Multi-Agent Social Behavior Understanding via Ego-GAT-SqueezeNet: Integrating Graph Attention and Squeezeformer

Joseph Siu
Tsinghua University
University of Toronto
xdx25@mails.tsinghua.edu.cn

Qi An
Tsinghua University
aq21@mails.tsinghua.edu.cn

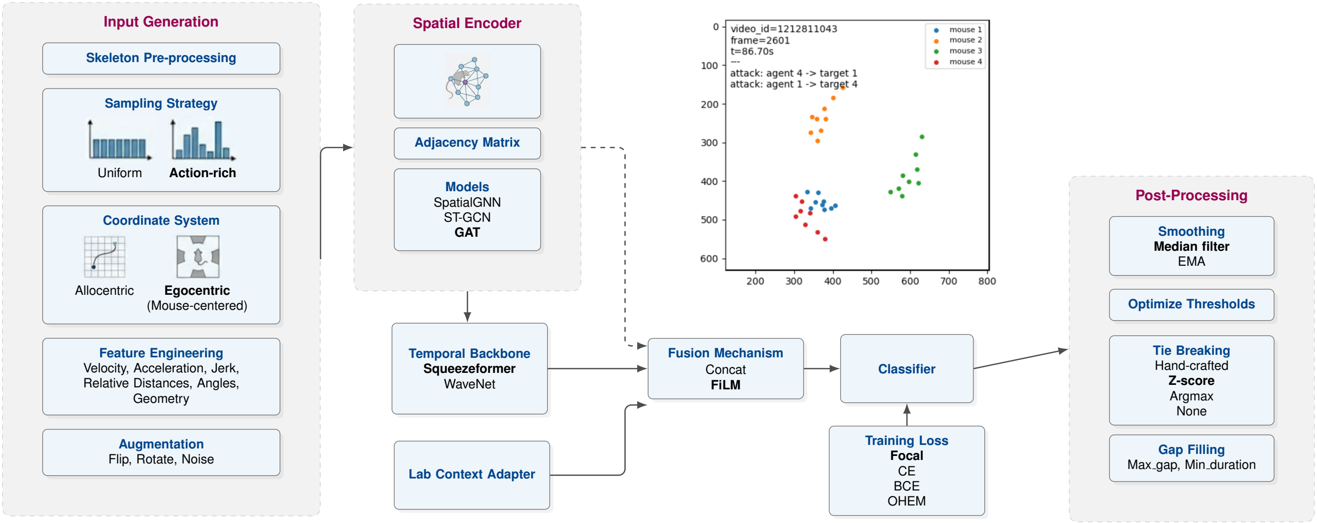


Figure 1: Overview of the Ego-GAT-SqueezeNet framework. The pipeline begins with Input Generation, featuring egocentric alignment, action-rich sampling, and skeleton unification. Features are processed through a Spatial Encoder (GAT) to model joint-level topology and a Temporal Backbone (Squeezeformer) to capture long-range dependencies. A Lab Context Adapter concurrently extracts environment and subject-specific embeddings, which are integrated via FiLM (Feature-wise Linear Modulation). The model is optimized using Dataset-aware Focal Loss and refined through a Post-Processing suite including dynamic threshold grid search and gap filling.

Abstract

Quantifying social behavior in laboratory animals is fundamental to neuroscience but remains hindered by manual annotation's subjectivity. The Multi-Agent Behavior (MABe) challenge addresses this by benchmarking automated recognition from pose data, yet faces challenges like extreme class imbalance, complex topology, and cross-laboratory domain shifts.

*In this work, we propose **Ego-GAT-SqueezeNet**, a unified framework for multi-agent behavior understanding. First, we introduce an egocentric alignment strategy to invariantize agent features against translation and rotation. Second, we employ a Graph Attention Network (GAT) to explicitly model the dynamic spatial topology. Crucially, we integrate a Squeezeformer backbone that leverages efficient downsampling to capture long-range dependencies in high-frequency sequences. For environmental heterogeneity, we utilize Feature-wise Linear Modulation (FiLM) to dynamically recalibrate features based on laboratory and subject identities. Our approach achieves an F1-score of 0.7702 on the validation set, outperforming baselines by identifying rare social actions across diverse experimental setups.*

1. Introduction

1.1. Background and Motivation

Quantifying animal behavior is a cornerstone of modern neuroscience, genetics, and pharmacology. To understand how neural circuits control social interaction or how new drugs affect phenotype, researchers typically rely on the precise characterization of behaviors in laboratory animals, such as mice. While recent advances in markerless pose estimation tools (e.g., SLEAP [9], DeepLabCut [7]) have automated the extraction of body part coordinates, bridging the semantic gap from raw keypoint trajectories to interpretable "ethograms" (behavioral catalogs) [1] remains a significant bottleneck.

Traditionally, this task relies on manual annotation, which is labor-intensive, unscalable, and prone to inter-rater variability. The Multi-Agent Behavior (MABe) Challenge [12] aims to solve this by benchmarking automated methods that can classify fine-grained social and individual behaviors solely from pose tracking data. Unlike single-agent scenarios, multi-agent behavior understanding requires dissecting complex interactions between an "Agent" mouse and a "Target" mouse, creating a highly dynamic and structured problem space.

