Multi-Agent Deep Learning for Visual T Cell Behavioral Modeling

Joseph Li
shoupei@stanford.edu

Sean Tsung
stsung@stanford.edu

Adrian Molofsky
molofsky@stanford.edu

Abstract

Chimeric Antigen Receptor (CAR) T cell therapy represents a revolutionary approach in cancer treatment, leveraging engineered T cells to target and eliminate cancer cells. Understanding the collaborative dynamics and movement patterns of these T cells is crucial for optimizing therapeutic efficacy. In this study, we employ state-of-the-art computer vision and deep learning architectures to model and predict the behavior of CAR T cells using time-lapse microscopy data. Our approach utilizes a Vision Transformer (ViT) as the spatial encoder and a Transformer as the temporal decoder, effectively capturing complex spatial-temporal interactions. The input to our models consists of segmented video frames depicting T cell interactions with cancer cells, and the output is a set of predicted future coordinates for the T cells. Through extensive experimentation, we achieved a high prediction accuracy of 98% *in coordinate predictions, demonstrating the potential of our method to enhance the understanding of T cell dynamics. These findings provide valuable insights into the mechanisms of T cell collaboration and offer a promising direction for improving CAR T cell therapy.*

1. Introduction

The collaborative dynamics of T cells, including spatial coordination, signaling, and adaptive movement patterns, are critical components in the development of effective immunotherapies such as Chimeric Antigen Receptor (CAR) T cells and T cell receptor (TCR) T cells [6]. These therapies offer promising anti-cancer treatments by engineering T cells to specifically recognize and target cancer cells. However, the underlying mechanisms of T-cell collaboration remain poorly understood, posing a significant challenge to optimizing these therapies for improved patient outcomes.

Our motivation for pursuing this problem stems from the potential to enhance the efficacy of T-cell-based immunotherapies by gaining a deeper understanding of T-cell dynamics. By modeling T cell attack strategies using time-lapse microscopy data, we aim to decode how genetic knockouts alter T cell behavior and predict T cell movement patterns in unseen scenarios. This research could provide valuable insights into the mechanisms of T-cell collaboration, ultimately contributing to the development of more effective cancer treatments. The input to our algorithm consists of consecutive frames from segmented live-cell microscopy videos, capturing the interactions between TCR T cells and cancer cells. We employ three distinct architectures—ResNet-LSTM, ViT-LSTM, and ViT-Transformer output predicted sets of future T cell coordinates. Each T cell is represented as an agent, and expert trajectories are computed from segmentation masks and cell tracking data provided by the Caliban and Occident pipelines.

Our objective is to predict future T cell positions by considering the locations of neighboring T cells and cancer cells, thereby decoding the signals driving coordinated behaviors such as aggregation, swarming, proliferation, and recruitment. By comparing the predictive accuracy of these models, we aim to identify the most effective approach for accurately modeling T cell dynamics. Our findings indicate that ViT-Transformer model achieved superior performance, with an accuracy of 93%, highlighting its potential for advancing the field of T-cell-based immunotherapy.

2. Related Work

The study of T cellular behavior modeling has seen significant advancements through various approaches, particularly in the context of live-cell imaging and machine learning techniques. This section categorizes existing research into three main areas: antigen sensitivity enhancement, T cell signaling regulation, and computational modeling of cellular interactions.

Carnevale et al. [3] demonstrated that RASA2 knockout can significantly improve antigen sensitivity and persistence in T cells. This work is pivotal as it highlights a genetic modification approach to enhance immune response. However, the study primarily focuses on in vitro experiments, which may not fully capture in vivo complexities. The strength of this approach lies in its potential for targeted genetic interventions, though its applicability in clinical settings requires further exploration.

Research on the CUL5 E3 ligase complex [8] has shown its role in enhancing anti-tumor responses by regulating CD8$^+$ T cell signaling. This study provides insights into the molecular mechanisms that can be leveraged to boost immune responses against tumors. While the findings are promising, the challenge remains in translating these molecular insights into therapeutic strategies. The work is commendable for its detailed mechanistic exploration, yet it lacks a comprehensive analysis of potential side effects in therapeutic applications.

Verma et al. [10] analyzed TCR T cell–cancer cell interactions using live-cell imaging, laying the groundwork for computational modeling of these interactions. Building on this, we draw inspiration from a predator-prey framework and coordinated multi-agent imitation learning, as described by Le et al. [7], to model T-cell cooperative behaviors. This approach is innovative in its use of policies derived from expert demonstrations, offering a robust framework for simulating complex cellular interactions.

Moen et al. [9] developed convolutional neural networks for subcellular structure identification, which is crucial for accurate modeling of cellular environments. Their work is notable for its high accuracy in identifying subcellular components, though it requires substantial computational resources. Similarly, Bochinski et al. [2] applied reconstruction methods for densely packed cell populations, providing a high-speed solution for cell tracking. While effective, these methods often struggle with scalability in larger datasets.

