APPENDIX FOR BSTT: A BAYESIAN SPATIAL-TEMPORAL TRANSFORMER FOR SLEEP STAGING

A APPENDIX

A.1 CALCULATION DETAILS

A.1.1 CALCULATION OF EVALUATION INDICATORS

We use accuracy (ACC), F1 Score, and KAPPA to evaluate the our BSTT model and baseline models, the specific calculations are follows:

$$\mathrm{ACC} = \frac{\sum_{\mathrm{C}-1}^{\mathrm{C}} \boldsymbol{TP}_{\mathrm{C}}}{\mathrm{N}} \tag{1}$$

$$\text{F1 Score} = \frac{\sum_{\mathrm{C}=1}^{\mathrm{C}} \boldsymbol{F1}_{\mathrm{C}}}{\mathrm{C}} \tag{2}$$

$$\mathrm{K}_{\mathrm{KAPPA}} = \frac{\boldsymbol{P}_{\mathrm{o}} - \boldsymbol{P}_{\mathrm{e}}}{1 - \boldsymbol{P}_{\mathrm{e}}} \tag{3}$$

where $\boldsymbol{TP}_{\mathrm{C}}$ is the true positive for class C, N is the total number of recordings, $\boldsymbol{F1}_{\mathrm{C}}$ is the F1 score of per class, C is the number of sleep stages, $\boldsymbol{P}_{\mathrm{o}}$ is the observed agreement, and $\boldsymbol{P}_{\mathrm{e}}$ is the agreement by chance.

A.1.2 CALCULATION OF POSITION EMBEDDING MATRIX

We use the position embedding function in groundbreaking work (Vaswani et al., 2017) to calculate the position embedding of Bayesian transformer module. The calculation of position embedding matrix is defined as:

$$\mathbf{P}^{\mathrm{ep}}_{\mathrm{i,2j}} = \sin\left(\frac{i}{10000^{2j/F}}\right) \tag{4}$$

$$\mathbf{P}^{\mathrm{ep}}_{\mathrm{i,2j+1}} = \cos\left(\frac{i}{10000^{2j/F}}\right) \tag{5}$$

A.2 BASELINE DETAILS

We compare the Bayesian spatial-temporal transformer with the following baselines:

- MCNN (Chambon et al., 2018): a deep learning sleep staging method utilizing multivariate multimodal PSG signals.
- MMCNN (Jia et al., 2020a): a multi-scale convolutional neural network for EEG signal classification.
- MLP+LSTM (Dong et al., 2017): a mixed neural network, which combines multilayer perceptron (MLP) and LSTM.
- DeepSleepNet (Supratak et al., 2017): a mixed sleep staging method utilizing CNN and Bi-LSTM.
- TinySleepNet (Supratak & Guo, 2020): an efficient and lightweight EEG sleep staging network.
- U-Time (Perslev et al., 2019): a temporal fully convolutional network based on U-Net architecture for sleep staging.

- GraphSleepNet (Jia et al., 2020b): a spatial-temporal graph convolutional neural network that can adaptively learn spatial-temporal features.
- Spatial-Temporal Transformer (Phan et al., 2022): a sleep staging method using transformer to extract spatial-temporal features.

A.3 VARIATIONAL INFERENCE IN BAYESIAN RELATION INFERENCE

Inspired by the variational RNN (VRNN)(Huang et al., 2020), we propose to employ variational inference to optimise our Bayesian relation inference component. VRNN introduces a latent variable $z_{i,t}$ to encode the uncertainty of input features at time t, which is assumed to have a Gaussian prior distribution $p\left(\mathbf{z}_{i,t} \mid \mathbf{h}_{i,t-1}\right)$ conditioned on the previous RNN hidden state $h_{i,t-1}$. The posterior distribution of this latent variable is approximated by a variational distribution $q\left(\mathbf{z}_{i,t} \mid \mathbf{x}_{i,t}, \mathbf{h}_{i,t-1}\right)$, allowing us to utilise the evidence lower bound (ELBO) for joint learning and inference. This can be formulated as follows:

$$\begin{aligned}\sum_{i=1}^{M} \{ & \mathrm{KL}\left(q\left(\mathbf{Z}_i \mid \mathbf{X}_i, \mathbf{H}_i\right) \| p\left(\mathbf{Z}_i \mid \mathbf{H}_i\right)\right) \\ & -\mathbb{E}_{\mathbf{Z}_i}\left[\log P\left(\mathbf{Y}_i \mid \mathbf{X}_i, \mathbf{Z}_i\right)\right]\}\end{aligned} \tag{6}$$

where $\mathbf{Z}$ is latent variables. In Bayesian relation inference component, random variables $\tilde{\mathbf{A}}$ and $\mathbf{S}$ can completely determines the random process in Bayesian relation inference as a whole. The Eq.(6) in VRNN can be rewritten as follow:

$$\sum_{i=1}^{M}\left\{\mathrm{KL}\left(q\left(\tilde{\mathbf{A}}, \mathbf{S} \mid \mathbf{X}_{0:\mathrm{i}}\right) \| p\left(\tilde{\mathbf{A}}, \mathbf{S} \mid \mathbf{X}_{0:\mathrm{i}}\right)\right)-\mathbb{E}_{\tilde{\mathbf{A}}, \mathbf{S}}\left[\log P\left(\mathbf{Y}_{\mathrm{i}} \mid \mathbf{X}_{\mathrm{i}}, \tilde{\mathbf{A}}, \mathbf{S}\right)\right]\right\} \tag{7}$$

A.4 PIPELINE GRAPH OF BAYESIAN SPATIAL-TEMPORAL TRANSFORMER

Figure 1 shows the overall of our Bayesian spatial-temporal transformer (BSTT). The input of BSTT is multi-channel EEG signals with time context. The embedded features first pass through the Bayesian spatial transformer module to infer the spatial relations among channels and model their spatial features. The Bayesian temporal transformer module infer the temporal relations and model the temporal features.

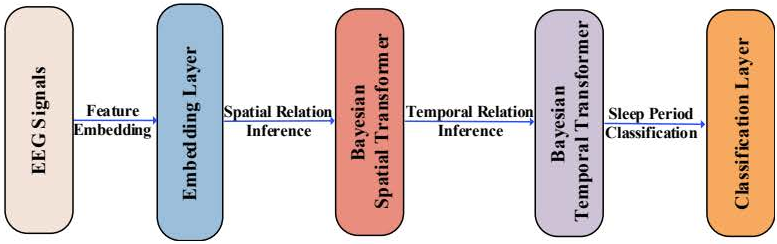

Figure 1: The pipeline graph of Bayesian spatial-temporal transformer.

.

A.5 EXPERIMENT CONFIGURATION AND DETAILS

A.5.1 TRAINING DETAILS

The details of the resources for training and the versions of the software are provided in Table 1. The details of the hyperparameters are provided in Table 2.

Table 1: The hardware and software configuration for training.

Software	Python	3.7.11
	PyTorch	1.10.0
	numpy	1.21.2
Hardware	CPU	Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz
	RAM	128 GB
	GPU	GeForce RTX 2080

Table 2: The configuration of hyper-parameters for training.

Hyper-parameter	Value
HiddenDim	256
GraphDim	256
Epoch	60
Layer_num_rnn	5
Spatial Lambda1	5e-5
Spatial Lambda2	4e-5
Temporal Lambda1	5e-5
Temporal Lambda2	4e-5
Heads	5
Time Series	5
Weight Decay	3e-5

A.6 Supplementary Experiment Results

A.6.1 Supplementary Visual Analysis

Figure 2 shows the intensity graphs of the temporal relation during some specific periods. Specifically, we select the temporal intensity graphs when performing mutual transitions between REM period and N2 period and during REM period for visual analysis. The proposed model pays more attention to the longer-interval relations when interpreting the duration of the REM period. The AASM standard states that the rule of continuous interpretation in REM period requires no rapid eye movement or significant waveform changes in one or more epochs after that (Berry et al., 2012). Hence, sleep specialists tend to focus on the long-term relation intensity and the association of epochs following the target sleep epoch, which is consistent with our experimental results. Furthermore, in our model, the process of mutual transition between REM period and N2 period tends to focus on the association between the last two sleep epochs. The AASM standard mentions that when interpreting the mutual transition between REM period and N2 period, it is necessary to ensure that there is no additional transition of the subsequent EEG signals, otherwise it may be reinterpreted back to the original sleep period (Berry et al., 2012). Our experimental results also confirm this result.

