【WP2.L303.3】

MULTI-VIEW VARIATIONAL RECURRENT NEURAL NETWORK FOR HUMAN EMOTION RECOGNITION USING MULTI-MODAL BIOLOGICAL SIGNALS

Yuya Moroto†, Keisuke Maeda ‡, Takahiro Ogawa ‡, Miki Haseyama‡

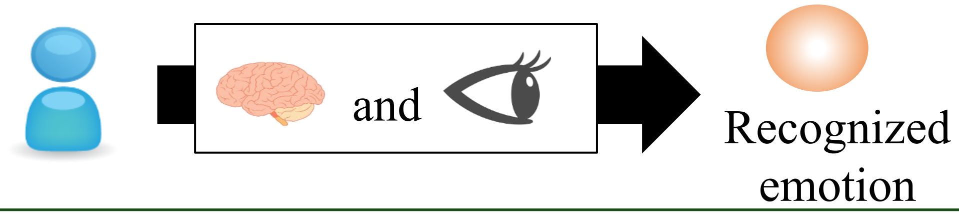
†Graduate School of Information Science and Technology, Hokkaido University, Japan
‡Faculty of Information Science and Technology, Hokkaido University, Japan

Introduction - background -

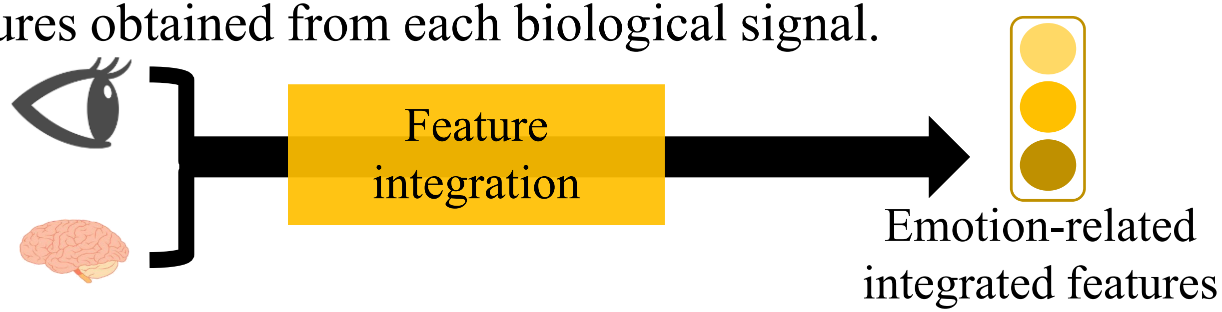
Human emotion recognition for visual stimuli

Human emotions are well-known to play an important role not only for human communications, but also for human-computer communications.

To make computers recognize human emotions while viewing images, multi-modal biological signals, which are eye gaze and brain activity, have been focused [5-8, 25].



Previous works realize the multi-modal human emotion recognition by integrating features obtained from each biological signal.



By using several biological signals, the accuracy of human emotion recognition has been improved.

[5] Wei Liu, *et al.*, in *Proc. Int'l Conf. Neural Information Processing*, 2016.
[6] Jie Qiu, *et al.*, in *Proc. Int'l Conf. Neural Information Processing*, 2018.
[7] Hao Tang, *et al.*, in *Proc. Int'l Conf. Neural Information Processing*, 2017.
[8] Yuya Moroto, *et al.*, *IEEE Access*, 2020.
[25] Yuya Moroto, *et al.*, in *Proc. IEEE Global Conf. Life Sciences and Technologies*, 2021.

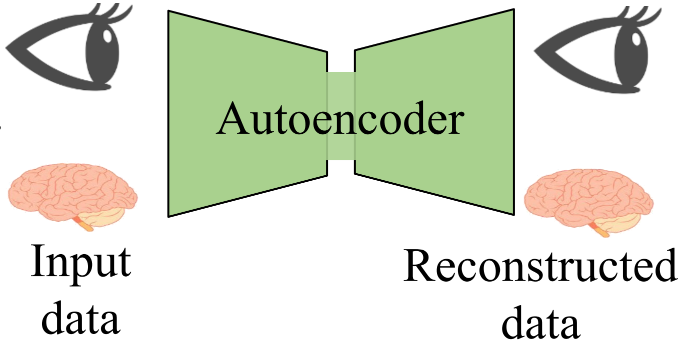
Related Works

Autoencoder-based method [5]

For treating multiple biological signals, the simple machine learning method is used.

Multi-modal human emotion recognition is realized.

The temporal change, that is an important factor, cannot be considered.

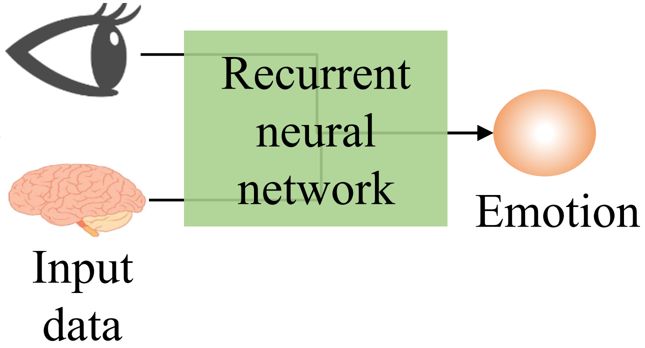


Recurrence-based method [7]

For treating biological signals as sequential data, the recurrent neural network is used.

The temporal changes can be considered.

Deterministic machine learning can be affected by noises in biological signals.



From the above related works, the following characteristics of biological signals should be considered.

i. **Information complementation by using several biological signals.**
ii. **Temporal changes in biological signals**
iii. **Effects of noises**

In general machine learning models, it is difficult to simultaneously consider the above characteristics.