The current state-of-the-art in T cellular behavior modeling involves a combination of genetic, molecular, and computational approaches [1] [4]. While many studies still rely on manual analysis, there is a clear trend towards automation and machine learning-driven methodologies. The integration of deep learning techniques, such as those by Moen et al. [9] and Greenwald et al. [5], represent a significant advancement in the field. However, challenges remain in terms of scalability and real-world applicability.

In conclusion, while each approach has its strengths and weaknesses, the combination of genetic insights and computational modeling offers a promising path forward. Future research should focus on integrating these methodologies to develop comprehensive models that can be applied in clinical settings.

3. Data

The dataset utilized in this study originates from three distinct medical laboratory experiments, designated as SafeHarbor, CUL5, and RASA2. Each experiment comprises five microscopic videos capturing the activities of T-cells over a 24-hour period, with frames recorded at four-minute intervals. Consequently, each video consists of approximately 350 frames, each with a resolution of 600x600 pixels. The laboratory has provided annotations indicating the pixel positions of T-cells and cancer cells within each frame. Our research specifically focuses on the CUL5 dataset, which includes certain test T-cells.

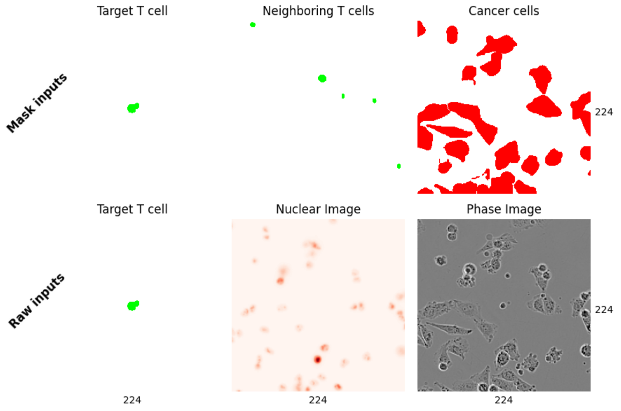

Figure 1. Segmented Live-Cell Image

Observations indicate that T-cells exhibit minimal dramatic movement. To enhance the sample size, we cropped the original 600x600 frames into 224x224 patches with a stride of 38 pixels. This preprocessing step yields approximately 35,000 samples per video. For our experiments, we utilized four videos, resulting in a total of 140,000 samples for the training dataset. Additionally, half a video, equating to 12,500 samples, was allocated for validation, and the remaining 12,500 samples were reserved for testing.

Given our objective to predict T-cell trajectories, the dataset inherently encodes both spatial and temporal information. We define a sample as a sequence of T consecutive frames (e.g., T = 5, 8, 12). The number of T-cells present in each frame varies due to annotation errors and the natural lifecycle of T-cells. To standardize our approach, we limit the number of T-cells to N (e.g., N = 50) per sample. We identify all unique T-cell IDs within the samples and randomly select N T-cells for model training.

The laboratory annotations also include T-cell positions within the frames, which we utilize to compute the centroid of each T-cell. These centroids serve as the coordinates for the T-cells. The centroid calculation involves summing the pixel positions annotated for each cell and computing the mean.

The processed dataset comprises cropped frame images and the computed T-cell coordinates. The dimensions for images in a single sample are T x H x W, and for coordinates, they are T x N x 2. During model training, we employ batches of samples, resulting in dimensions of B x T x H x W for images and B x T x N x 2 for coordinates, where B represents the batch size, T is the number of consecutive frames, and H = W = 224. N denotes the maximum number of T-cells in a sample, set to 50 based on the observed maximum in a 224x224 patch. If fewer than N T-cells are present in a frame, zero-padding is applied.

We noted that the initial frames were suboptimal in quality and thus excluded them from analysis. Additionally, some T-cells were not consistently annotated across consecutive frames. To address these inconsistencies, we trained models using different frame numbers for N = 5, 8, and 12.

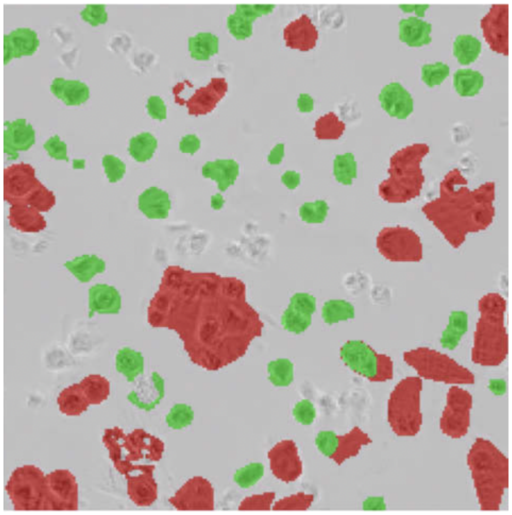

Figure 2. Masked Live-Cell Image

Additionally, the annotated positions of T-cells and cancer cells are employed to create masks for the images, facilitating illustration and visualization purposes, as depicted in the accompanying figures.