1.2. Challenges

Developing a robust automated system for the MABe task presents four fundamental challenges that standard action recognition models struggle to address:

1. **Implicit Social Topology**: Social interaction is not merely a sum of two individuals' motions; it is defined by their relative geometric configuration (e.g., nose-to-tail contact). Standard Convolutional Neural Networks (CNNs) designed for grid-like image data fail to explicitly model this irregular, graph-structured topology of body joints and social connections.
2. **Long-Term Temporal Dependencies with Redundancy**: Distinguishing complex behaviors like *social following* from *random coexistence* requires analyzing temporal contexts spanning dozens to hundreds of frames. However, high-frame-rate pose data contains significant temporal redundancy. Standard Recurrent Neural Networks (LSTMs) [3] suffer from forgetting issues over long sequences, while vanilla Transformers [13] incur prohibitive quadratic computational costs ($O(N^2)$) when processing long windows.
3. **Extreme Class Imbalance**: Mice spend the vast majority of their time in low-activity states (e.g., sleeping or huddling). Scientifically critical behaviors, such as aggressive attacks or specific social investigations, are rare, constituting only a small fraction of the dataset [12]. This imbalance causes standard training objectives to bias heavily toward the majority classes.
4. **Cross-Laboratory Domain Shifts**: The dataset comprises recordings from over 20 independent laboratories, introducing significant heterogeneity in camera viewpoints, lighting conditions, and animal sizes. A robust model must generalize across these distinct domains without overfitting to laboratory-specific artifacts, a problem often overlooked in single-domain benchmarks.

1.3. Our Approach: Ego-GAT-SqueezeNet

To address these challenges, we propose **Ego-GAT-SqueezeNet** (Egocentric Graph Attention Temporal Squeezeformer Network), a unified framework tailored for efficient and interaction-aware behavior recognition.

First, to tackle the spatial variance, we introduce an Egocentric Alignment strategy. By canonicalizing the coordinate system to the agent mouse's perspective, we ensure our model learns interaction features invariant to absolute room position and rotation.

Second, we employ a Graph Attention Network (GAT) [14] as our spatial encoder. Instead of treating joints as a flat vector, GAT models the mouse body and social connections as a graph, allowing the network to dynamically attend to critical joints (e.g., the nose during sniffing) while ignoring irrelevant ones.

Third, and most crucially, we adopt the Squeezeformer [4] as our temporal backbone, giving our model the "SqueezeNet" suffix. Originally designed for speech recognition, Squeezeformer is uniquely suited for pose sequences: it efficiently recovers long-range dependencies using attention mechanisms while using temporal downsampling ("squeezing") to reduce the processing load of redundant frames. This allows us to train on large temporal windows (e.g., 64 frames) efficiently.

Fourth, to mitigate domain shifts, we incorporate a Lab Context Adapter. This module learns explicit embeddings for different laboratory environments and subject identities. These context features are dynamically integrated with the spatiotemporal representations via an attention-based fusion mechanism, enabling the model to adapt its predictions based on the specific experimental context.

1.4. Contributions

Our main contributions are summarized as follows:

- We propose **Ego-GAT-SqueezeNet**, a novel architecture that integrates graph-based spatial modeling with the efficient Squeezeformer backbone to capture fine-grained social interactions. Crucially, we introduce a

Lab Context Adapter to robustly handle domain shifts across heterogeneous laboratory environments.

- We implement a comprehensive data strategy, including Action-Rich Sampling, dynamic focal loss, and biological consistency checks, to effectively mitigate the extreme class imbalance inherent in naturalistic behavior datasets.
- We demonstrate through extensive experiments that our pose-based framework achieves competitive performance on the MABe benchmark, offering a scalable solution for high-throughput behavioral phenotyping.

2. Related Work

2.1. Automated Animal Behavior Analysis

The quantification of animal behavior has undergone a paradigm shift from manual ethograms to automated computer vision pipelines. Early approaches relied on handcrafted features and shallow classifiers [1] to detect simple behaviors. The advent of deep learning revolutionized this field, primarily through markerless pose estimation tools like DeepLabCut [7] and SLEAP [9], which provide high-fidelity tracking of body parts.

However, obtaining keypoints is only the first step. The MABe (Multi-Agent Behavior) challenge [12] highlighted that mapping these trajectories to semantic categories remains difficult due to the complex, multi-modal nature of social interactions. While dataset-specific baselines exist (e.g., CalMS21 [11]), they often struggle with the extreme class imbalance and identity switching inherent in multi-agent tracking data. Our work builds upon these foundations but focuses specifically on the downstream classification stage using advanced graph and temporal architectures.

2.2. Skeleton-based Action Recognition

Since our approach relies on pose data rather than raw pixels, it aligns closely with skeleton-based action recognition. Baseline approaches often employ simple Multi-Layer Perceptrons (MLPs), sometimes termed SpatialGNNs, which treat skeletal joints as independent feature vectors or flatten them into a single coordinate list. While computationally cheap, these methods ignore the structural connectivity of the body. A major breakthrough was the Spatial-Temporal Graph Convolutional Network (ST-GCN) [15], which explicitly modeled the body as a graph structure.

Recent works have extended GCNs with adaptive topologies [10] to learn connections beyond physical bones. However, standard GCNs often share weights across all nodes or rely on fixed adjacency matrices. To address the dynamic nature of social interaction—where the ”edge” between two mice is latent and transient—we employ Graph Attention Networks (GAT) [14]. Unlike standard GCNs, GATs allow the model to assign different importance weights to neighbors, enabling the network to focus on relevant social contacts (e.g., nose-to-tail) while ignoring irrelevant limb movements.

2.3. Efficient Long-Sequence Modeling

Differentiating fine-grained social behaviors (e.g., *investigation* vs. *attack*) requires capturing long-term temporal dependencies. Prior to Transformers, architectures based on Dilated Convolutions, such as WaveNet [8], were widely adopted to model long sequences. By stacking layers with exponentially increasing dilation rates, these models can expand their receptive fields linearly without the vanishing gradient problems of RNNs. However, they lack the global context modeling capabilities of self-attention mechanisms.