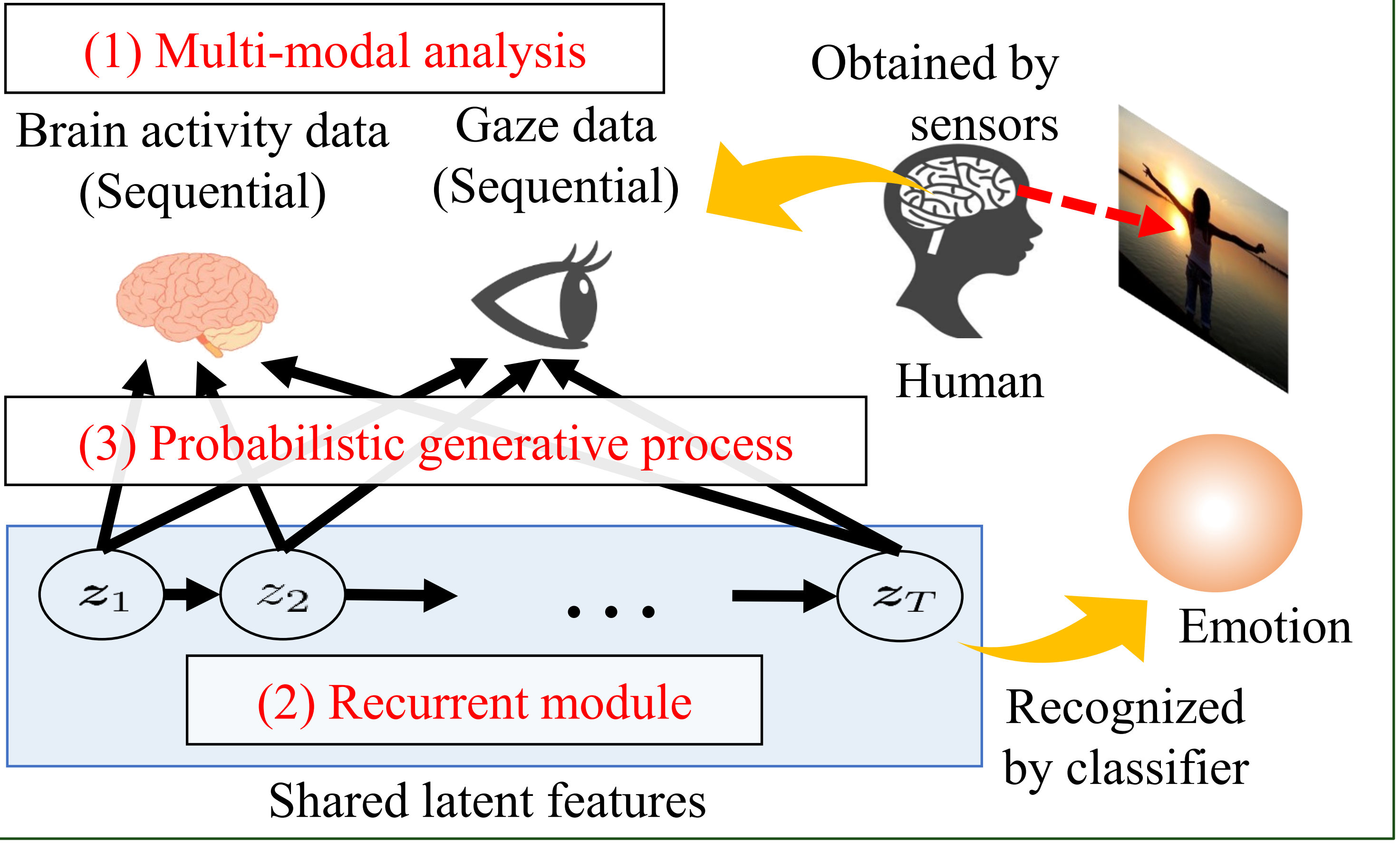
[5] Wei Liu, *et al.*, in *Proc. Int'l Conf. Neural Information Processing*, 2016.
[7] Hao Tang, *et al.*, in *Proc. Int'l Conf. Neural Information Processing*, 2017.

Approach

Conceptual diagram in this presentation

We derive a machine learning model that can simultaneously consider the following points.

i. **Multi-modal analysis**
ii. **Recurrent module for sequential data**
iii. **Probabilistic generative process**



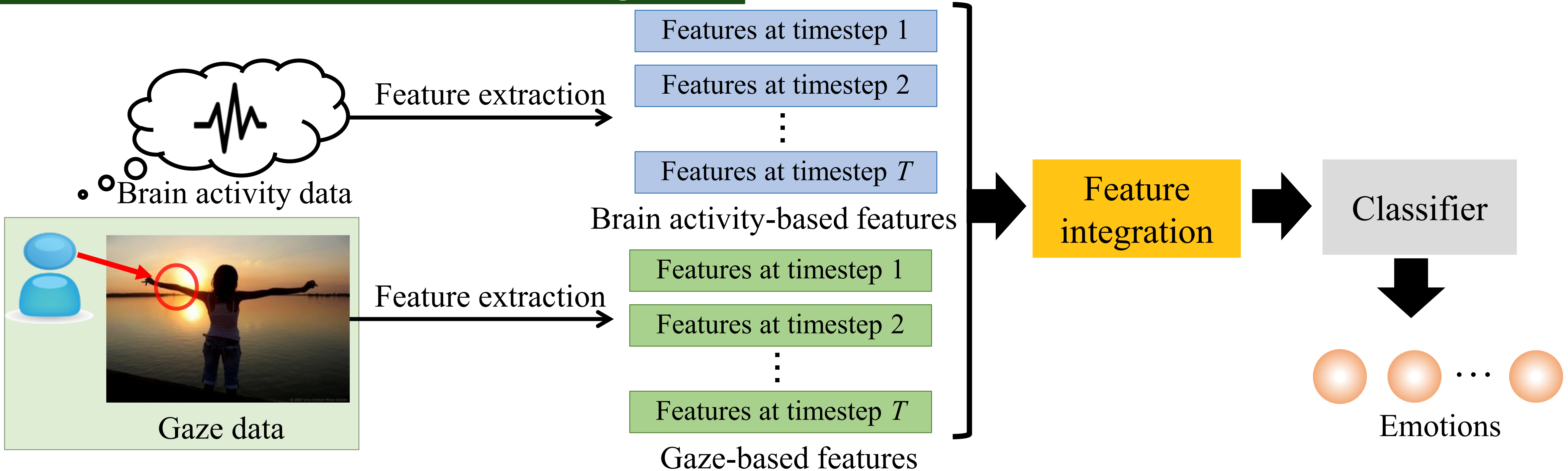
Novelty

The above mechanisms are simultaneously realized in a single machine learning model.

By deriving the novel machine learning model, we realize the improvement of recognition accuracy.

Position of the proposed method

Flow of multi-modal human emotion recognition



We newly propose the **feature integration** method with the following mechanisms.

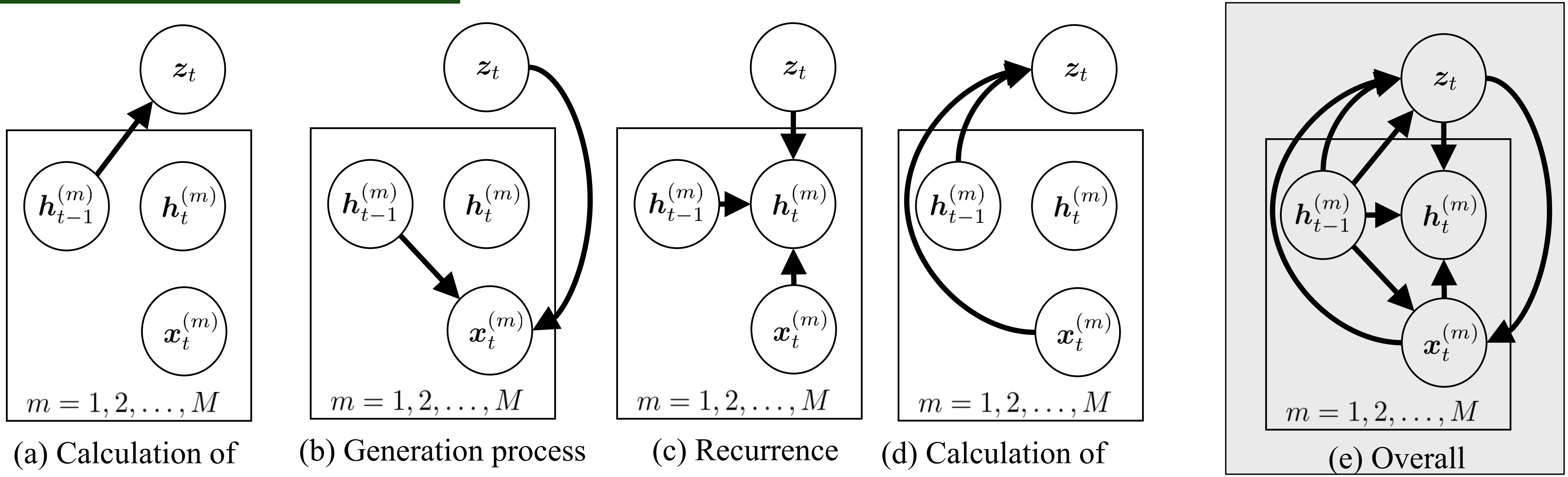
i. Multi-modal analysis
ii. Recurrent module for sequential data
iii. Probabilistic generative process

We newly derive the feature integration method suitable for multi-modal human emotion recognition.

