APPENDIX

A DATASETS

A.1 UEA DATASETS

Table 1: Selected datasets from UEA Multivariate Time Series Classification archive.

Dataset	Train	Test	Channels	Length	Classes
MotorImagery	278	100	64	3000	2
SelfRegSCP2	200	180	7	1152	2
FaceDetection	5890	3524	144	62	2
Ethanol	261	263	3	1751	4

A.2 NEURAL DATASETS

Table 2: Neural datasets.

Dataset	Train	Test	Units	Length	Classes
NLB-Maze	1721	574	182	140	-
NLB-RTT	810	270	130	120	-
Mihi-Day 1	167	42	162	76/84	8
Mihi-Day 2	172	43	152	73/87	8
Chewie-Day 1	127	32	163	81	8
Chewie-Day 2	144	36	148	75/77	8

B EXPERIMENT DETAILS

B.1 MODEL IMPLEMENTATION AND HYPERPARAMETERS

For the patching tokenization layer of PatchTST and GAFormer, we set the patch window as 10 for SelfRegSCP2, MotorImagery, and Ethanol, set the patch window as 2 for FaceDetec since the data length is short. For all neural datasets (Mihi-Chewie, NLB-Maze, NLB-RTT), we set the patch window as 1 to better evaluate the potential of our model for studying fine-grained dynamics of neural activities. For all experiments and all models including GRU, TCN, MVTS, AutoTrans, the token dimension and embedding size are 256. For Transformer layers, the number of head is set as 16. In GAFormer, we set the depth of spatial/temporal group embedding module and spatial transformer encoder as 3, and the depth for final temporal transformer encoder as 6. For the added TGE module to other baseline models, we kept the depth and dimension consistent with the TGE in GAFormer. For fair comparison, the depth of MVTS, AutoTrans, NDT and EIT are kept the same as the total depths of spatial and temporal Transformer layers in GAFormer. We set the number of groups to be 10 in all group embedding modules. Following Zerveas et al. (2021) and Nie et al. (2022), we adopt batch normalization rather than layer normalization in all Transfomer architectures.

B.2 TRAINING

We did not use pretraining or data augmentations for all experiments except for the MotorImagery dataset. In the MotorImagery dataset, the number of samples is limited and the length of each trial

TGE	T-Encoder	Dim	Head	Group K	Patch size
3	6	256	8	10	10

Table 3: *Hyperparameters used for univariate datasets.*

Dataset	Channel-independent Encoder	SGE	S-Encoder	TGE	T-Encoder	Dim	Group K	Patch Size
FaceDetect	3	3	3	3	6	256	8	10
MotorImagery	3	3	3	3	6	256	8	10
SelfRegSCP2	3	3	3	3	8	256	8	10
Ethanol	3	3	3	3	8	256	8	10
Mihi-Chewie	3	3	3	3	6	256	16	1
NLB	3	3	3	3	8	256	8	1

Table 4: *Hyperparameters used for multivariate datasets.* The first 5 columns represent the depth of each transformer component.

(3000 timepoints) is too long even with patching tokenization. Thus we split each test sample from 3000 timepoints to be 3 samples of 1000 timepoints and formed a larger dataset. In the training stage, we randomly select a segment of 1000 timepoints from each training trial as a strong data augmentation. For models with patching tokenization, we set the batch size as 32. For models with no patching tokenization (MVTS, AutoTrans), we set the batch size as 16 since the number of tokens is large.