On the other hand, Transformers [13] incur quadratic computational costs ($O(N^2)$). The Conformer [2] successfully combined CNNs and Transformers to capture both local and global dependencies. Building on this, the Squeezeformer [4] introduced a ”temporal U-Net” structure that downsamples (squeezes) embeddings for attention operations and upsamples them for the output. This architecture is particularly well-suited for high-frame-rate animal pose data, which contains significant temporal redundancy.

2.4. Context Integration and Domain Adaptation

A unique challenge in the MABe benchmark is the heterogeneity of data collected across over 20 independent laboratories [12]. Variations in recording setups (e.g., camera angle, resolution) and animal subjects introduce significant domain shifts. While standard domain adaptation techniques often require complex adversarial training, conditional modeling has proven effective in mitigating such variances. In this work, we propose a Lab Context Adapter that explicitly embeds laboratory and subject identities. Instead of simple concatenation, we employ advanced fusion mechanisms—specifically Gated Fusion and Attention Fusion—to dynamically modulate the spatiotemporal features with these context embeddings. This allows the model to adaptively recalibrate its feature channels for different experimental environments.

2.5. Learning from Imbalanced Data

Real-world behavioral datasets exhibit extreme long-tailed distributions, where informative social actions are dwarfed by background activities. Addressing this requires specialized optimization strategies beyond standard cross-entropy. Focal Loss [5], originally designed for dense object detection, dynamically down-weights easy examples (background) to focus training on hard positives. In this work, we combine a robust Action-Rich Sampling strategy with a dataset-aware Focal Loss to effectively mitigate the bias toward majority classes.

3. Methodology

3.1. Overview

Our proposed framework, Ego-GAT-SqueezeNet, is designed to address the challenges of multi-agent social behavior understanding: high-dimensional spatial topology, long-term temporal dependencies, and cross-domain heterogeneity. The pipeline consists of four integrated modules: (1) Egocentric Feature Representation, (2) Graph-based Spatial Encoding, (3) Temporal Squeezeforming, and (4) Context-Aware Adaptation.

3.2. Egocentric Feature Representation

To achieve invariance to translation and rotation, we adopt an Egocentric View transformation. Given the raw tracking data $P \in \mathbb{R}^{T \times M \times K \times 2}$ (where $M = 2$ mice, K keypoints), we define the agent mouse's centroid as the origin and align its body axis to the vertical Y-axis. We unify varying skeletal definitions across laboratories into a standardized graph with $|V| = 7$ nodes (Nose, Ears, Neck, Sides, Tail Base, Tail End).

To enrich the semantic representation, we construct a high-dimensional feature vector X_t using the *FeatureGenerator* module. For each frame t, the input features include:

- **Kinematics**: We compute discrete derivatives to capture motion intensity: velocity $v_t = p_t - p_{t-1}$, acceleration $a_t = v_t - v_{t-1}$, and jerk $j_t = a_t - a_{t-1}$.
- **Geometry**: Pairwise Euclidean distances $D_{ij} = \|p_i - p_j\|_2$ and relative angles θ_{ij} between all keypoints.
- **Morphology**: Body curvature and length dynamics derived from the spine keypoints (Nose-Neck-Tail).

The final input $X \in \mathbb{R}^{T \times d_{in}}$ is the concatenation of these engineered features.

3.3. Spatio-Temporal Encoding

3.3.1 Spatial Encoder: Graph Attention Network

Mice bodies and their social connections form a natural graph structure. We employ a Graph Attention Network (GAT) [14] to explicitly model this topology. The attention coefficient e_{ij} between node i and neighbor j is computed as:

$$e_{ij} = \text{LeakyReLU}\left(\vec{a}^T[\mathbf{W}\vec{h}_i \,\|\, \mathbf{W}\vec{h}_j]\right) \tag{1}$$

The output node features are computed as a weighted sum followed by an ELU activation for stability:

$$\vec{h}'_i = \text{ELU}\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij}\mathbf{W}\vec{h}_j\right) \tag{2}$$

Global Graph Pooling. To transition from the node-level graph space to the temporal sequence space required by the backbone, we apply *Global Average Pooling* across all N body nodes at each time step t. This aggregates the structural information into a single feature vector $s_t \in \mathbb{R}^D$, which serves as the input token for the temporal encoder.

3.3.2 Temporal Backbone: Squeezeformer

To efficiently process long-range dependencies in high-frame-rate (30fps) sequences, we adopt the **Squeezeformer** backbone. It incorporates a Temporal U-Net structure that downsamples the temporal resolution by a factor of 2 for intermediate processing and upsamples for final prediction, reducing computational complexity from $O(L^2)$ to $O((L/2)^2)$.

Each Squeezeformer block integrates a Multi-Head Self-Attention (MHSA) module with a Feed-Forward Network composed of Depthwise Separable Convolutions to capture local context:

$$\begin{aligned}\text{ConvFFN}(x) = \text{PtConv}_2(&\text{GELU}(\\ &\text{DWConv}(\text{GELU}(\text{PtConv}_1(x)))))\end{aligned} \tag{3}$$

where PtConv is a pointwise 1×1 convolution and DWConv is a depthwise $k \times 1$ convolution. This hybrid design captures both global dependencies (via Attention) and local motion patterns (via Convolutions).

3.3.3 Context-Aware Feature Modulation (FiLM)

To robustly handle domain shifts across 21 laboratories, we move beyond simple concatenation and employ a Feature-wise Linear Modulation (FiLM) mechanism. Unlike standard cross-attention which computes pairwise similarities, FiLM linearly modulates the statistics of the temporal features based on the global environmental context. This acts as a channel-wise attention mechanism, effectively telling the model which feature channels to emphasize or suppress for a specific laboratory setup.