Proposed feature integration - overview -

Proposed feature integration Multi-view Variational Recurrent Neural Network (MvVRNN)

Graphical models of MvVRNN

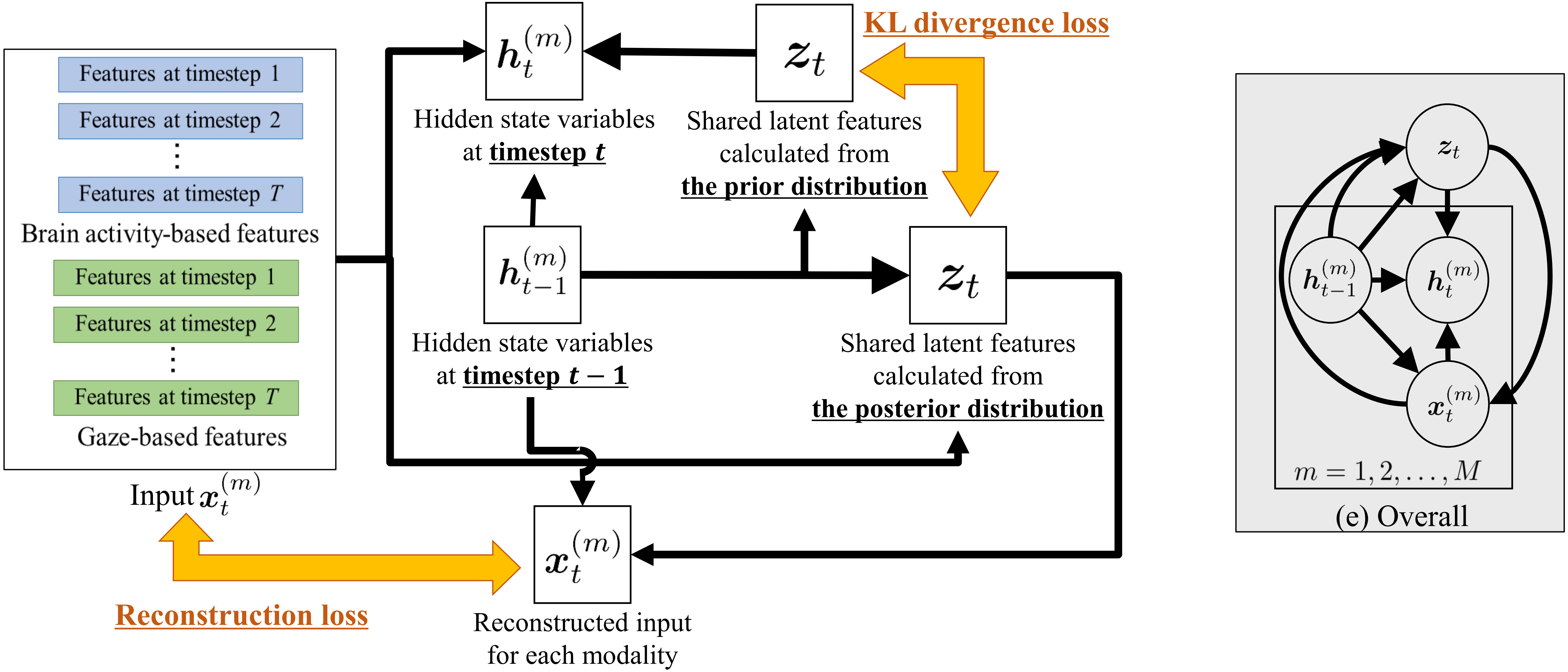


(a) Calculation of the prior distribution (b) Generation process (c) Recurrence (d) Calculation of the posterior distribution (e) Overall

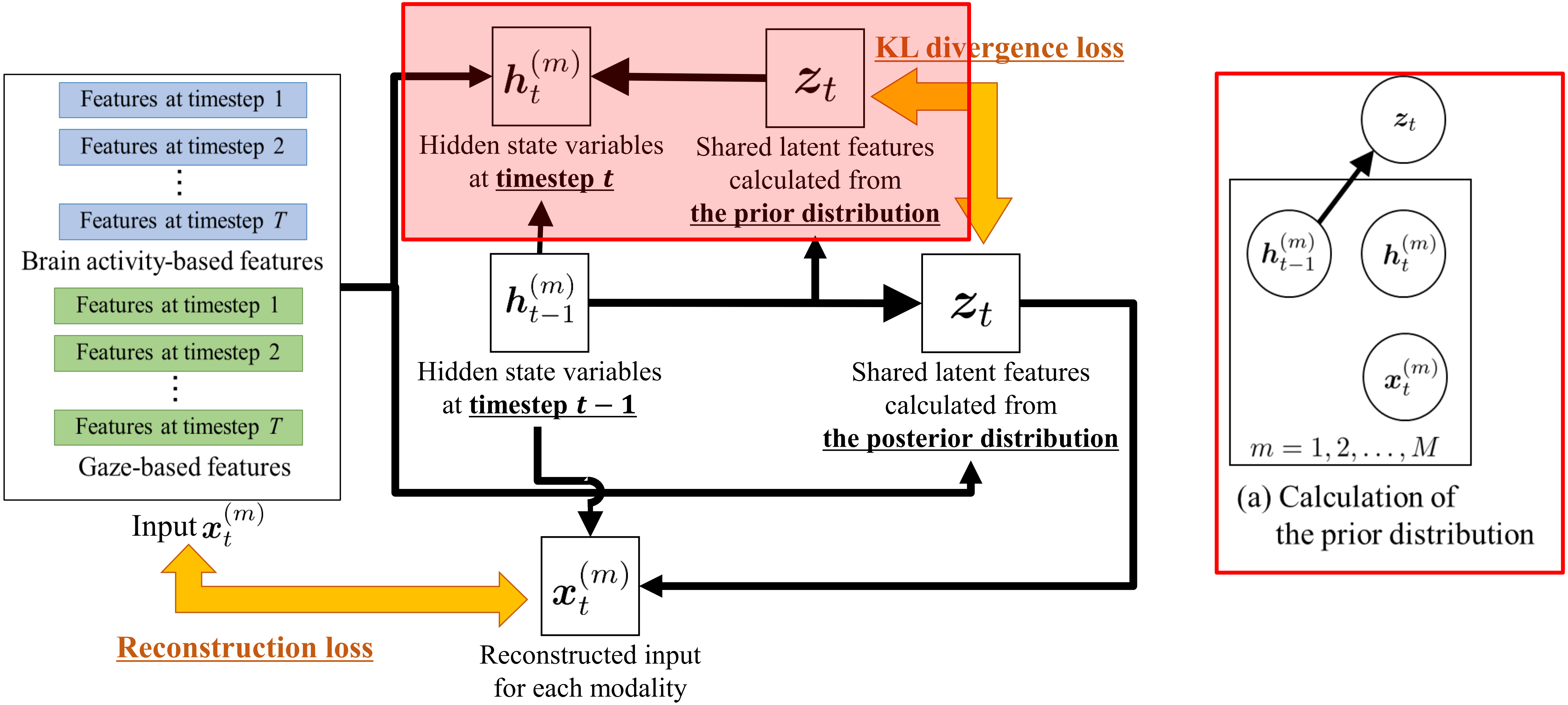
h : hidden state variables M : number of modalities t : timestep x : input data z : shared latent features

By deriving the MvVRNN, we can treat the characteristics of biological signals.