In addition to randomly selecting T-cells across frames, we also experimented with selecting T-cells that consistently appear in all frames within a sample. While this approach ensures temporal consistency in the data, it is computationally more intensive and reduces the available sample size for training.

4. Methods

The objective of this study is to model the behavior of T-cells and predict their movement trajectories, a task that inherently involves both spatial and temporal dimensions. To address this challenge, we propose a model architecture that incorporates a spatial encoder and a temporal decoder. Convolutional Neural Networks (CNNs) are well-suited for spatial feature extraction due to their ability to capture local patterns effectively. Recently, Vision Transformers (ViTs) have also emerged as a promising alternative for spatial feature encoding, offering advantages in capturing global context.

For temporal decoding, Long Short-Term Memory (LSTM) networks are traditionally employed for sequence modeling, given their capability to handle temporal dependencies. Additionally, transformer decoders have demonstrated significant power in sequence prediction tasks. In this section, we introduce and evaluate three model architectures: CNN-LSTM, ViT-LSTM, and ViT-Transformer, analyzing their potential effectiveness in predicting T-cell trajectories.

4.1. CNN-LSTM Model

In this approach, we model T-cell trajectories using a Convolutional Neural Network (CNN) for spatial feature extraction and a Long Short-Term Memory (LSTM) network for capturing temporal dynamics.

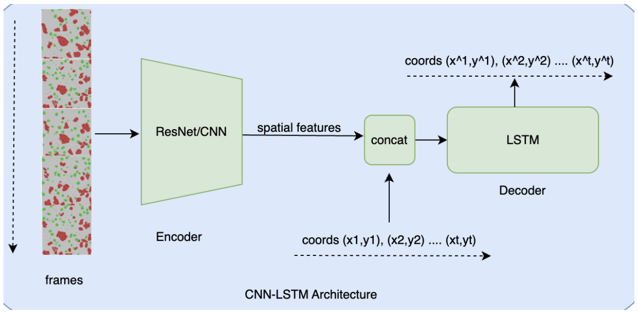
coords (x^1,y^1), (x^2,y^2) .... (x^t,y^t)
ResNet/CNN
spatial features
concat
LSTM
Decoder
Encoder
coords (x1,y1), (x2,y2) .... (xt,yt)
frames
CNN-LSTM Architecture

Figure 3. ResNet LSTM Hybrid Architecture

4.1.1 Spatial Encoder (CNN)

The spatial encoder employs a pretrained ResNet-50 to extract spatial features from video frames. The input to the CNN is a sequence of video frames, and the output is a sequence of feature maps:

$$\mathbf{F}_t = CNN(\mathbf{I}_t) \in \mathbb{R}^{C' \times H' \times W'}$$

where $\mathbf{F}_t$ is the feature map at time t.

4.1.2 Temporal Encoder (LSTM)

The LSTM processes a sequence that combines CNN-extracted spatial features and coordinate embeddings to model temporal dependencies.For each time step t, the feature map $\mathbf{F}_t$ and the coordinate embedding $\mathbf{e}_t$ are concatenated:

$$x_t = \text{Concat}(\mathbf{F}_t, \mathbf{e}_t)$$

The LSTM updates its hidden state $\mathbf{h}_t$ using:

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_c x_t + U_c h_{t-1} + b_c)$$
$$h_t = o_t \odot tanh(c_t)$$

where i_t, f_t, o_t are the input, forget and output gates. The coordinate is predicted by $\hat{\mathbf{a}}_t = \text{Linear}(h_t)$.

The CNN-LSTM model benefits from the robust feature extraction capabilities of the pretrained ResNet-50, enhancing spatial representation. However, the sequential nature of LSTM can limit scalability and efficiency, particularly with longer sequences. This model is effective for capturing local spatial patterns but may struggle with complex temporal dependencies.

During prediction, the LSTM operates in an autoregressive manner. Starting with the initial coordinates, the model predicts the next set of coordinates, which are then fed back as input for subsequent predictions. The LSTM's hidden and cell states are maintained across steps to ensure temporal continuity.

4.2. ViT-LSTM model

Similar to CNN-LSTM model, the ViT-LSTM model combines the strengths of Vision Transformers (ViT) for spatial encoding with LSTMs for temporal sequence modeling.

4.2.1 Spatial Encoder (ViT)

The Vision Transformer processes video frames using a pretrained ViT model. Each frame is divided into non-overlapping patches, which are flattened and linearly projected into a d-dimensional space. Positional encodings are added to retain spatial order. The transformer encoder applies multi-head self-attention to these embeddings, capturing global spatial dependencies.

4.2.2 Temporal Encoder (LSTM)

The LSTM processes the sequence of spatial features, similar to the CNN-LSTM model, but benefits from the ViT's ability to capture more comprehensive spatial information.