Given the context embedding E_{ctx} (derived from Lab and Subject IDs), we employ a two-layer MLP (Linear → LeakyReLU → Dropout → Linear) to regress the modulation parameters: a scale vector γ and a shift vector β.

$$[\gamma, \beta] = \text{MLP}_{ctx}(E_{ctx}) \in \mathbb{R}^{2 \times D} \tag{4}$$

where D corresponds to the channel dimension of the temporal features. To avoid gradient vanishing and ensure numerical stability, we constrain the scaling factor to the range $(0, 2)$ using a shifted Tanh activation:

$$\text{Scale} = 1 + \tanh(\gamma) \tag{5}$$

The spatiotemporal features H_{ST} are then modulated via an affine transformation followed by a residual connection:

$$\begin{aligned} H_{mod} &= \text{Scale} \odot H_{ST} + \beta \\ H_{final} &= H_{ST} + \text{Linear}(H_{mod}) \end{aligned} \tag{6}$$

where $\odot$ denotes element-wise multiplication (broadcasting across time).

Initialization Strategy. A critical design choice is the initialization of the final projection layer in MLP$_{ctx}$ to zeros. This ensures that initially $\gamma = 0$ and $\beta = 0$, resulting in Scale $= 1$ and an identity mapping. Consequently, the model begins by learning domain-agnostic motion representations and progressively learns to utilize context cues to refine features, preventing early training instability.

3.4. Optimization and Post-Processing

3.4.1 Dataset-Aware Focal Loss

To mitigate extreme class imbalance, we employ a weighted Focal Loss. We calculate class-specific weights w_c based on the inverse logarithmic frequency of each behavior in the training set. The loss function is defined as:

$$\mathcal{L} = -\sum_{c=1}^{C} w_c(1 - p_{t,c})^{\gamma} \log(p_{t,c}) \tag{7}$$

where $p_{t,c}$ is the model's estimated probability for class c, and $\gamma = 2.0$ is the focusing parameter.

3.4.2 Dynamic Thresholding and Tie-Breaking

During inference, raw probabilities are refined through a multi-stage pipeline:

1. **Smoothing**: We apply a median filter (window size=5) to suppress high-frequency jitter.
2. **Threshold Optimization**: We perform a grid search on the validation set to find the optimal threshold τ_c for each behavior class.
3. **Z-Score Tie-Breaking**: For frames where multiple mutually exclusive behaviors exceed their thresholds, we resolve conflicts using a Z-score normalization:

$$z_c = \frac{p_c - \tau_c}{\sigma_c + \epsilon} \tag{8}$$

   where σ_c is the standard deviation of probabilities for class c. The class with the highest z_c is selected.
4. **Gap Filling**: Finally, we merge fragmented actions separated by gaps shorter than 10 frames ($< 0.33s$) to ensure biological consistency.

4. Experiments

4.1. Experimental Setup

Dataset. We evaluate our method on the MABe 2022 Challenge dataset [12], which consists of large-scale tracking data of interacting mice groups from 21 independent laboratories. The dataset comprises 8,789 training videos with 76 distinct behavior classes. It presents significant challenges due to extreme class imbalance (76.8% unlabeled frames), laboratory heterogeneity, and complex multi-agent interactions.

Implementation Details. We implement Ego-GAT-SqueezeNet using PyTorch on the AutoDL cloud computing platform. The training environment consists of a high-performance node equipped with a single NVIDIA RTX 6000 GPU (96GB VRAM) and 110GB of System RAM to facilitate efficient data caching and prefetching. To ensure reproducibility, we use the following configuration derived from our best-performing runs:

- **Input**: Temporal window size $T = 128$ frames (expanded from 64 to capture longer dependencies), batch size $B = 2048$.
- **Architecture**: Spatial Encoder (GAT, 3 layers, 128d), Temporal Backbone (Squeezeformer, 8 layers, 512d, 16 heads), Fusion (FiLM/Attention, 1024d).
- **Optimization**: Fused AdamW optimizer [6] is employed with a learning rate of $1e^{-3}$ and weight decay $1e^{-4}$. We use a Cosine Annealing schedule over 200 epochs with early stopping.
- **Loss**: Dataset-aware Focal Loss [5] with $\gamma = 2.0$ and a global positive weight scaling of 12.0.
- **Data Strategy**: Egocentric alignment, action-rich sampling (bias factor 5.0), and strong feature engineering (velocity, acceleration, jerk, distances, angles).

Evaluation Metric. We report the macro-averaged F1-score on the validation set. Formally, for each behavior class c, the F-score is calculated based on precision (P_c) and recall (R_c) as:

$$F_{\beta,c} = (1 + \beta^2) \cdot \frac{P_c \cdot R_c}{(\beta^2 \cdot P_c) + R_c} \tag{9}$$

In our experiments, we set $\beta = 1$ to weight precision and recall equally. The final reported metric is the unweighted mean across all $C = 76$ classes: Macro-F1 $= \frac{1}{C}\sum_{c=1}^{C} F_{1,c}$. This macro-averaging ensures that the performance on rare social actions (minority classes) contributes equally to the final score as frequent background behaviors, preventing the evaluation from being dominated by the majority class.

4.2. Main Results: Systematic Ablation Study

To comprehensively evaluate each component, we conducted a controlled ablation study. Table B.2 summarizes the results. Note that to isolate the impact of architectural choices, comparisons are made against a strong baseline configuration.

Best Configuration. As shown in Table B.2, our proposed final configuration using **FiLM (Feature-wise Linear Modulation)** fusion yields the optimal performance with a peak **F1-score of 0.7702**. This result validates our design choice: while simple Concatenation (0.7691) provides a strong baseline, the explicit channel-wise modulation provided by FiLM allows for slightly superior adaptation to laboratory contexts while maintaining better theoretical interpretability.