MvVRNN - model architecture -

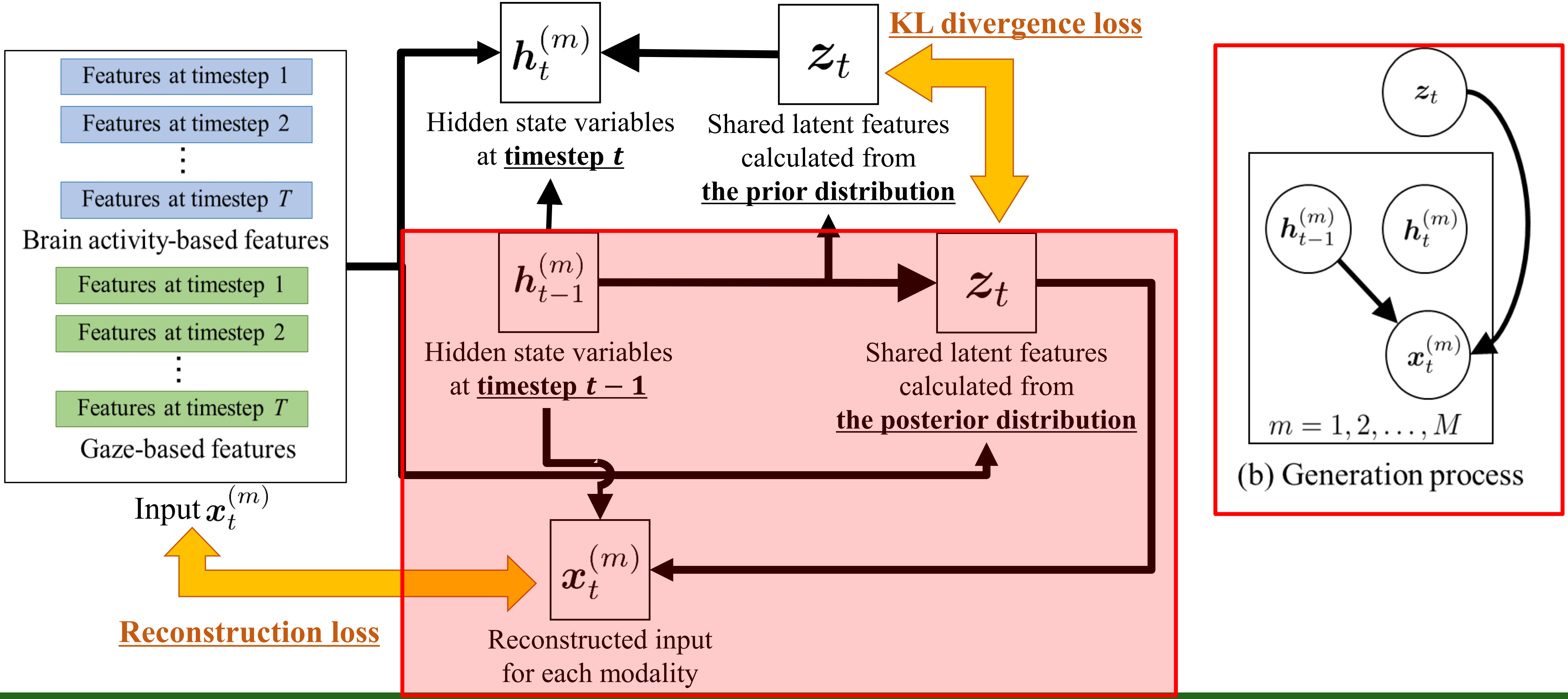


MvVRNN - model architecture -

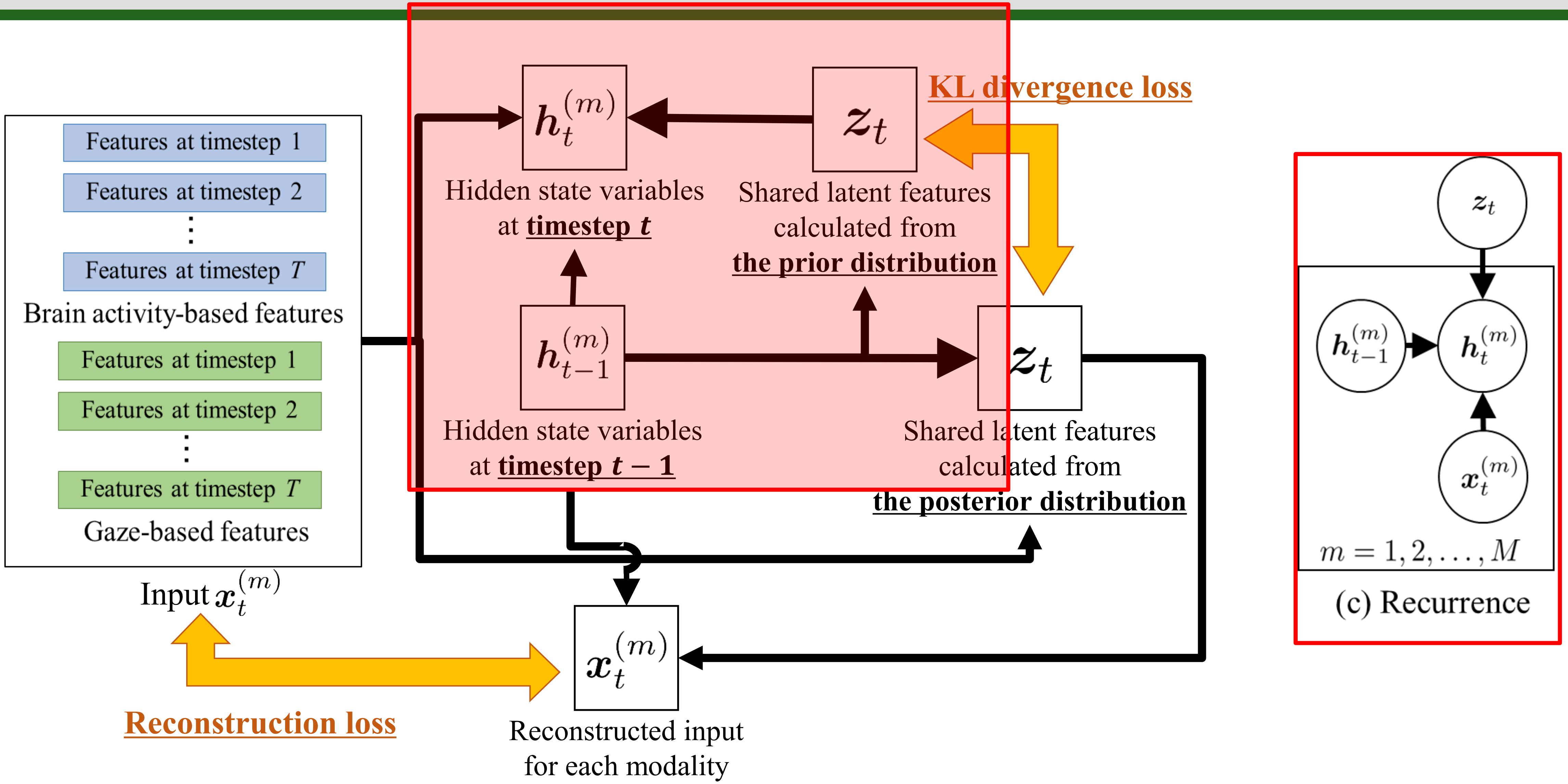


(a) Calculation of the prior distribution

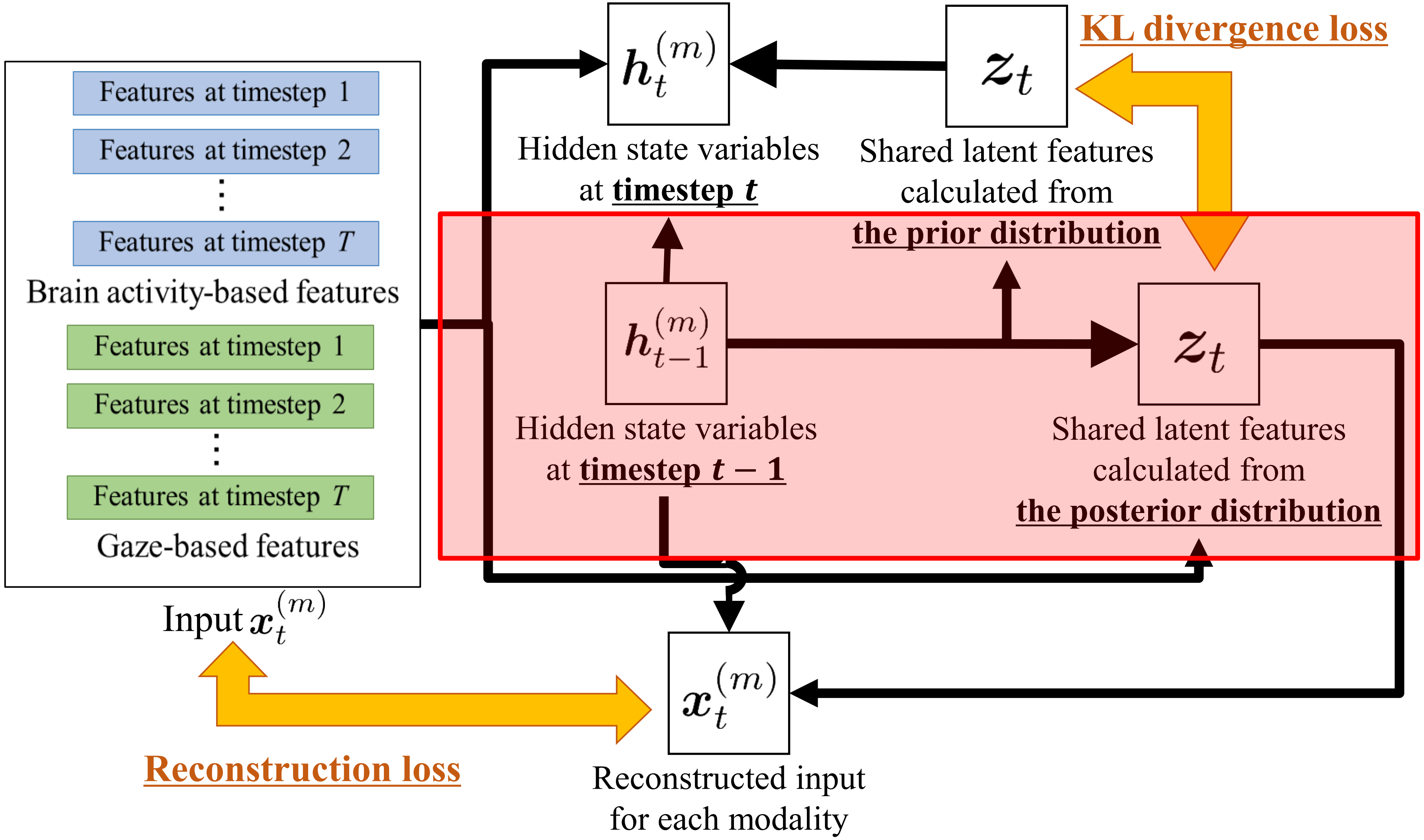
MvVRNN - model architecture -

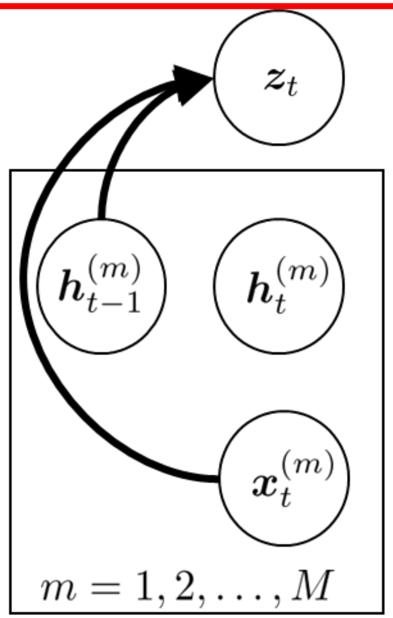


MvVRNN - model architecture -



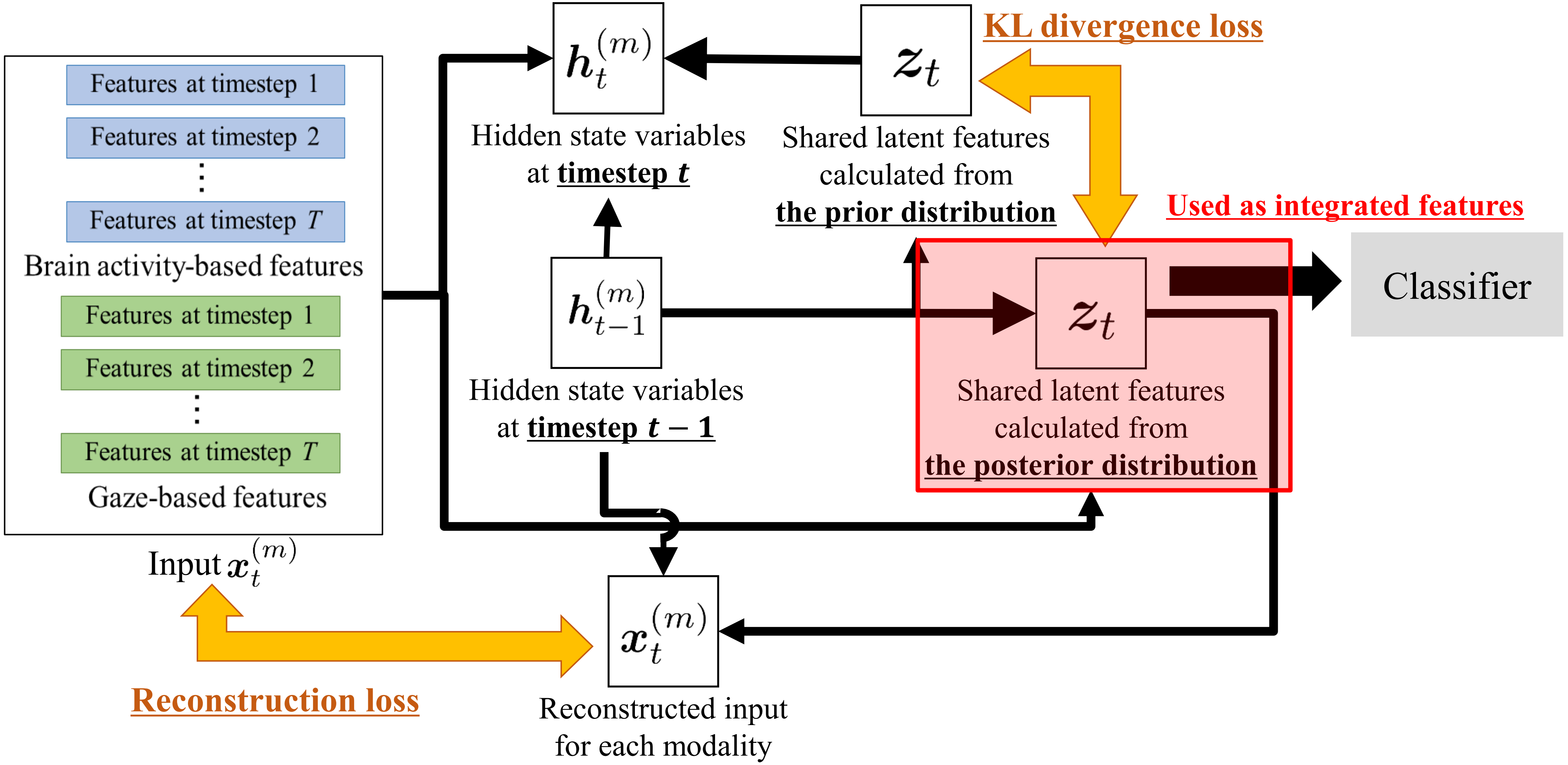
MvVRNN - model architecture -





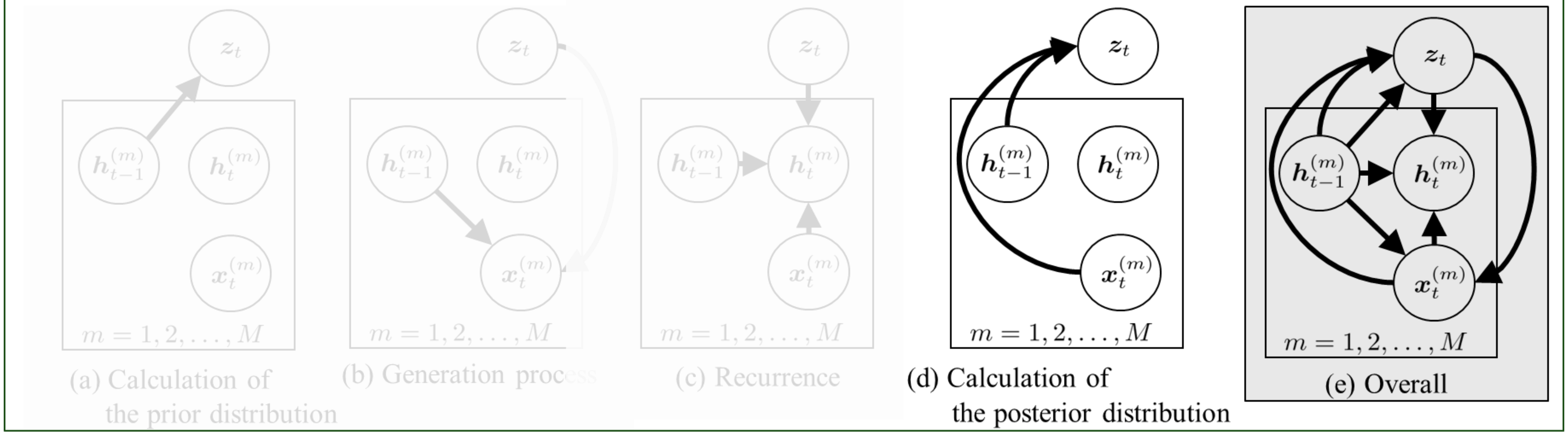