C Additional Visualizations

C.1 Visualizations on real-world datasets

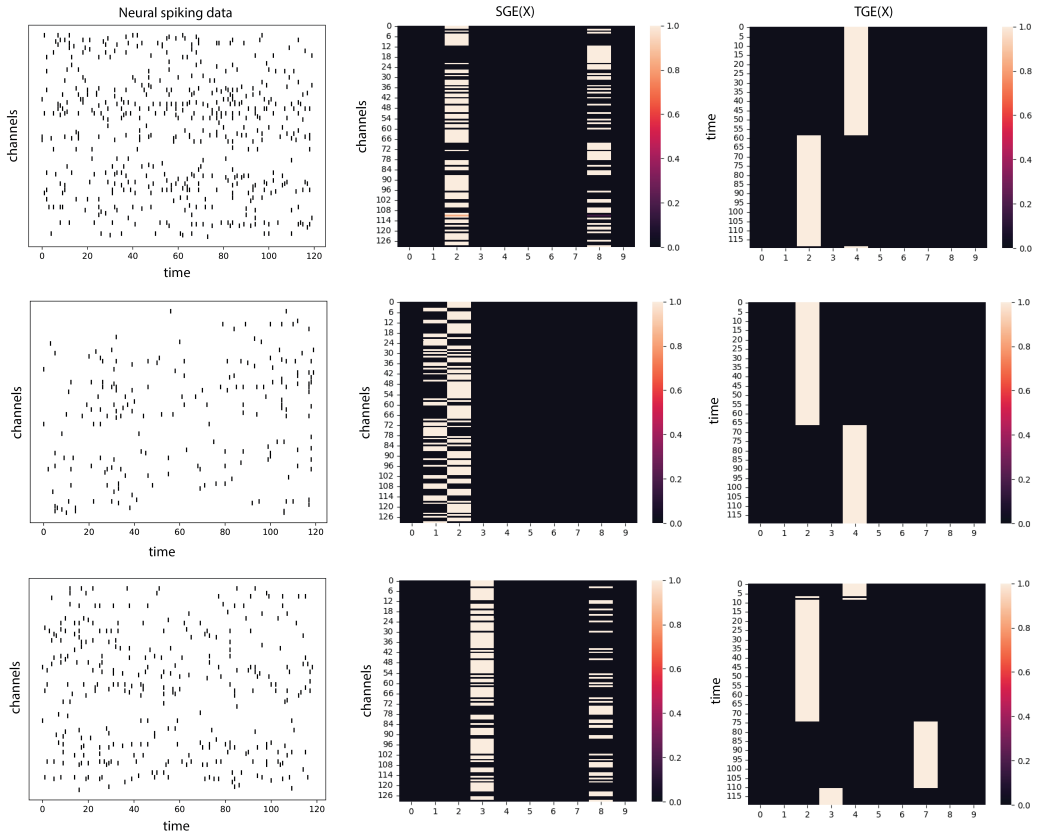

Figure 1: *Visualization of Group Embeddings for Neural Spiking Datasets.* (Left) Spiking data, (Middle) Spatial group embeddings, and (Right) Temporal Group Embeddings for the Neural Latents Benchmark Random Target Task.

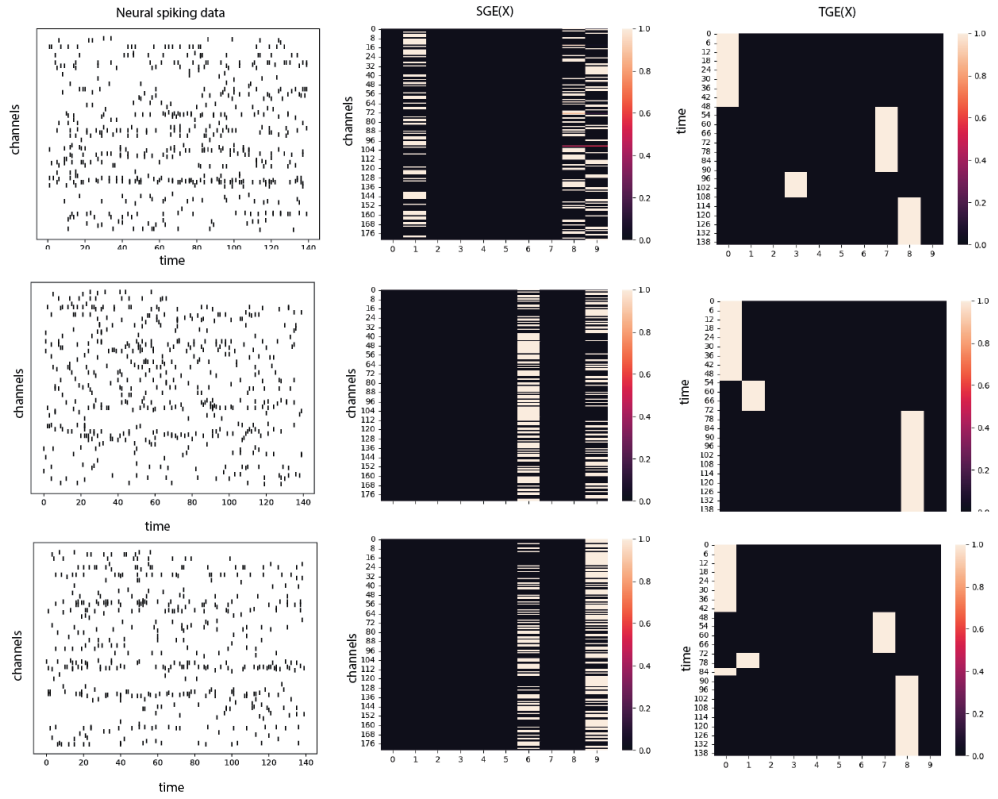

Figure 2: *Visualization of Group Embeddings for Neural Spiking Datasets.* (Left) Spiking data, (Middle) Spatial group embeddings, and (Right) Temporal Group Embeddings for the Neural Latents Maze Task.

