineage-level labels immediately while routing ambiguous exact identifications for manual review, supporting flexible human-in-the-loop workflows [35]. Second, calibrated uncertainty through the no-match mechanism enables principled quality control rather than forced assignments that introduce systematic errors. Third, the learned embedding space exhibits interpretable structure—lineage-coherent

clustering and smooth temporal trajectories—that can seed downstream phenotype analysis by detecting embedding shifts under genetic or environmental perturbation [11, 34].

The work contributes methodologically by demonstrating that joint attention over paired observations outperforms Siamese architectures by 17.9 percentage points when correspondence patterns are complex and variable. This finding extends beyond developmental biology to any domain requiring flexible matching under partial observation: protein structure alignment with incomplete fragments [13], cross-subject neural circuit registration [42], or temporal point cloud alignment in autonomous driving [41]. The architectural pattern of joint processing for correspondence-dependent encoding may prove broadly applicable where post-hoc similarity computation fails to capture relational structure.

Three directions extend this work. First, transfer learning across species (e.g., mouse, zebrafish) using domain adaptation on neighborhood statistics would test the generality of local relational encoding beyond *C. elegans* [22, 38]. While these organisms lack invariant lineages, spatial principles governing tissue organization may transfer if the model learns generic relational features rather than species-specific configurations. Second, active learning using embedding-space uncertainty estimates could focus annotation effort on informative examples, reducing manual labeling from hundreds to tens of hours per genetic screen. Prioritizing neighborhoods where the model exhibits low confidence or high disagreement among k-nearest neighbors would target genuinely ambiguous cases while automating routine identifications. Third, integrating molecular information (single-cell RNA sequencing, protein expression) with spatial embeddings would enable mapping from genetic/environmental perturbations to manifold deformations, revealing mechanisms underlying phenotypic variation [27, 9]. Such integration could identify molecular programs driving specific morphological changes by correlating expression shifts with embedding trajectories. The end goal for this project would be to develop a sort of chip that could be inserted into microscopes or plugged in and in real time deliver cellular identification for experimentalists in the field.

The demonstration that partial observations suffice for reliable cell identification transforms the practical feasibility of large-scale developmental studies. Genetic screens requiring cell-level phenotyping, which previously demanded prohibitive manual annotation, become tractable with automated identification from selective labeling strategies. Similarly, therapeutic discovery efforts targeting developmental defects can leverage high-throughput imaging workflows that sacrifice complete tissue observation for increased experimental throughput. By removing the annotation bottleneck, this approach enables systematic, quantitative investigation of genetic regulatory networks, environmental perturbation responses, and disease mechanisms at cellular resolution across hundreds to thousands of individuals [35].

References

[1]

[2] Bao, Z., Murray, J. I., Boyle, T., Ooi, S. L., Sandel, M. J., & Waterston, R. H. (2006). Automated cell lineage tracing in *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences*, 103(8), 2707–2712.

[3] Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. *Proceedings of the 26th Annual International Conference on Machine Learning*, 41–48.

[4] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1), 289–300.

[5] Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? *Proceedings of the 38th International Conference on Machine Learning*, 813–824.

[6] Besl, P. J., & McKay, N. D. (1992). A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2), 239–256.

[7] Bonazzola, R., Ravikumar, N., Attar, R., Ye, C., Piechnik, S. K., Neubauer, S., Petersen, S. E., & Frangi, A. F. (2024). Unsupervised ensemble-based phenotyping enhances discoverability of genes related to left-ventricular morphology. *Nature Machine Intelligence*, 6, 820–833.

[8] Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., & Shah, R. (1993). Signature verification using a "Siamese" time delay neural network. *Advances in Neural Information Processing Systems*, 6, 737–744.

[9] Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S. N., Steemers, F. J., Adey, A., Waterston, R. H., Trapnell, C., & Shendure, J. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352), 661–667.

[10] Coifman, R. R., & Lafon, S. (2006). Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1), 5–30.

[11] Du, Z., Santella, A., He, F., Tiongson, M., & Bao, Z. (2014). De novo inference of systems-level mechanistic models of development from live-imaging-based phenotype analysis. *Cell*, 156(1–2), 359–372.

[12] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1), 1–26.

[13] Engel, N., Belagiannis, V., & Dietmayer, K. (2021). Point transformer. *IEEE Access*, 9, 134826–134840.

[14] Friedl, P., & Alexander, S. (2011). Cancer invasion and the microenvironment: Plasticity and reciprocity. *Cell*, 147(5), 992–1009.

[15] Hao, Y., Stuart, T., Kowalski, M. H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., & Satija, R. (2024). Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology*, 42(2), 293–304.

[16] Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3), 535–547.

[17] Kaletta, T., & Hengartner, M. O. (2006). Finding function in novel targets: *C. elegans* as a model organism. *Nature Reviews Drug Discovery*, 5(5), 387–399.