(d) Calculation of the posterior distribution

MvVRNN - model architecture -



Proposed feature integration - multi-modal analysis -

Graphical models of MvVRNN



To capture the relationship between multiple biological signals, we construct (d) Calculation of the posterior distribution.
Shared latent variables are calculated from the input data as follows:

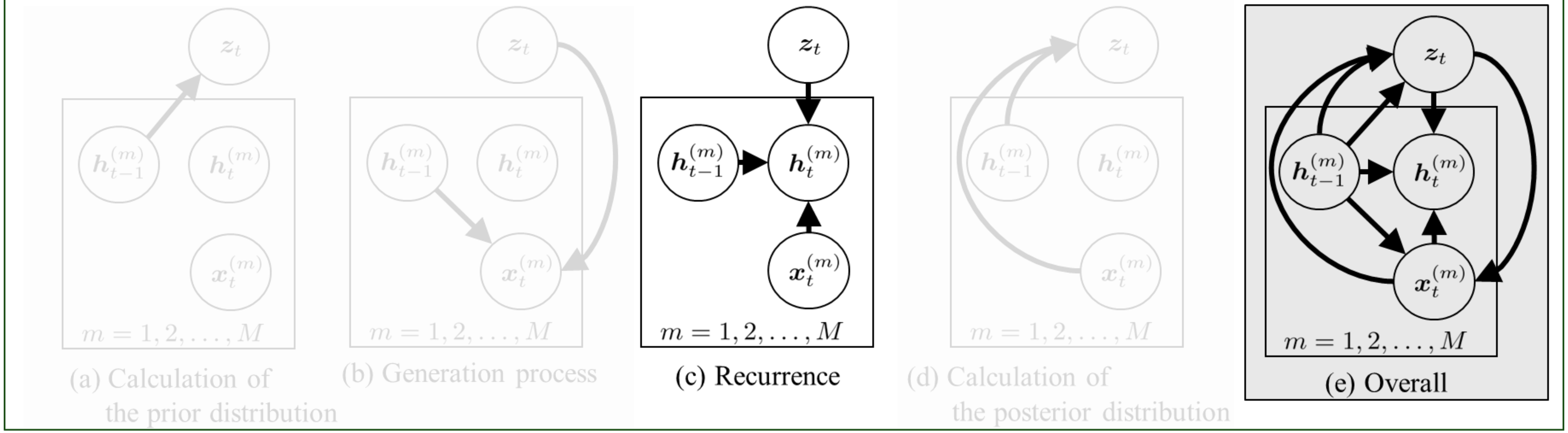
$$\boldsymbol{z}_t | \boldsymbol{x}_t^{(1)}, \boldsymbol{x}_t^{(2)}, \ldots, \boldsymbol{x}_t^{(M)} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{post},t}, \mathrm{diag}(\boldsymbol{\sigma}^2_{\mathrm{post},t}))$$

$\boldsymbol{\mu}_{\mathrm{post},t}, \boldsymbol{\sigma}_{\mathrm{post},t}$: mean and standard deviation of $\boldsymbol{z}_t$ calculated from input data

By calculating the shared latent features from several input data, we realize the feature integration.