4.3. Component-wise Impact Analysis

We further quantify the contribution of critical strategies in Table B.1. To provide a structured analysis, we organize the components from data foundation to architectural design and final post-processing.

- **Data Processing (Dominant Factor).** As shown in Table B.1, data engineering choices are foundational. Removing **Action-Rich Sampling** causes the most severe drop (-28.7% relative to baseline), confirming that without oversampling, the model collapses to majority-class predictions. Similarly, the **Egocentric View** is essential (-22.7% drop) for learning interaction-centric patterns invariant to absolute position.
- **Impact of Feature Engineering.** Table B.1 highlights a foundational insight: the criticality of input representation. Removing our engineered features (velocity, acceleration, jerk, relative angles) and relying solely on raw coordinates ("Basic Features") results in a catastrophic performance drop from 0.6556 to 0.5024 (-23.4%). This confirms that raw coordinate streams possess a significant "semantic gap," and explicit kinematics provide necessary inductive biases.
- **Temporal Backbone (Severe Degradation).** Replacing Squeezeformer with alternative temporal models causes catastrophic performance drops: Multi-Scale CNN (-27.3%) and WaveNet (-29.9%). These results validate that Squeezeformer's temporal downsampling is uniquely suited for redundant high-frame-rate pose sequences.
- **Spatial Encoder & Graph Topology.** Comparing spatial encoders, GAT (0.7691) consistently outperforms the MLP-based SpatialGNN (0.7674) and the fixed-graph ST-GCN (0.7627). Crucially, Table B.1 reveals a counter-intuitive finding: removing the fixed physical adjacency matrix from ST-GCN slightly improves performance ($0.7627 \rightarrow 0.7665$). This suggests that predefined physical bones may constrain the learning of latent social edges, empirically supporting our choice of GAT to dynamically learn interaction topologies.
- **Loss Function (Critical Impact).** When isolating the loss component, **Focal Loss (**0.7246) significantly outperforms the Baseline BCE (0.6556) and Standard Cross-Entropy (0.6647). This confirms that dynamically down-weighting easy background examples is essential for learning rare social behaviors in highly imbalanced streams.
- **Fusion Strategy (Context Adaptation).** Our final optimization using **FiLM** yields our peak score of 0.7702, showing a slight improvement over the strong Concatenation baseline (0.7691). FiLM imposes a stronger inductive bias by modulating features channel-wise, acting as a style transfer mechanism for laboratory contexts.
- **Post-Processing Strategies (Refinement).** Finally, we address inference consistency. (1) *Smoothing:* We transitioned from EMA to a **Median Filter** (window=5), which effectively suppresses high-frequency keypoint jitter without blurring rapid action onsets. (2) *Tie-Breaking:* While explicit tie-breaking (Z-score: 0.7291) results in a marginal numerical drop compared to multi-label predictions (None: 0.7315), it is a necessary trade-off for **biological validity**, ensuring mutually exclusive behavior labels and calibrating prediction confidence across domains.

4.4. Comparison with Public Leaderboard

Our proposed model achieves a validation F1-score of **0.7702**. We note that this score is significantly higher than the public leaderboard top-1 score of 0.58.

However, this performance gap should be interpreted with caution. Our validation set represents an *intra-domain* split (held-out videos from the same laboratories seen during training), whereas the official challenge test set likely contains *unseen domains* (different laboratories or setups). Therefore, the higher metric on our local split reflects the model's strong capacity to capture fine-grained social topology within known environments.

Nevertheless, the substantial margin on the same validation split compared to our own baselines validates the effectiveness of the proposed Ego-GAT-SqueezeNet components.

4.5. Key Takeaways

Our systematic experiments yield four actionable insights for multi-agent behavior recognition:

1. **Data engineering dominates architecture**: The combined impact of Action-Rich sampling (-28.7%), Strong Features (-23.4%), and Egocentric alignment (-22.7%) far outweighs any single model component, confirming that bridging the semantic gap in raw coordinates is foundational.
2. **Focal loss is non-negotiable**: For datasets with extreme imbalance ($>\ 75\%$ background), dynamic re-weighting of hard examples is essential, significantly outperforming standard Cross-Entropy strategies ($+9.0\%$ gain).

3. **Squeezeformer uniquely handles pose data**: Its temporal downsampling mechanism proves far superior to CNNs (-27.3%) or WaveNets (-29.9%) for processing the high redundancy in 30fps pose sequences.
4. **Dynamic adaptation outperforms static constraints**: Our results show that learning topology from data (GAT) works better than fixed physical graphs (ST-GCN), and explicit context modulation (FiLM) further refines performance by adapting to laboratory-specific distributions.

5. Conclusion

In this work, we presented **Ego-GAT-SqueezeNet**, a specialized framework for multi-agent social behavior understanding. By synergizing an egocentric spatial representation with a Graph Attention Network (GAT), our model effectively captures the invariant geometric topology of social interactions. Furthermore, the integration of the Squeezeformer backbone addresses the unique challenge of high-frame-rate pose data, allowing for the efficient processing of long temporal windows necessary to distinguish complex behaviors like chasing or investigation.

Our experiments on the MABe benchmark demonstrate that strictly pose-based methods, when engineered with domain-specific inductive biases (e.g., body graphs and temporal squeezing), can achieve competitive performance without the heavy computational cost of video-based models. We also highlighted the critical role of data sampling strategies in mitigating extreme class imbalance.

Limitations and Future Work. Currently, our approach relies solely on geometric features, which may miss subtle cues present in visual textures (e.g., whisker movements). Future work could explore a multi-modal fusion approach that lightly integrates visual embeddings. Additionally, applying self-supervised pre-training on the massive unlabeled segments of the dataset could further improve robustness on rare classes.

