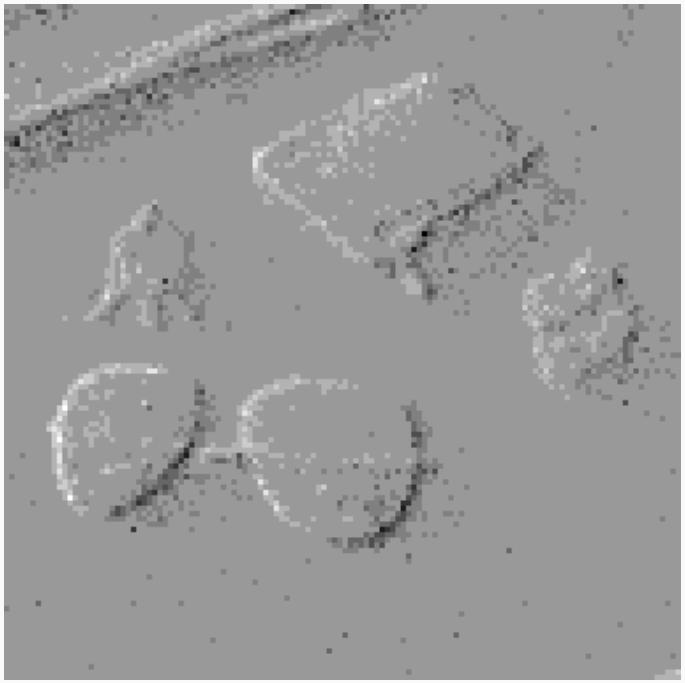
Gesture Recognition with a Neuromorphic Vision Sensor and Deep Learning