Proposed feature integration - recurrent module -

Graphical models of MvVRNN



To capture temporal changes in the biological signals, we construct (c) Recurrence.
Shared latent variables are also used to consider the relationships of shared latent variables across timesteps as follows:

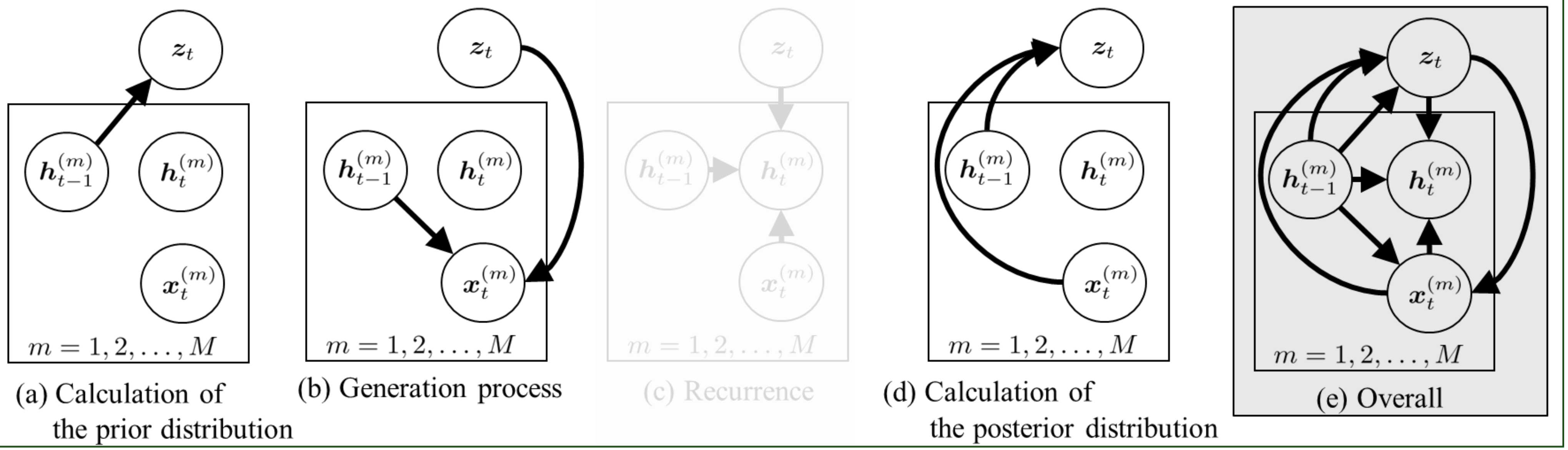
$$\boldsymbol{h}_t^{(m)} = g_h(g_{x^{(m)}}(\boldsymbol{x}_t^{(m)}), g_z(\boldsymbol{z}_t), \boldsymbol{h}_{t-1}^{(m)})$$

g_h, g_x, g_z : function for feature transformation

By introducing the recurrence module, we realize the time-series analysis.

Proposed feature integration - probabilistic generative process -

Graphical models of MvVRNN



To consider the effects of noises in the biological signals, a generative model via the Bayesian inference framework is adopted. Important probability distributions are calculated in the following models:

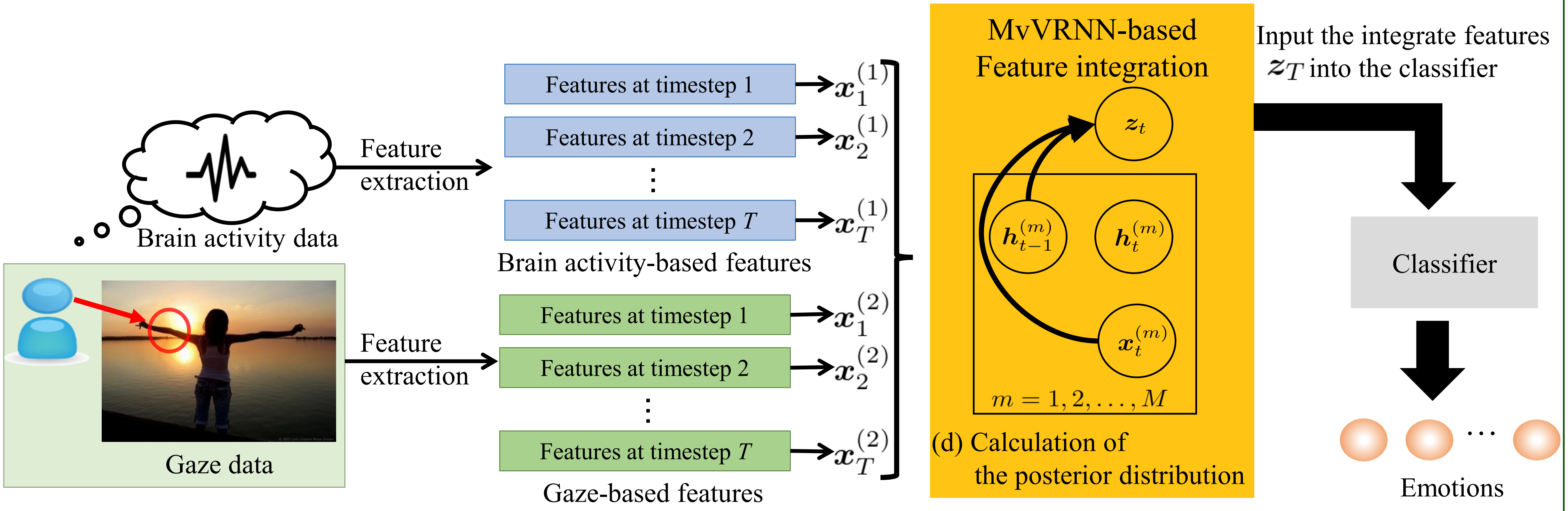
Prior distribution $p(z_t)$: (a) Calculation of the prior distribution.
Likelihood $p(x_t^{(m)}|z_t)$: (b) Generation process
Posterior distribution $q(z_t|x_t^{(1)}, x_t^{(2)}, \ldots, x_t^{(M)})$: (d) Calculation of the posterior distribution

The probabilistic generative model can reduce the effects of noises.

Proposed feature integration - test phase -

Flow of multi-modal human emotion recognition

Although the MvVRNN consists of several modules, we use the shared latent features as integrated features.



By using MvVRNN, it is expected to calculate integrated features with high expressive power for emotions.

Experiment - settings -

Data acquisition

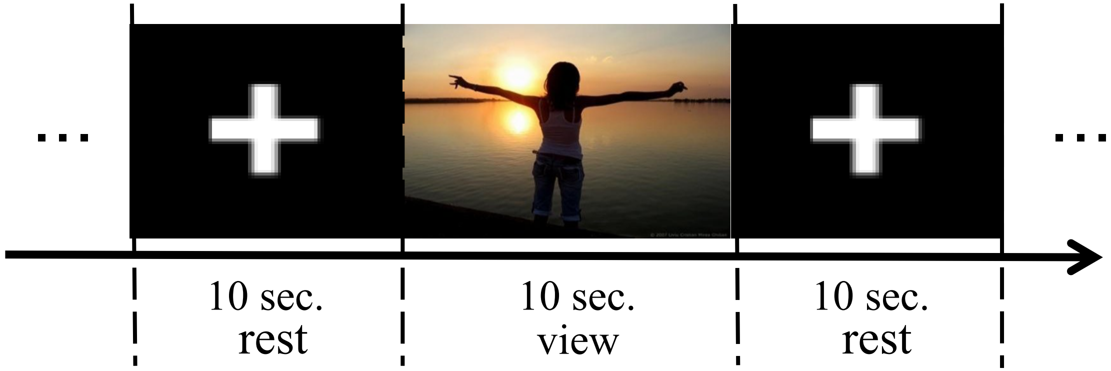
Experimental design : Block design (10 seconds for viewing images,
10 seconds for resting)

Visual Stimuli : 80 images obtained from Art photo dataset [12]
(Training for 80%, Test for 20%)

Number of Participants : 10

Instrument : 1. Tobii eye tracker 4C for eye tracking data
2. LIGHTNIRS for functional near-infrared spectroscopy (fNIRS) data

Ground Truth : Emotions (Positive / Negative) recalled for each image were obtained through a questionnaire survey.