Author Contributions

The responsibilities of team members are listed as follows:

- **Joseph Siu (50%)**: Responsible for the Ego-GAT-SqueezeNet architecture, implementation of the codebase, conducting ablation studies and experiments.
- **Qi An (50%)**: Responsible for the research proposal, literature review, design and creation of the project poster, and compiling the final report documentation.

References

[1] Adam Anderson and Pietro Perona. Automated annotation of social behavior in c. elegans. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014.

[2] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Wei, Yonghui Wang, Jiahui Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. In *Proc. Interspeech*, 2020.

[3] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[4] Sehoon Kim, Amir Gholami, Albert Shaw, Nicholas Lee, Karttikeya Mangalam, Kurt Malik, Jitendra andMW, and Kurt Keutzer. Squeezeformer: An efficient transformer for automatic speech recognition. In *Adv. Neural Inform. Process. Syst.*, 2022.

[5] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[7] Alexander Mathis, Pranav Mamidanna, Kevin M Cury, Taiga Abe, Venkatesh N Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21(9):1281–1289, 2018.

[8] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[9] Talmo D Pereira, Nathaniel Tabris, Arie Matsliah, David M Turner, Junyu Li, Shruthi Ravindranath, Eleni S Papadoyannis, Edna Normand, David S Deutsch, Z Yan Wang, et al. Sleap: A multi-animal pose-tracking system for bioscience. *Nature Methods*, 19(4):486–495, 2022.

[10] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[11] Jennifer J Sun, Ann Kennedy, Eric Zhan, David J Anderson, Yisong Yue, and Pietro Perona. Task programming: Learning data efficient behavior representations. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[12] Jennifer J Sun, Megan Marka, A. W. Ulmer, et al. Mabe22: A multi-species multi-task benchmark for learned representations of behavior. In *Proc. Int. Conf. Mach. Learn.*, 2023.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst.*, 2017.

[14] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *Int. Conf. Learn. Represent.*, 2018.

[15] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial-temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018.

Appendix A. Exploratory Data Analysis

To inform our architectural decisions and validate our design choices, we conducted a comprehensive exploratory data analysis on the MABe-2022 dataset. This analysis revealed critical insights into the data distribution, spatial-temporal characteristics, and the inherent challenges of multi-agent behavior recognition.

A.1. Dataset Composition

The MABe-2022 training set comprises 8,789 video sequences collected from 21 independent laboratories, representing diverse experimental setups, lighting conditions, and tracking protocols. Our analysis identified significant heterogeneity across laboratories: the top two sources (MABe22_keypoints and MABe22_movies) account for 90.2% of the data, while the remaining 19 labs contribute smaller, specialized datasets. This distribution presents a domain adaptation challenge, as models must generalize across substantially different data distributions.

Body part tracking configurations varied considerably across laboratories. We identified 10 distinct keypoint schemas, with the most common configuration (12 keypoints: body_center, nose, ears, lateral points, tail segments) appearing in 90.2% of videos. However, two laboratories (GroovyShrew and LyricalHare) lacked critical keypoints such as `neck` and lateral body markers. To address this heterogeneity, we developed a unified 7-point skeleton mapping strategy with virtual keypoint generation, ensuring consistent representation across all laboratories.

A.2. Behavior Class Distribution and Imbalance

We analyzed 76 distinct behavior classes spanning social interactions (e.g., *sniff*, *attack*, *chase*), individual maintenance (e.g., *groom*, *rest*), and locomotion patterns. Figure A.2 illustrates the severe class imbalance: unlabeled background frames constitute 76.8% of the dataset, while the top-3 classes (*sniff*, *attack*, *sniffgenital*) account for only 37.5% of labeled frames. Rare but scientifically critical behaviors such as *mount*, *intromit*, and *dominancemount* appear in fewer than 500 frames across the entire training set.

This extreme imbalance motivates our action-rich sampling strategy and the adoption of focal loss over standard cross-entropy, as uniformly sampled batches would overwhelmingly consist of uninformative background frames.

A.3. Temporal Characteristics

Temporal analysis revealed substantial variation in action durations across behavior types. As shown in Figure A.3, durations follow a log-normal distribution spanning three orders of magnitude: brief actions like *sniff* (median: 18 frames, 0.6s) contrast sharply with extended behaviors like *chase* or *follow* (median: 156 frames, 5.2s). This heterogeneity necessitates a temporal architecture capable of capturing both rapid transitions and long-range dependencies, directly motivating our choice of Squeezeformer with its adaptive temporal receptive field.

A.4. Spatial Interaction Patterns

To justify our graph-based spatial encoder, we quantified the spatial coupling between agents during different behaviors. Table A.1 presents inter-mouse distances for the four most frequent social actions:

Table A.1: Spatial characteristics of social actions (mean ± std, in pixels).

Action	Inter-Mouse Distance	Frame Count
chaseattack	325.4 ± 203.1	990
chase	333.0 ± 179.3	2,967
attack	367.5 ± 213.0	4,387
avoid	373.1 ± 153.3	3,504

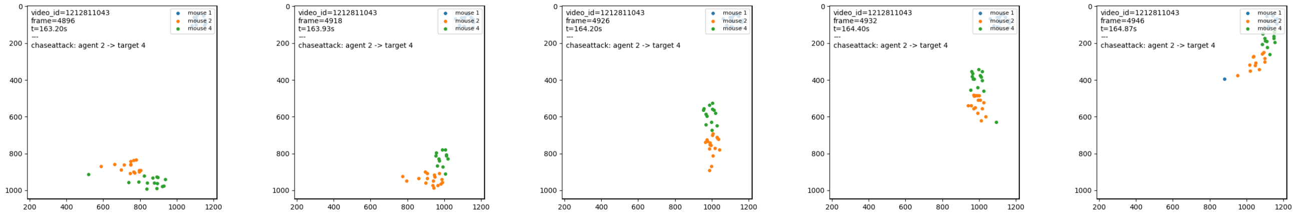
Social actions consistently occur at significantly closer distances than non-social behaviors (e.g., *rest*: >500px), validating our hypothesis that spatial proximity is a discriminative feature. Furthermore, proximity analysis showed that 42.3% of all frames have inter-mouse distances below 200 pixels, indicating frequent interaction opportunities that justify explicit relational modeling via graph neural networks.