The ViT-LSTM model enhances spatial feature extraction through the Vision Transformer, offering improved spatial context understanding. However, it still relies on LSTMs for temporal modeling, which may not fully exploit the temporal dynamics present in the data.

Similar to the CNN-LSTM model, the ViT-LSTM uses an autoregressive approach during testing. The initial coordinates are used to start the prediction, and each predicted set of coordinates is fed back into the LSTM for the next prediction step.

4.3. ViT-Transformer model

This model employs a Vision Transformer for spatial encoding and a Transformer decoder for temporal modeling, aiming to fully leverage the transformer architecture's capabilities.

The ViT-Transformer model excels in capturing complex temporal patterns and interactions, offering superior scalability and parallelization compared to LSTM-based models.

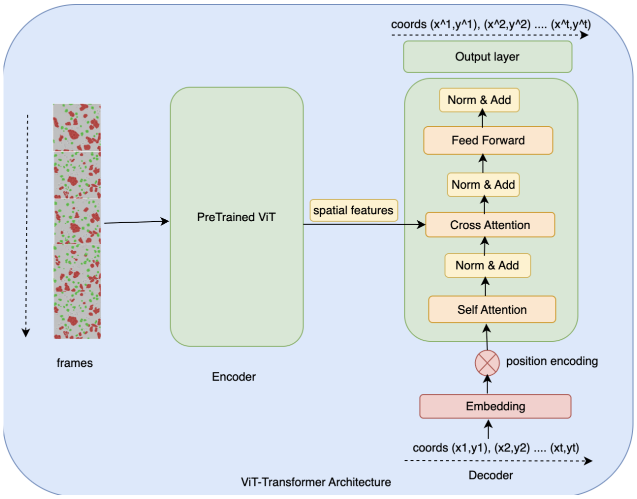

Figure 4. ViT Encoder Transformer Decoder Architecture

Its ability to process sequences in parallel and capture long-range dependencies makes it particularly advantageous for T-cell prediction tasks, where intricate spatial-temporal interactions are prevalent. The model architecture allows it to efficiently handle large datasets and complex patterns, making it a promising choice for accurately predicting T-cell trajectories.

4.3.1 Spatial Encoder (ViT)

The spatial encoder is identical to that in the ViT-LSTM model, utilizing a pretrained ViT for robust spatial feature extraction. Here we provide a high level description of the encoder for later experiment analysis.

Input The encoder receives video frames with dimensions size $B \times T \times H \times W \times C$ (e.g. $16 \times 5 \times 244 \times 244 \times 3$).

Patch Embedding Each frame is divided into N non-overlapping patches of size $P \times P$. This division allows the model to process smaller, manageable sections of the image, capturing local features. Each patch $I_t^{(i)}$ from frame t is flattened into a vector and linearly projected into a d-dimensional embedding space.

$$z_t^{(i)} = \text{Linear}(\text{Flatten}(I_t^{(i)})) \in \mathbb{R}^d$$

This transformation enables the model to represent each patch as a point in a high-dimensional space, facilitating the learning of complex spatial patterns.

Positional Encoding To retain the spatial order of patches, learnable positional embeddings $E_{pos} \in \mathbb{R}^{N \times d}$ are added to the patch embeddings. This step is crucial for maintaining the spatial context that is lost during the flattening process.

$$Z_t = [z_t^{(1)} + E_{pos}^{(1)}, \ldots, z_t^{(N)} + E_{pos}^{(N)}]$$

The positional encoding helps the model understand the relative positions of patches within the frame, which is essential for tasks involving spatial relationships.

MultiHead Attention The sequence of patch embeddings, now enriched with positional information, is processed using a multi-head self-attention mechanism. This mechanism allows the model to weigh the importance of different patches relative to each other, capturing global spatial dependencies.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

Here Q, K and V represent the query, key, and value matrices, respectively, and W^O is the output projection matrix. The multi-head attention enables the model to focus on different parts of the input simultaneously, enhancing its ability to learn complex spatial patterns.

4.3.2 Temporal Decoder (Transformer)

Input The temporal decoder receives historical T-cell coordinates, represented as a sequence $\mathbf{a}_{1:t} = (x_1, y_1), \ldots, (x_t, y_t)$, which are embedded into a d-dimensional space. This embedding allows the model to process the coordinates in a format compatible with the transformer architecture.

Causal Masking Causal masking is applied to ensure that predictions $\hat{\mathbf{a}}_{t+1}$ depend only on past coordinates $\mathbf{a}_{1:t}$.This masking prevents information from future time steps from influencing the current prediction, maintaining the autoregressive nature of the model.

Cross-Attention The cross-attention mechanism fuses spatial features from the ViT encoder with temporal embeddings from the coordinate sequence. This integration allows the model to leverage both spatial and temporal information when making predictions.

$$\text{CrossAttention}\{Q_{dec}, K_{enc}, V_{enc}\}$$

Here, Q_{dec} represents the query from the decoder, while K_{enc} and V_{enc} are the key and value from the encoder. This mechanism enables the model to align spatial features with temporal dynamics effectively.