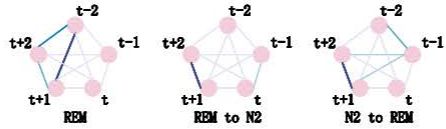

Figure 2: Temporal relation intensity graphs under specific sleep periods or sleep transition processes.

.

A.6.2 ABLATION EXPERIMENT RESULTS

Table 3 and 4 show the comparison of the results of BSTT ablation experiments on the two datasets. Specifically, we design three variants of the Bayesian spatial-temporal transformer, including:

- Bayesian Spatial Transformer (BST), which removes the Bayesian temporal transformer module to determine the impact of modeling temporal relations on model performance.
- Bayesian Temporal Transformer (BTT), which removes the Bayesian spatial transformer module to determine the impact of modeling spatial relations on model performance.
- Spatial-Temporal Transformer (STT), which removes the relational inference component to determine the impact of Bayesian relational inference on model performance.

Table 3: Ablation experiments on the ISRUC dataset.

method	ACC(%)	F1 scroe(%)	KAPPA(%)
BST	80.76	78.24	75.13
BTT	80.57	78.13	75.02
STT	80.35	78.05	74.71
BSTT	**81.96**	**80.30**	**76.78**

Table 4: Ablation experiments on the MASS dataset.

method	ACC(%)	F1 scroe(%)	KAPPA(%)
BST	88.76	83.64	83.28
BTT	88.62	83.64	83.02
STT	88.64	83.53	83.16
BSTT	**89.50**	**85.00**	**84.37**

A.6.3 COMPARISON OF SOME MAIN BASELINES WITH BSTT AT CLASS F1 SCORE

Table 5 and 6 show the comparison of class F1 score of BSTT and some main baseline models.

Table 5: Class F1 score for some main baselines and BSTT on ISRUC dataset.

Method	F1_Wake(%)	F1_N1(%)	F1_N2(%)	F1_N3(%)	F1_REM(%)
U-Time	84.63	52.43	79.18	86.32	75.47
GraphSleepNet	87.19	58.92	79.98	89.70	78.71
ST-Transformer	87.53	57.69	80.83	88.63	78.79
BSTT (Our)	**88.89**	**59.81**	**82.15**	**89.75**	**80.71**

Table 6: Class F1 score for some main baselines and BSTT on MASS dataset.

Method	F1_Wake(%)	F1_N1(%)	F1_N2(%)	F1_N3(%)	F1_REM(%)
U-Time	86.48	51.81	88.32	80.10	87.95
GraphSleepNet	89.92	64.18	**92.81**	82.10	90.89
ST-Transformer	**91.28**	**65.32**	92.09	83.14	90.81
BSTT (Our)	90.92	64.28	92.77	**85.14**	**91.88**

A.6.4 RELATION INTENSITY GRAPHS OF BAYESIAN RELATION INFERENCE COMPONENT

To further demonstrate the effectiveness of our proposed Bayesian spatial-temporal relation inference component, we display other spatial-temporal relationship intensity matrices generated by this

component. Evidently, brain spatial connectivity is very strong during Wake period and REM period, whereas brain spatial connectivity is gradually weakened during NREM periods and N3 period comes to the weakest. Besides, the temporal relationship intensity matrices also have interpretability. For example, when the sleep transition from N2 to N3 occurscomponent focus more on the relation between the second and third time slices and fourth time slices. Same as sleep spacialists, the proposed component is more concerned with the persistence of the sleep signal after the transition has occurred during this period.

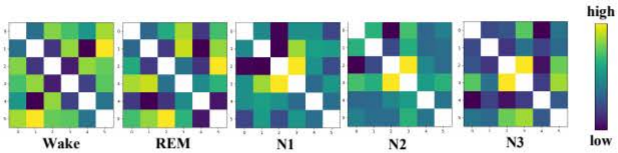

Figure 3: Spatial relation intensity graphs on ISRUC dataset (6 channels).

.