Flow of block design

Numbers of images for each emotion recalled by participants.

ID	1	2	3	4	5	6	7	8	9	10
Negative	45	43	42	35	44	50	38	41	34	45
Positive	35	37	38	45	36	30	42	39	46	35

[12] Shaojing Fan, *et al.*, in *Proc. IEEE/CVF CVPR*, 2018.

Experiment - validation procedure -

Purpose

Validation of MvVRNN for multi-modal human emotion recognition

Comparative methods

We adopted following 2 comparative methods for ablation study.

- VRNN using only gaze (VRNN-gaze)
- VRNN using only fNIRS (VRNN-brain)

Moreover, we adopted 5 other integration methods.

- Bi-modal Deep AutoEncoder (BDAE) [5]
- Deep Canonical Correlation Analysis (Deep CCA) [6]
- Bi-modal Long Short-Term Memory (BLSTM) [7]
- Time-considered CCA (TC-CCA) [8]
- Time-considered multi-modal variational autoencoder (TC-MVAE) [25]

	Multi-modality	Recurrence	Variational
VRNN-gaze		✓	✓
VRNN-brain		✓	✓
TC-MVAE [25]	✓		
TC-CCA [8] Deep CCA [6] BDAE [5]	✓		
BLSTM [7]	✓	✓	
MvVRNN	✓	✓	✓

Other settings

Classifier : Support Vector Machine (SVM) [26]

Evaluation metrics : Recall, Precision, F1-score, Accuracy

[5] Wei Liu, *et al.*, in *Proc. Int'l Conf. Neural Information Processing*, 2016.
[6] Jie Qiu, *et al.*, in *Proc. Int'l Conf. Neural Information Processing*, 2018.
[7] Hao Tang, *et al.*, in *Proc. Int'l Conf. Neural Information Processing*, 2017.
[8] Yuya Moroto, *et al.*, *IEEE Access*, 2020.
[25] Yuya Moroto, *et al.*, in *Proc. IEEE Global Conf. Life Sciences and Technologies*, 2021.
[26] Vladimir Vapnik, *Automation and Remote Control*, 1963.

Experiment - experimental results-

	Recall	Precision	F1-score	Accuracy
VRNN-gaze	0.29	0.28	0.25	0.49
VRNN-brain	0.58	0.45	0.49	0.54
TC-MVAE [15]	0.49	0.55	0.52	0.57
TC-CCA [14]	0.63	**0.85**	0.67	0.74
Deep CCA [18]	0.57	0.54	0.53	0.58
BDAE [11]	0.29	0.64	0.55	0.57
BLSTM [21]	0.37	0.31	0.44	0.44
MvVRNN	**0.77**	0.67	**0.70**	**0.75**

MvVRNN vs VRNN-gaze, VRNN-brain

Confirming the effectiveness of using multiple biological signals for human emotion recognition.

MvVRNN vs BLSTM

Confirming the effectiveness of feature integration based on the probabilistic generative model.

MvVRNN vs TC-MVAE

Confirming the effectiveness of the introduction of a reccurent module that can take into account temporal changes.

MvVRNN vs TC-CCA, Deep CCA, BDAE

Confirming the effectiveness of MvVRNN compared to other feature integration methods.

We confirm the effectiveness of MvVRNN for multi-modal human emotion recognition.

Conclusion

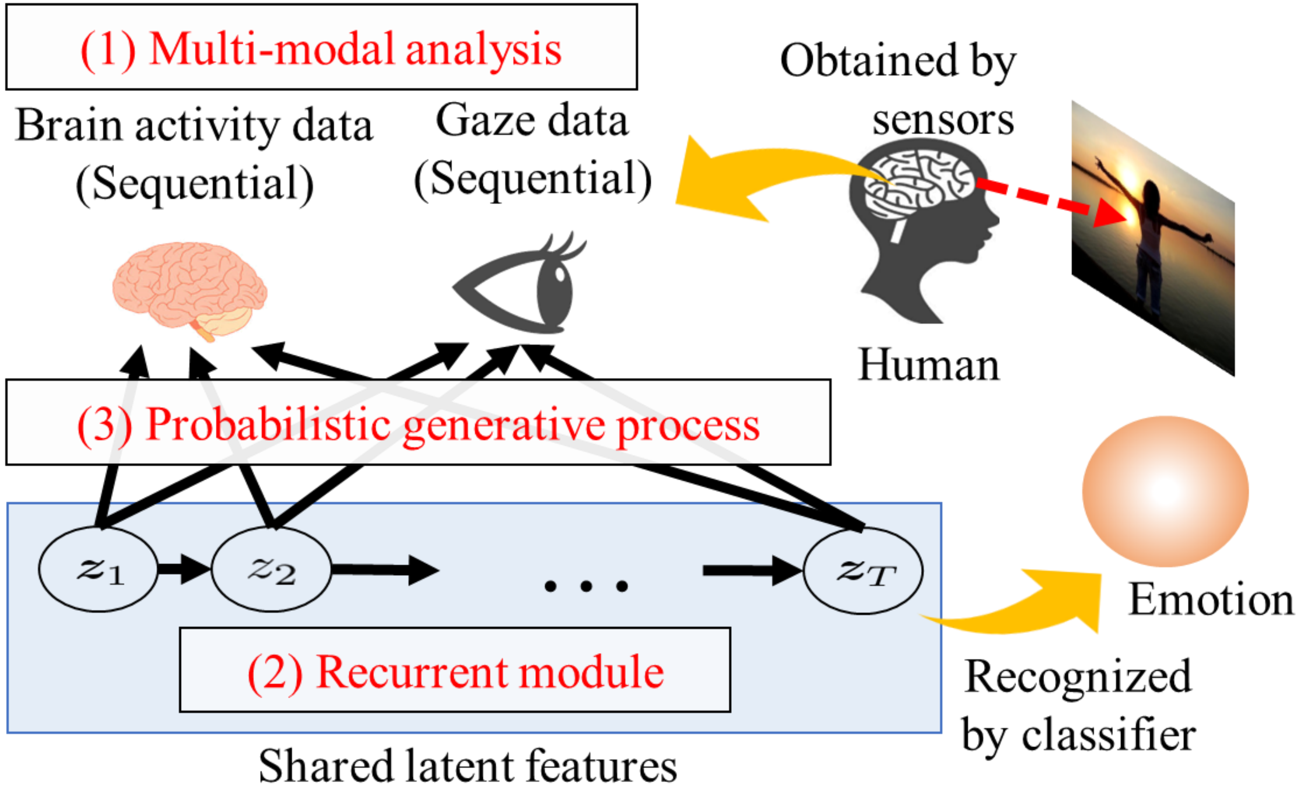
Proposed feature integration

Multi-view Variational Recurrent Neural Network (MvVRNN)

Conceptual diagram in this presentation

We derive a machine learning model that can simultaneously consider the following points.

i. **Multi-modal analysis**
ii. **Recurrent module for sequential data**
iii. **Probabilistic generative process**



Novelty

The above mechanisms are simultaneously realized in a single machine learning model.

Experimental results show the effectiveness of MvVRNN for multi-modal human emotion recognition.