Output The final output of the temporal decoder is the predicted coordinates $\hat{\mathbf{a}}_{t+1}$. These predictions are generated by considering both the historical trajectory and the spatial context provided by the ViT encoder.

During testing, the ViT-Transformer model uses an autoregressive approach with causal masking. Starting with the initial coordinates, the model predicts the next set of coordinates, which are then used as input for subsequent predictions. The use of causal masking ensures that each prediction is based only on past information, maintaining the autoregressive nature of the process.

5. Experiments

5.1. Setup

The experiments were conducted using a cluster equipped with four NVIDIA A30 GPUs, which provided the necessary computational power to efficiently train the models. We initially selected a batch size of 4, but found that this underutilized the GPUs' parallel processing capabilities. To improve training efficiency and better leverage the available hardware, we increased the batch size to 16.

The models were trained using a scheduled learning rate strategy, beginning with an initial learning rate of 1×10^{-4}. The Adam optimizer was employed to facilitate convergence, leveraging its adaptive learning rate capabilities to optimize the training process. The duration of training varied between 2 to 3 hours per model, contingent upon the number of consecutive frames included in each sample.

The maximum number of T-cells per sample was identified as a critical parameter influencing both model quality and training duration. Initially, the models were configured with 80 T-cells per sample, which substantially increased the training time. To mitigate this, we decided to reduce the number of T-cells tracked to 50 per sample when training models with shorter sequence lengths (5 frames and 8 frames; see below for details), since shorter sequences contain fewer cells on average. However, when training the models with a 12-frame sequence length, we ran the model with 80 cells per sample due to the higher average number of observed cells. This adjustment significantly decreased the training time for the models run on shorter sequence lengths while maintaining a satisfactory level of model performance.

We trained models with three different sequence lengths: 5, 8, and 12 consecutive frames. We hypothesized that, within the range of sequence lengths considered, model performance might increase with longer sequence lengths. However, we knew that the longer the sequence length, the longer it would take to train the models.

The choice of batch size and the number of T-cells per sample were pivotal in balancing training efficiency and model accuracy. The scheduled learning rate and the use of the Adam optimizer contributed to stable convergence across different model architectures. These experiments underscore the importance of parameter tuning in optimizing the training process for complex spatial-temporal models.

5.2. Metrics

In the context of T-cell trajectory prediction, it is crucial to define appropriate metrics that account for the spatial nature of the data. Since T-cells are not point entities, we employ **accuracy within a specified pixel radius** as a success criterion.

Specifically, an accuracy of 90% within 10 pixels indicates that the predicted trajectory falls within 10 pixels of the ground truth 90% of the time. This metric provides a spatial tolerance that is essential for evaluating predictions in biological imaging contexts. In addition to spatial accuracy, we utilize several standard metrics for trajectory prediction to comprehensively assess model performance:

Mean Absolute Error (MAE): MAE measures the average magnitude of errors between predicted and true trajectories, without considering their direction. It is calculated as the mean of the absolute differences between predicted and actual positions over all time steps. MAE provides a straightforward measure of prediction accuracy, with lower values indicating better performance.

Average Displacement Error (ADE): ADE is the average Euclidean distance between predicted and true trajectories over all time steps. It is computed by averaging the displacement errors at each time step across the entire trajectory. ADE is particularly useful for evaluating the overall accuracy of a predicted trajectory, as it considers the entire sequence of predictions.

Final Displacement Error (FDE): FDE measures the Euclidean distance between the predicted and true positions at the final time step of the trajectory. This metric focuses on the endpoint accuracy of the prediction, which is critical in applications where the final position is of particular importance.

These metrics collectively provide a comprehensive evaluation framework for trajectory prediction models, allowing for nuanced assessments of both spatial accuracy and temporal prediction quality. By employing these metrics, we ensure that our models are rigorously evaluated and capable of producing reliable predictions in complex biological environments.

5.3. Results

The performance of the models was evaluated on the final validation set using three different sequence lengths (5, 8, and 12 frames). The results are summarized in Tables 1–3, which report the accuracy, Mean Absolute Error (MAE), Average Displacement Error (ADE), and Final Displacement Error (FDE) for each model configuration. Following model training and validation, we selected the best-performing model and evaluated its performance on a hold-out test set.

5-Frame Sequence Evaluation As shown in Table 1, the ViT-Transformer model significantly outperformed the other architectures in the 5-frame sequence evaluation. It achieved an impressive accuracy of 93.61%, with MAE, ADE, and FDE values of 2.0287, 3.1328, and 3.1453, respectively. This indicates that the ViT-Transformer model is highly effective in capturing the spatial and temporal dynamics of the T-cell trajectories.

In contrast, the ResNet-LSTM and ViT-LSTM models demonstrated considerably lower performance, with accuracies of 8.71% and 8.02%, respectively. Their MAE, ADE, and FDE metrics were substantially higher, reflecting less precise trajectory predictions. These results suggest that the combination of Vision Transformer (ViT) and Transformer architectures provides a superior framework for modeling complex trajectory data.