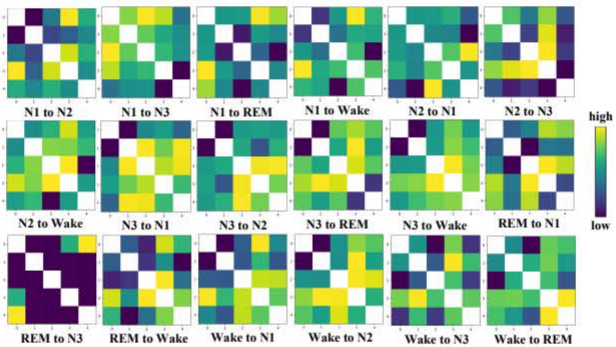

Figure 4: Temporal relation intensity graphs on ISRUC dataset.

.

A.6.5 Experiment of multi-head Bayesian relation inference component

We replace the self-attention component in traditional transformer with Bayesian relation inference component we proposed. In traditional transformer, multi-head self-attention is proposed to generate attention graph from different angles and improve the performance. Inspired by multi-head self-attention component, we propose multi-head Bayesian relation inference component to better infer spatial-temporal relation and generate relation graph. The calculation is defined as follows:

$$head_i = \boldsymbol{f}_{\mathrm{BRI}}(\boldsymbol{E}) \tag{8}$$

$$MultiHead(E) = \mathrm{Concat}(\,\mathrm{head}_1, \ldots, \mathrm{head}_n)\, W^O \tag{9}$$

Each head represents a single Bayesian relation inference component. We have tried some values of n, and finally chose $n = 3$ to train the model. Table 7 shows the result on the ISRUC dataset when we set n to different values.

A.7 Code and Dataset

The code of BSTT is available at: https://github.com/YuchenLiu1225/BSTT/tree/main/BSTT.

The list of open-source baseline model code used in the experiments is as follows:

Table 7: The influence of n on the ISRUC dataset.

n	ACC(%)	F1 scroe(%)	KAPPA(%)
1	80.66	78.19	75.16
3	**81.96**	**80.20**	**76.78**
5	81.76	80.05	76.48

- DeepSleepNet: https://github.com/akaraspt/deepsleepnet
- TinySleepNet: https://github.com/akaraspt/tinysleepnet
- GraphSleepNet: https://github.com/ziyujia/GraphSleepNet
- U-Time: https://github.com/perslev/U-Time
- MMCNN: https://github.com/ziyujia/ECML-PKDD_MMCNN

REFERENCES

Richard B Berry, Rita Brooks, Charlene E Gamaldo, Susan M Harding, C Marcus, Bradley V Vaughn, et al. The aasm manual for the scoring of sleep and associated events. *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine*, 176: 2012, 2012.

Stanislas Chambon, Mathieu N Galtier, Pierrick J Arnal, Gilles Wainrib, and Alexandre Gramfort. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(4): 758–769, 2018.

Hao Dong, Akara Supratak, Wei Pan, Chao Wu, Paul M Matthews, and Yike Guo. Mixed neural network approach for temporal sleep stage classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(2):324–333, 2017.

Hengguan Huang, Fuzhao Xue, Hao Wang, and Ye Wang. Deep graph random process for relational-thinking-based speech recognition. In *International Conference on Machine Learning*, pp. 4531–4541. PMLR, 2020.

Ziyu Jia, Youfang Lin, Jing Wang, Kaixin Yang, Tianhang Liu, and Xinwang Zhang. Mmcnn: A multi-branch multi-scale convolutional neural network for motor imagery classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 736–751. Springer, 2020a.

Ziyu Jia, Youfang Lin, Jing Wang, Ronghao Zhou, Xiaojun Ning, Yuanlai He, and Yaoshuai Zhao. Graphsleepnet: Adaptive spatial-temporal graph convolutional networks for sleep stage classification. In *IJCAI*, pp. 1324–1330, 2020b.

Mathias Perslev, Michael Jensen, Sune Darkner, Poul Jørgen Jennum, and Christian Igel. U-time: A fully convolutional network for time series segmentation applied to sleep staging. *Advances in Neural Information Processing Systems*, 32, 2019.

Huy Phan, Kaare B Mikkelsen, Oliver Chen, Philipp Koch, Alfred Mertins, and Maarten De Vos. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Transactions on Biomedical Engineering*, 2022.

Akara Supratak and Yike Guo. Tinysleepnet: An efficient deep learning model for sleep stage scoring based on raw single-channel eeg. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 641–644. IEEE, 2020.

Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, 2017.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *arXiv*, 2017.