As shown on the right in Figure 2, we observed a consistent cutoff in the temporal group assignments, where the initial 50 timepoints were mostly categorized as group 1. We notice that this corresponds to the onset time (at 250ms in each trial, which is binned as the 50th timepoints with bin size of 50ms) of the movements of monkey subjects in the studied **MC_Maze** subdataset (Ye & Pandarinath, 2021). This grouping structure indicates that `GE` facilitates the encoder's ability to discern nuanced, common structural patterns throughout the dataset.

C.2 Visualizations on synthetic datasets

We provide additional visualizations of token representations on the synthetic many-body datasets, as shown in Figure 3, Figure 4, and Figure 5.

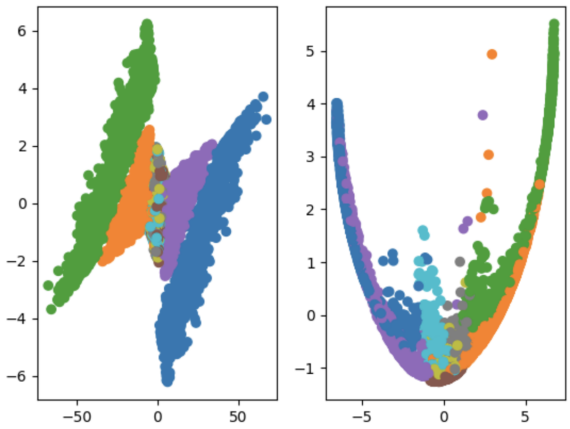

Figure 3: Token embeddings at the input (left) and output (right) layers for a transformer encoder trained with learned position embeddings (PE). Different colors represent the x and y axis of object-1 (green and orange), object-2 (blue and purple), and two random noise objects.

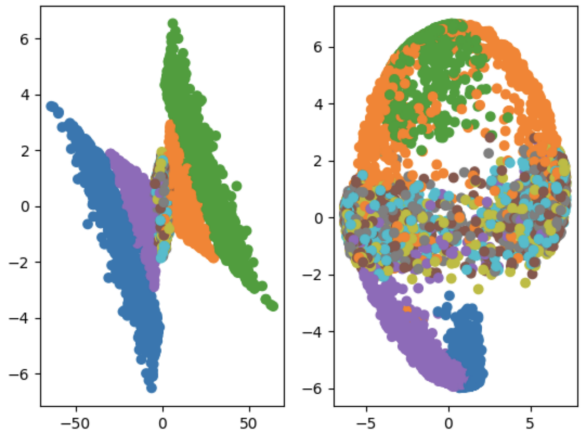

Figure 4: Token embeddings at the input (left) and output (right) layers for a transformer encoder trained with group embeddings (GE). Different colors represent the x and y axis of object-1 (green and orange), object-2 (blue and purple), and two random noise objects.

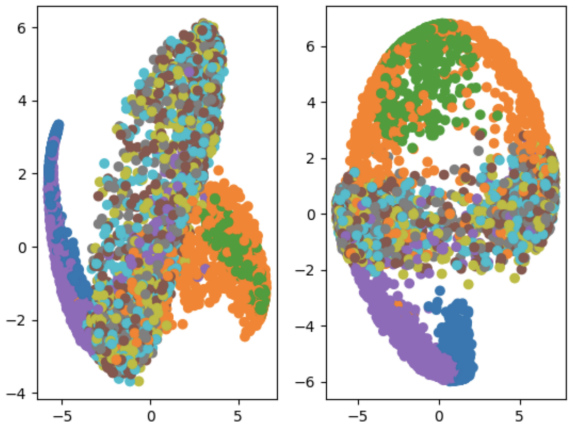

Figure 5: Token embeddings at the output for the same frozen transformer backbone without (left) and with (right) GE. Surprisingly, with the same backbone, applying group embeddings at the input can produce representations that are more separable. Different colors represent the x and y axis of object-1 (green and orange), object-2 (blue and purple), and two random noise objects.