Table 1. Final validation set metrics: 5 frames

Model Type	Accuracy	MAE	ADE	FDE
ResNet–LSTM	8.71%	41.9796	63.2818	63.3383
ViT–LSTM	8.02%	43.5380	65.5375	65.1746
ViT–Transformer	93.61%	2.0287	3.1328	3.1453

8-Frame Sequence Evaluation The evaluation with 8-frame sequences, detailed in Table 2, further highlights the robustness of the ViT-Transformer model. It achieved an accuracy of 96.79%, with MAE, ADE, and FDE values of 1.2261, 1.8884, and 1.8742, respectively. These metrics underscore the model's ability to maintain high prediction accuracy over longer sequences, which is crucial for applications requiring extended temporal analysis.

The ResNet-LSTM and ViT-LSTM models showed improved performance compared to the 5-frame evaluation, with accuracies of 16.20% and 16.31%, respectively. However, their error metrics remained significantly higher than those of the ViT-Transformer, indicating that while they benefit from longer sequences, they still fall short in terms of precision and reliability.

Table 2. Final validation set metrics: 8 frames

Model Type	Accuracy	MAE	ADE	FDE
ResNet–LSTM	16.20%	31.0874	46.7935	46.8892
ViT–LSTM	16.31%	30.6151	46.0481	45.4517
ViT–Transformer	96.79%	1.2261	1.8884	1.8742

12-Frame Sequence Evaluation Table 3 presents the results for the 12-frame evaluation, where the ViT-

Transformer again demonstrated exceptional performance, achieving an accuracy of 97.74% and even lower MAE, ADE, and FDE metrics (0.8787, 1.3592, and 1.3354). The performance gap with LSTM-based baselines is still wide, with their accuracies only reaching 20.77% and 24.10% and error metrics still an order of magnitude higher.

Table 3. Final validation set metrics: 12 frames

Model Type	Accuracy	MAE	ADE	FDE
ResNet–LSTM	20.77%	26.4346	39.6137	39.7748
ViT–LSTM	24.10%	22.9939	34.5097	33.8738
ViT–Transformer	97.74%	0.8787	1.3592	1.3354

In summary, the ViT-Transformer achieved remarkable accuracy levels of 93.61%, 96.79%, and 97.74% for the 5-frame, 8-frame, and 12-frame sequences, respectively. These results underscore the model's ability to consistently deliver precise predictions, even as the sequence length increases. The high accuracy indicates that the ViT-Transformer effectively captures the intricate spatial-temporal dependencies inherent in T-cell movement, setting a new benchmark for trajectory prediction tasks.

The results clearly demonstrate the exceptional effectiveness of the ViT-Transformer model in trajectory prediction tasks. Its superior performance—evidenced by high accuracy and low error metrics across the 5-frame, 8-frame, and 12-frame sequence evaluations—shows that it effectively leverages the strengths of both Vision Transformers and Transformer architectures to capture complex spatial-temporal patterns. The substantial performance gap between the ViT-Transformer and other models underscores the critical importance of architectural choices in achieving state-of-the-art results. The following section delves deeper into the model's capabilities and provides a visual analysis of its predictions.

The validation step accuracy, depicted in Figure 5, illustrates the model's robust performance across different validation scenarios. The figure highlights the model's ability to maintain high accuracy throughout the validation process, demonstrating its reliability and generalization capabilities.

Figure 6 presents visualizations of the trajectories predicted by the ViT-Transformer model. These visualizations provide qualitative insights into the model's predictive prowess. The predicted trajectories closely align with the ground truth, showcasing the model's proficiency in accurately forecasting both the direction and magnitude of T-cell movements. The visualizations reveal that the ViT-Transformer not only predicts the overall trajectory path but also captures subtle variations in movement patterns. This precision is crucial for applications requiring detailed and accurate modeling of cellular dynamics, such as in drug development and immunotherapy research.