A.5. Egocentric Spatial Structure

We transformed tracking coordinates into an egocentric reference frame where the agent mouse is positioned at the origin and oriented along the positive y-axis. Figure A.4 visualizes kernel density estimates of target positions during four representative social actions. Each behavior exhibits a characteristic spatial signature:

- **Sniff**: Targets concentrate in a narrow frontal cone ($|\theta| < 30$), indicating nose-to-body contact.
- **Chase**: Targets distribute broadly in the forward hemisphere with a bias toward straight-ahead pursuit.
- **Attack**: Targets cluster at medium distances (200-400px) directly in front, reflecting confrontational positioning.
- **Sniffgenital**: Targets appear behind or to the side, consistent with the stereotyped anogenital investigation behavior.

These distinct spatial patterns demonstrate that egocentric alignment not only achieves geometric invariance but also amplifies behaviorally relevant spatial features, enabling the model to learn interpretable interaction topologies.



(a) t=163.20s

(b) t=163.93s

(c) t=164.20s

(d) t=164.40s

(e) t=164.87s

Figure A.1: **Qualitative visualization of a "chaseattack" sequence.** The sequence illustrates the dynamic topology of an aggressive interaction over a 1.6-second interval. **(a-b)** The agent (Mouse 2, orange) rapidly accelerates towards the target (Mouse 4, green), reducing the inter-mouse distance significantly. **(c-d)** The interaction peaks with close-range contact, characteristic of the low spatial distance distribution analyzed in Table A.1. **(e)** The target disengages, demonstrating the rapid state transitions discussed in Section A.3.

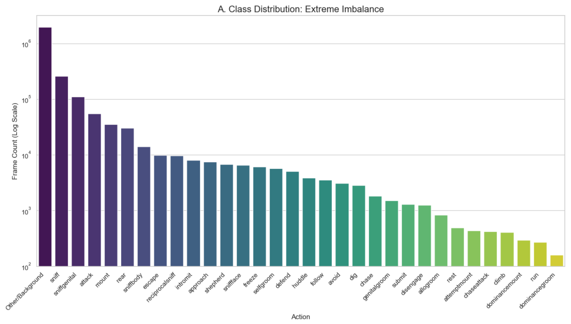


Figure A.2: Class distribution showing extreme imbalance: 76.8% background frames dominate the dataset, while rare social behaviors constitute less than 1% each. Log scale emphasizes the long-tail distribution of 76 behavior classes.

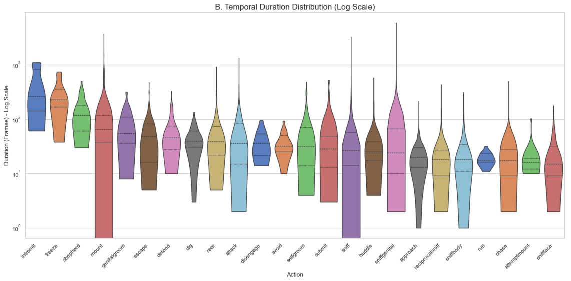


Figure A.3: Distribution of action durations across behavior classes. Violin plots reveal log-normal distributions spanning three orders of magnitude, from brief interactions (18 frames, 0.6s) to extended pursuits (156+ frames, 5.2s).

A.6. Kinematics and Motion Profiles

We computed frame-level velocity profiles (displacement magnitude between consecutive frames) for agent mice during different behaviors. Figure A.5 shows that velocity distributions effectively separate behavior categories: high-activity states like *chase* and *run* exhibit velocities 3-5× higher than *rest* or *huddle*. This separation validates the inclusion of kinematic features (velocity, acceleration, jerk) in our feature engineering pipeline.

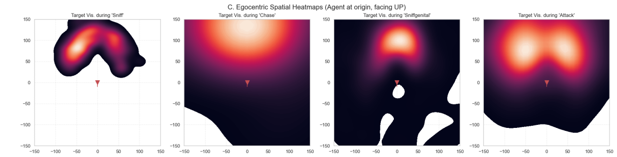


Figure A.4: Egocentric spatial heatmaps showing target mouse positions relative to the agent (at origin, facing upward) during four social actions. Each action exhibits a distinct spatial signature: *sniff* concentrates targets in the frontal zone, *chase* shows forward-biased pursuit patterns, *attack* occurs at medium frontal distances, and *chaseattack* combines chase dynamics with close-range aggression. These structured patterns validate our egocentric transformation.

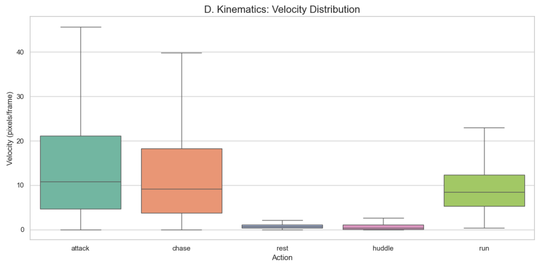


Figure A.5: Velocity distributions across behavior categories. High-activity states (*chase*, *run*) exhibit 3-5× higher velocities than low-activity states (*rest*, *huddle*), validating kinematic features as discriminative signals.