[18] Lai, C. H., Chou, C. Y., Ch'ang, L. Y., Liu, C. S., & Lin, W. (2000). Identification of novel human genes evolutionarily conserved in *Caenorhabditis elegans* by comparative proteomics. *Genome Research*, 10(5), 703–713.

[19] Lalit M, Handberg-Thorsager M, Hsieh YW, Jug F, Tomancak P. in Computer Vision – ECCV 2020. Workshops Lecture Notes in Computer Science Ch. Chapter 30; 2020. pp. 458–473.

[20] Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *International Conference on Learning Representations*.

[21] Ma, J., Zhao, J., Yang, P., Yu, M., Lee, D., Jiao, Y., Chen, J., Gong, H., & Liu, Q. (2021). A new cell tracking algorithm for large-scale 4D microscopy images of *C. elegans* embryos. *Nature Methods*, 18(8), 1047–1050.

[22] McDole, K., Guignard, L., Amat, F., Berger, A., Malandain, G., Royer, L. A., Turaga, S. C., Branson, K., & Keller, P. J. (2018). In toto imaging and reconstruction of post-implantation mouse development at the single-cell level. *Cell*, 175(3), 859–876.e33.

[23] Moore, J. L., Du, Z., & Bao, Z. (2013). Systematic quantification of developmental phenotypes at single-cell resolution during embryogenesis. *Development*, 140(15), 3266–3274.

[24] Moyle, M.W., Barnes, K.M., Kuchroo, M. et al. Structural and developmental principles of neuropil assembly in C. elegans. Nature 591, 99–104 (2021). https://doi.org/10.1038/s41586-020-03169-5

[25] Murray, J. I., Bao, Z., Boyle, T. J., & Waterston, R. H. (2006). The lineaging of fluorescently-labeled *Caenorhabditis elegans* embryos with StarryNite and AceTree. *Nature Protocols*, 1(3), 1468–1476.

[26] Myronenko, A., & Song, X. (2010). Point set registration: Coherent point drift. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(12), 2262–2275.

[27] Packer, J. S., Zhu, Q., Huynh, C., Sivaramakrishnan, P., Preston, E., Dueck, H., Stefanik, D., Tan, K., Trapnell, C., Kim, J., Waterston, R. H., & Murray, J. I. (2019). A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution. *Science*, 365(6459), eaax1971.

[28] Pan, X., Xia, Z., Song, S., Li, L. E., & Huang, G. (2021). 3D object detection with PointFormer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7463–7472.

[29] Santella, A., Haus, E., & Bao, Z. (2025). Twin attention for learning of spatiotemporal structure and dynamics in complex tissues. *Manuscript in preparation.*

[30] Schapiro, D., Jackson, H. W., Raghuraman, S., Fischer, J. R., Zanotelli, V. R. T., Schulz, D., Giesen, C., Catena, R., Varga, Z., & Bodenmiller, B. (2017). histoCAT: Analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nature Methods*, 14(9), 873–876.

[31] Shah, P. K., Santella, A., Jacobo, A., Siletti, K., Hudspeth, A. J., & Bao, Z. (2019). An in toto approach to dissecting the genetic circuitry of development. *Nature*, 569, E5–E6.

[32] Smith, L. N., & Topin, N. (2019). Super-convergence: Very fast training of neural networks using large learning rates. *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, 11006, 1100612.

[33] Sulston, J. E., Schierenberg, E., White, J. G., & Thomson, J. N. (1983). The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Developmental Biology*, 100(1), 64–119.

[34] Toulany, N., Zha, Z., Gu, Y., Prakash, M., & Mueller, F. (2023). Uncovering developmental time and tempo using deep learning. *Nature Methods*, 20, 815–823.

[35] Ulman, V., Maška, M., Magnusson, K. E. G., Ronneberger, O., Haubold, C., Harder, N., et al. (2017). An objective comparison of cell-tracking algorithms. *Nature Methods*, 14(12), 1141–1152.

[36] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.

[37] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3), 261–272.

[38] Wagner, D. E., Weinreb, C., Collins, Z. M., Briggs, J. A., Megason, S. G., & Klein, A. M. (2018). Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, 360(6392), 981–987.

[39] Wang, Z., Ramsey, B. J., Wang, D., Wong, K., Li, H., Wang, E. W., Bao, Z. (2016). An Observation-Driven Agent-Based Modeling and Analysis Framework for C. elegans Embryogenesis. 11(11), e0166551–e0166551. https://doi.org/10.1371/journal.pone.0166551

[40] Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., & Liu, T. (2020). On layer normalization in the transformer architecture. *Proceedings of the 37th International Conference on Machine Learning*, 10524–10533.

[41] Yin, T., Zhou, X., & Krähenbühl, P. (2020). Center-based 3D object detection and tracking. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11784–11793.

[42] Yu, M., Zhao, J., Cao, Y., Santella, A., Colón-Ramos, D. A., Shroff, H., Mohler, W. A., & Bao, Z. (2021). Versatile neuronal cell matching and registration in *C. elegans* with LEVERAGER. *eLife*, 10, e69089.