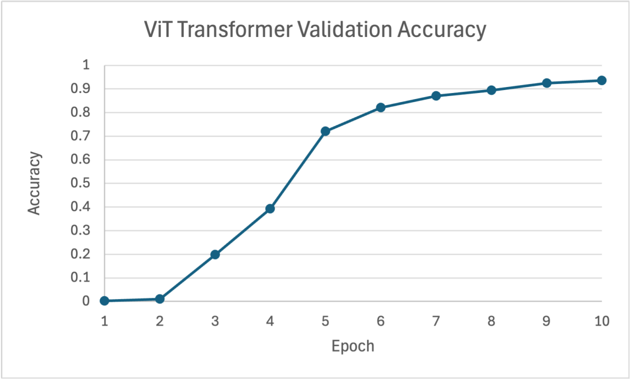

Figure 5. ViT-Transformer Validation Accuracy

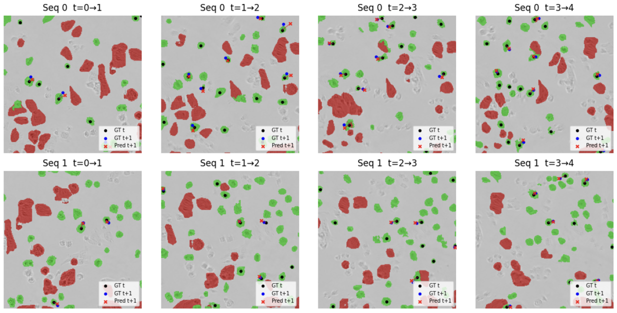

Figure 6. Predicted Trajectory vs. Ground Truth

Test Set Evaluation Finally, we evaluated the test-set performance of our best-performing model, the 12-frame ViT-Transformer. The model achieves 98.20% accuracy and extremely low MAE, ADE, and FDE values (0.8145, 1.2607, and 1.2981). We used a holdout test set containing frames that were never seen during training or validation. The excellent performance on the test set strongly suggests that our model did not overfit to the training data and has the potential to produce accurate trajectory predictions on unseen live cell microscopy samples.

Table 4. Test set metrics for ViT-Transformer

Accuracy	MAE	ADE	FDE
98.20%	0.8145	1.2607	1.2981

5.4. Discussion

The ViT-Transformer's superior performance can be attributed to its architectural design, which leverages the strengths of Vision Transformers for spatial feature extraction and Transformers for temporal sequence modeling. This combination allows the model to effectively handle the complex, high-dimensional data characteristic of T-cell trajectories.

The substantial performance gap between the ViT-Transformer and other models, such as ResNet-LSTM and ViT-LSTM, highlights the importance of selecting appropriate architectures for trajectory prediction tasks. The ViT-Transformer's ability to maintain high accuracy and low error rates across varying sequence lengths positions it as a state-of-the-art solution in the field.

Future research could explore the integration of additional contextual information, such as environmental factors or cell-cell interactions, to further enhance the model's predictive capabilities. Additionally, optimizing hyperparameters and exploring alternative training strategies could yield further improvements in performance.

6. Conclusion

In this study, we explored the efficacy of various deep learning architectures for predicting T-cell trajectories, with a particular focus on the ViT-Transformer model. Our comprehensive evaluation demonstrated that the ViT-Transformer significantly outperformed other models, such as ResNet-LSTM and ViT-LSTM, in 5-frame, 8-frame, and 12-frame sequence evaluations. The ViT-Transformer achieved remarkable accuracy levels while maintaining low error metrics across all evaluated scenarios. Indeed, the best-performing ViT-Transformer model was found to achieve an extremely high test-set accuracy of 98.20%. These results underscore the model's ability to effectively capture the complex spatial-temporal dependencies inherent in T-cell movement.

The superior performance of the ViT-Transformer can be attributed to its architectural design, which combines the strengths of Vision Transformers for spatial feature extraction with Transformers for temporal sequence modeling. This synergy allows the model to handle high-dimensional data and accurately predict both the direction and magnitude of T-cell movements. In contrast, the ResNet-LSTM and ViT-LSTM models struggled to achieve comparable accuracy, highlighting the importance of selecting appropriate architectures for trajectory prediction tasks.

Looking forward, there are several avenues for future research that could further enhance the predictive capabilities of the ViT-Transformer model. With additional time, team members, or computational resources, we would explore the integration of contextual information, such as environmental factors or cell-cell interactions, to provide a more holistic understanding of T-cell dynamics. Additionally, optimizing hyperparameters and experimenting with alternative training strategies could yield further improvements in model performance. Finally, extending the model to predict longer sequences or incorporating real-time data processing capabilities could broaden its applicability in clinical and research settings, ultimately contributing to advancements in immunotherapy and personalized medicine.

7. Contribution

Joseph worked on live-cell image processing and cell coordinates calculation for labels. Joseph authored the ResNet-LSTM, Vit-Transformer, and Vit-LSTM models. Joseph also drafted the final report.

Sean worked on the training framework, model development, and script setup. Sean migrated the model to the lightning framework, trained all models, and compiled the results.

Adrian worked on the data pipleline and spearheaded on additional model architecture (e.g. 3D CNN), and explored some other data processing approaches that resulted in the milestone results.

All three of them contributed equally to the discussion.

References

[1] M. Alieva, A. K. L. Wezenaar, E. J. Wehrens, A. Cleven, J. Dekkers, et al. Bridging live-cell imaging and next-generation cancer treatment. *Nature Reviews Cancer*, 23:731–745, 2023.

[2] E. Bochinski, V. Eiselein, and T. Sikora. High-speed tracking-by-detection without using image information. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2017.

[3] J. Carnevale, E. Shifrut, N. Kale, et al. Rasa2 ablation in t cells boosts antigen sensitivity and long-term function. *Nature*, 609:174–182, 2022.