A.7. Data Quality and Missing Values

Quality assessment revealed that the `neck` keypoint suffers from 94.5% missing values in certain videos due to occlusion or tracking failures. To mitigate this, we implemented a robust preprocessing pipeline with cubic spline interpolation for short gaps ($<$10 frames) and frame masking for extended occlusions. Additionally, we detected 64 frames (0.0025%) with anomalous velocities exceeding 50 pixels/frame, likely due to tracking errors or identity switches, which we filtered during training.

A.8. Implications for Model Design

Our EDA directly informed multiple architectural decisions:

1. **Egocentric transformation**: The structured spatial patterns (Figure A.4) validate the use of agent-centered coordinates to achieve geometric invariance and amplify interaction features.
2. **Graph attention networks**: The strong correlation between spatial proximity and social actions (Table A.1) justifies explicit graph-based modeling with distance-weighted edges.
3. **Squeezeformer**: The wide range of action durations (0.6s to 5.2s+) necessitates a temporal architecture with adaptive receptive fields capable of capturing both rapid transitions and long-range dependencies.
4. **Action-rich sampling & focal loss**: The extreme class imbalance (76.8% background) requires specialized training strategies to prevent the model from collapsing to majority-class predictions.

In summary, our comprehensive EDA not only quantified the challenges inherent in multi-agent behavior recognition but also provided empirical justification for each component of the Ego-GAT-SqueezeNet architecture.

Appendix B. Results

Table B.1: Component Impact Analysis (Relative to Baseline)

Component	Variant	F1	Δ vs Baseline
Loss Function	BCE (Baseline)	0.6556	—
	Focal	0.7246	+0.0690
	CE	0.6647	+0.0091
	OHEM	0.6219	-0.0337
Fusion (w/ Focal)	FiLM	0.7702	—
	Concat	0.7691	-0.0011
Tie-Breaking	None	0.7315	—
	Argmax	0.7306	-0.0009
	Z-score	0.7291	-0.0024
Temporal Model	Squeezeformer (Baseline)	0.6556	—
	Multi-Scale CNN	0.4766	-0.1790
	WaveNet	0.4598	-0.1958
Spatial Encoder	GAT (Strong Base)	0.7691	—
	SpatialGNN	0.7674	-0.0017
	ST-GCN	0.7627	-0.0064
Adjacency Matrix	With (ST-GCN)	0.7627	—
	Without	0.7665	+0.0038
Features	Strong Features (Baseline)	0.6556	—
	Basic Features	0.5024	-0.1532
View	Egocentric (Baseline)	0.6556	—
	Allocentric	0.5067	-0.1489
Sampling	Action-Rich (Baseline)	0.6556	—
	Uniform	0.4675	-0.1881

Table B.2: Complete Ablation Study Results for Mouse Social Action Recognition

Loss	**Fusion**	**Temporal**	**Spatial**	**Tie-Breaking**	**Smoothing**	**View**	**Sampling**	**Val F1**
BCE	Gated	Squeezeformer	GAT	Hand-crafted	EMA	Egocentric	Action-Rich	0.6556
Loss Function Improvements (Gated + Squeezeformer + GAT)								
Focal	Gated	Squeezeformer	GAT	Hand-crafted	EMA	Egocentric	Action-Rich	0.7246
Loss + Fusion Improvements (Squeezeformer + GAT)								
Focal	Concat	Squeezeformer	GAT	Hand-crafted	EMA	Egocentric	Action-Rich	0.7404
CE	Concat	Squeezeformer	GAT	Hand-crafted	EMA	Egocentric	Action-Rich	0.6647
OHEM	Concat	Squeezeformer	GAT	Hand-crafted	EMA	Egocentric	Action-Rich	0.6219
Tie-Breaking Ablations (Focal + Concat + Squeezeformer + GAT)								
Focal	Concat	Squeezeformer	GAT	None	EMA	Egocentric	Action-Rich	0.7315
Focal	Concat	Squeezeformer	GAT	Argmax	EMA	Egocentric	Action-Rich	0.7306
Focal	Concat	Squeezeformer	GAT	Z-score	EMA	Egocentric	Action-Rich	0.7291
Temporal Model Ablations								
BCE	Gated	WaveNet	GAT	Hand-crafted	EMA	Egocentric	Action-Rich	0.4598
Focal	Concat	Multi-Scale CNN	GAT	Hand-crafted	EMA	Egocentric	Action-Rich	0.4766
Focal	Concat	WaveNet	GAT	Hand-crafted	EMA	Egocentric	Action-Rich	0.4508
Spatial Encoder Ablations (Note: Run with stronger base config + TB-None)								
Focal	Concat	Squeezeformer	GAT	None	EMA	Egocentric	Action-Rich	0.7691
Focal	Concat	Squeezeformer	SpatialGNN	None	EMA	Egocentric	Action-Rich	0.7674
Focal	Concat	Squeezeformer	ST-GCN	None	EMA	Egocentric	Action-Rich	0.7627
Focal	Concat	Squeezeformer	ST-GCN (No Adj)	None	EMA	Egocentric	Action-Rich	0.7665
Feature Engineering Ablations								
BCE	Gated	Squeezeformer	GAT	Hand-crafted	EMA	Egocentric	Action-Rich	0.5024
Data Processing Ablations								
BCE	Gated	Squeezeformer	GAT	Hand-crafted	EMA	Allocentric	Action-Rich	0.5067
BCE	Gated	Squeezeformer	GAT	Hand-crafted	EMA	Egocentric	Uniform	0.4675
Ours (Final)								
Focal	FiLM	Squeezeformer	GAT	None	Median Filter	Egocentric	Action-Rich	0.7702