[4] J. F. Dekkers, M. Alieva, A. Cleven, et al. Uncovering the mode of action of engineered t cells in patient cancer organoids. *Nature Biotechnology*, 41:60–69, 2023.

[5] N. F. Greenwald, G. Miller, E. Moen, A. Kong, A. Kagel, C. C. Fullaway, B. J. McIntosh, K.-H. Leow, M. S. Schwartz, T. Dougherty, C. Pavelchek, S. Cui, I. Camplisson, O. Bar Tal, J. Singh, M. Fong, G. Chaudhry, Z. Abraham, J. Moseley, and D. van Valen. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nature Biotechnology*, 40(4):555–565, 2022.

[6] A. C. Johnston, G. M. Alicea, C. C. Lee, P. V. Patel, E. A. Hanna, E. Vaz, A. Forjaz, Z. Wan, P. R. Nair, Y. Lim, T. Chen, W. Du, D. Kim, T. D. Nichakawade, V. W. Rebecca, C. L. Bonifant, R. Fan, A. L. Kiemen, P. H. Wu, and D. Wirtz. Engineering self-propelled tumor-infiltrating car t cells using synthetic velocity receptors. *bioRxiv*, Mar. 2024. PMID: 38168186; PMCID: PMC10760159.

[7] H. M. Le, Y. Yue, P. Carr, and P. Lucey. Coordinated multi-agent imitation learning, 2018.

[8] X. Liao, W. Li, H. Zhou, B. Rajendran, A. Li, J. Ren, Y. Luan, D. Calderwood, B. Turk, W. Tang, Y. Liu, and D. Wu. The cul5 e3 ligase complex negatively regulates central signaling pathways in cd8$^+$ t cells. *Nature Communications*, 15(1):603, 2024. PMID: 38242867; PMCID: PMC10798966.

[9] E. Moen, D. Bannon, T. Kudo, W. Graf, C. Ward, and D. van Valen. Deep learning for cellular image analysis. *Nature Methods*, 16(12):1233–1246, 2019.

[10] A. Verma, C. Yu, S. Bachl, I. Lopez, M. Schwartz, E. Moen, N. Kale, C. Ching, G. Miller, T. Dougherty, E. Pao, W. Graf, C. Ward, S. Jena, A. Marson, J. Carnevale, D. Van Valen, and B. E. Engelhardt. Cellular behavior analysis from live-cell imaging of tcr t cell–cancer cell interactions. *bioRxiv*, Nov. 2024. PMID: 39605616; PMCID: PMC11601648.

8. Appendix

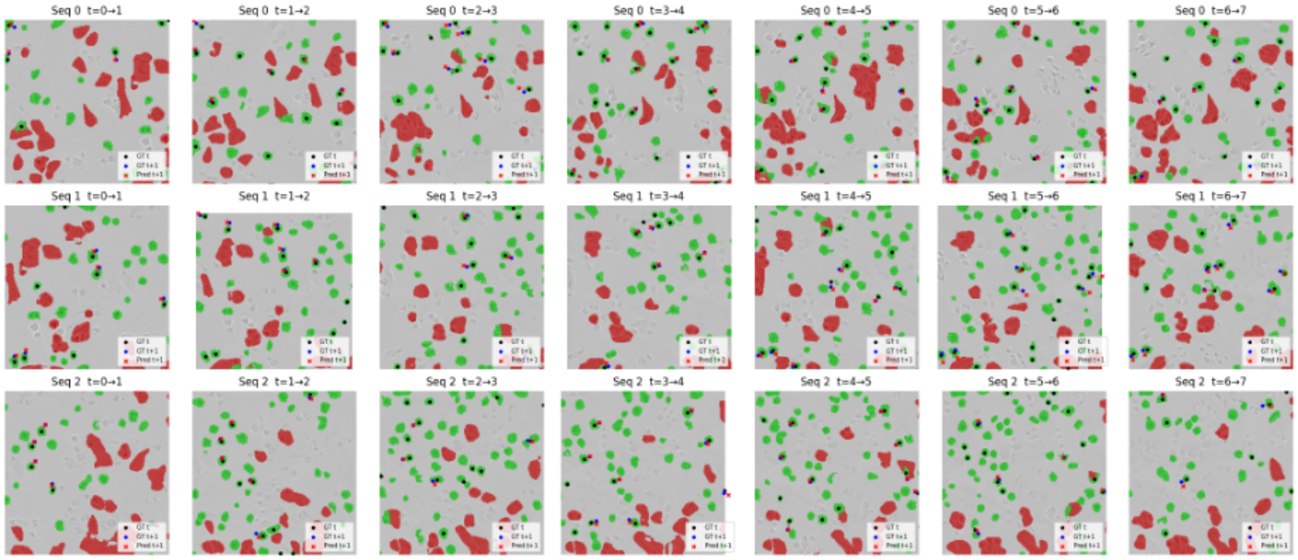

Figure 7. ViT-Transformer Predicted Trajectory