Marten Lienen

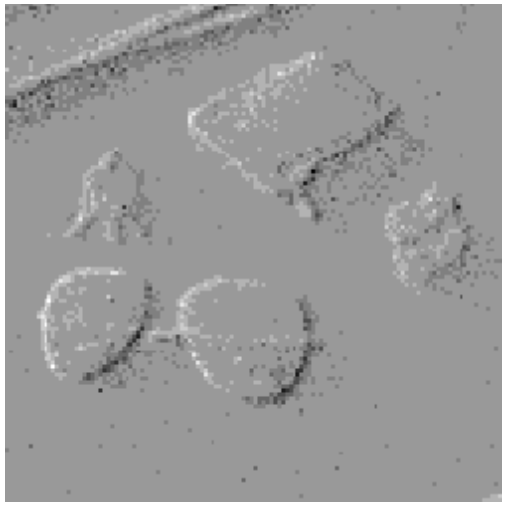
Image vs. Event Histogram

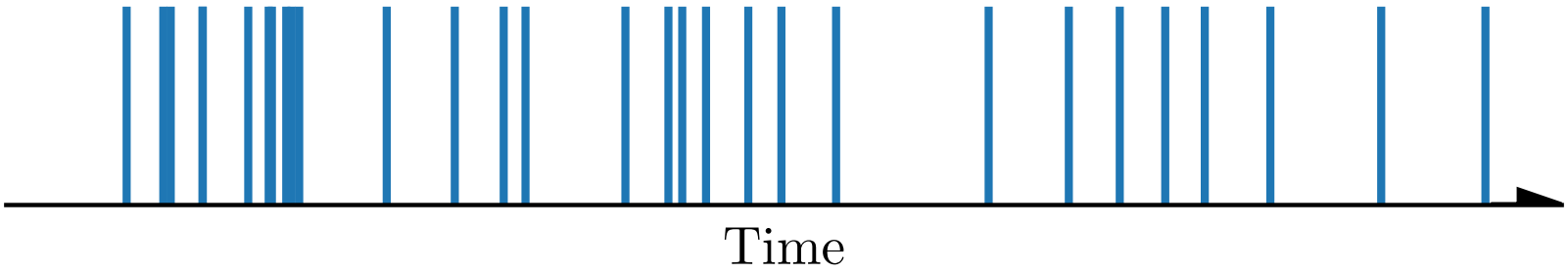


The Event Stream

Event

Timestamp t
Coordinates x, y
Polarity p





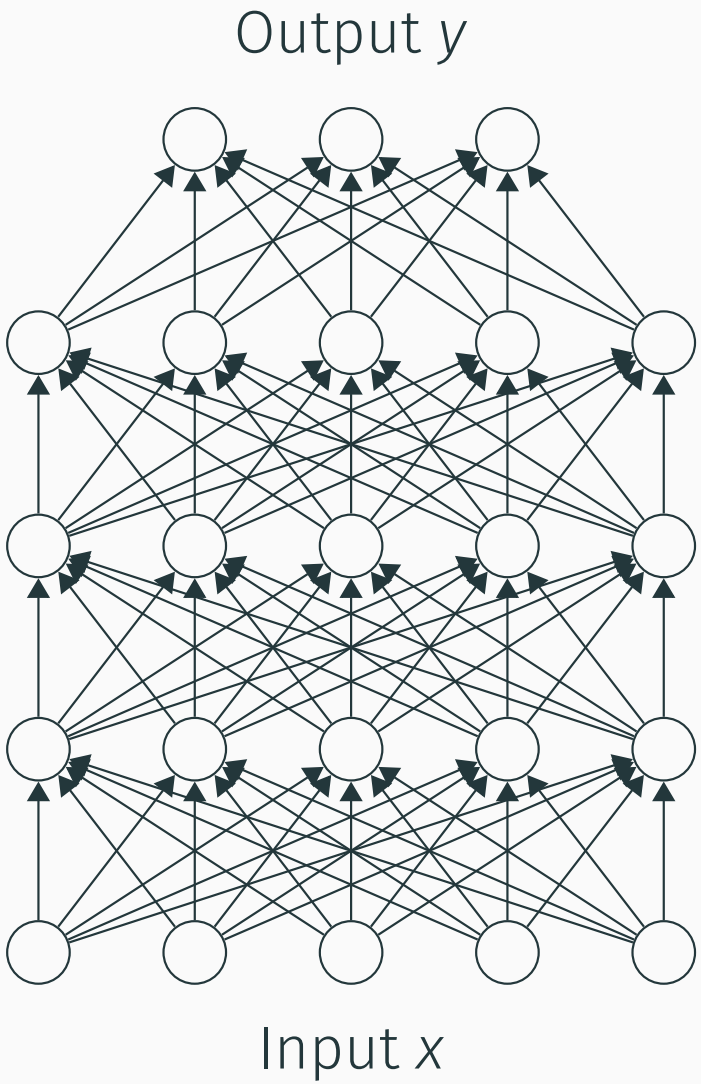
Dataset

- 16 gestures
- 2 subjects
- 40 segmented recordings
- 40 minutes and 640 examples in total

Preprocessing

- Translation invariance in time $\Delta t_i \leftarrow t_i - t_{i-1}$
- Translation invariance in space
  - Means with exponentially decaying weights as point of reference
  - Halflife of 50 ms and 1 s
- Feature Normalization

Neural Networks



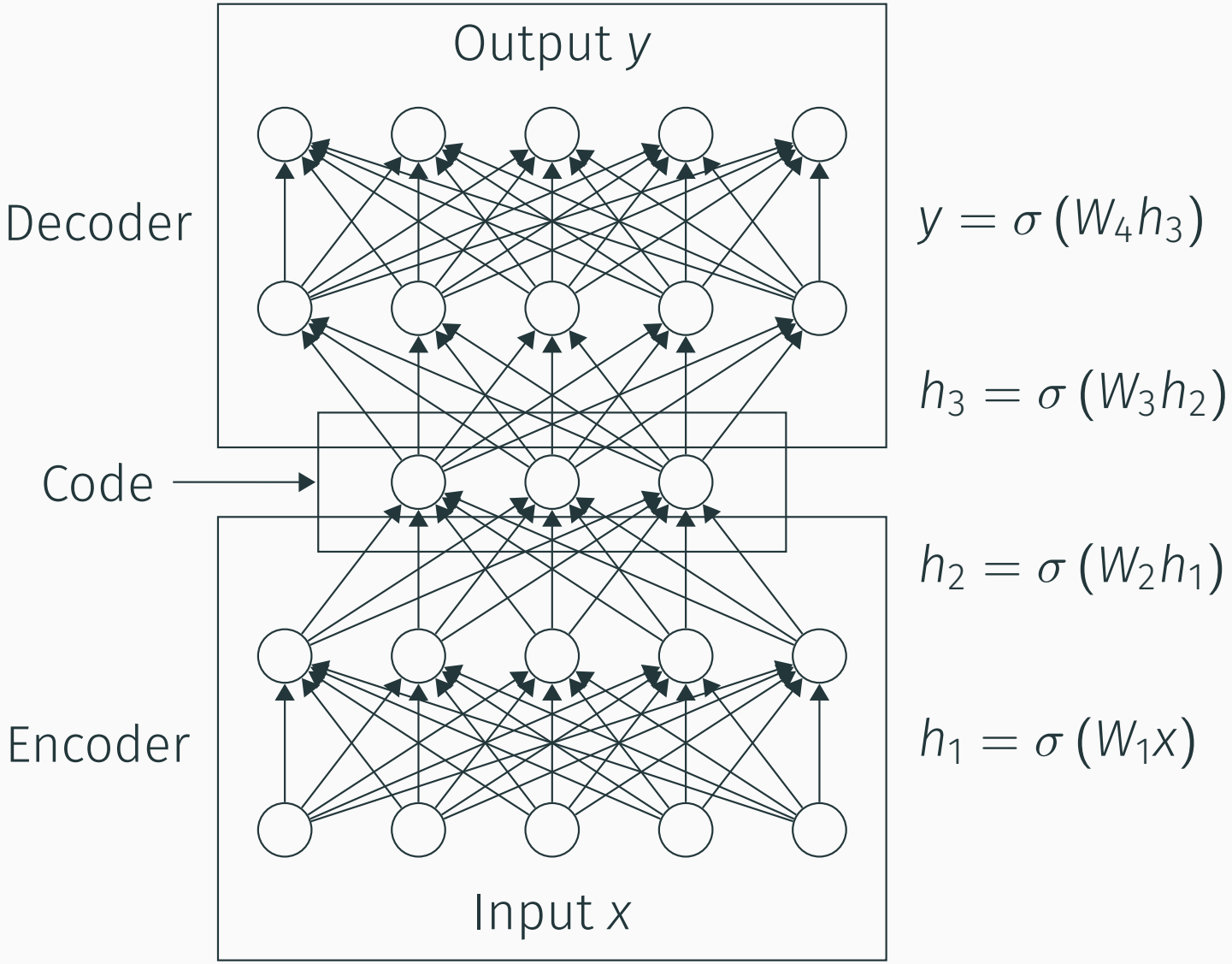
$$y = \sigma(W_4 h_3)$$

$$h_3 = \sigma(W_3 h_2)$$

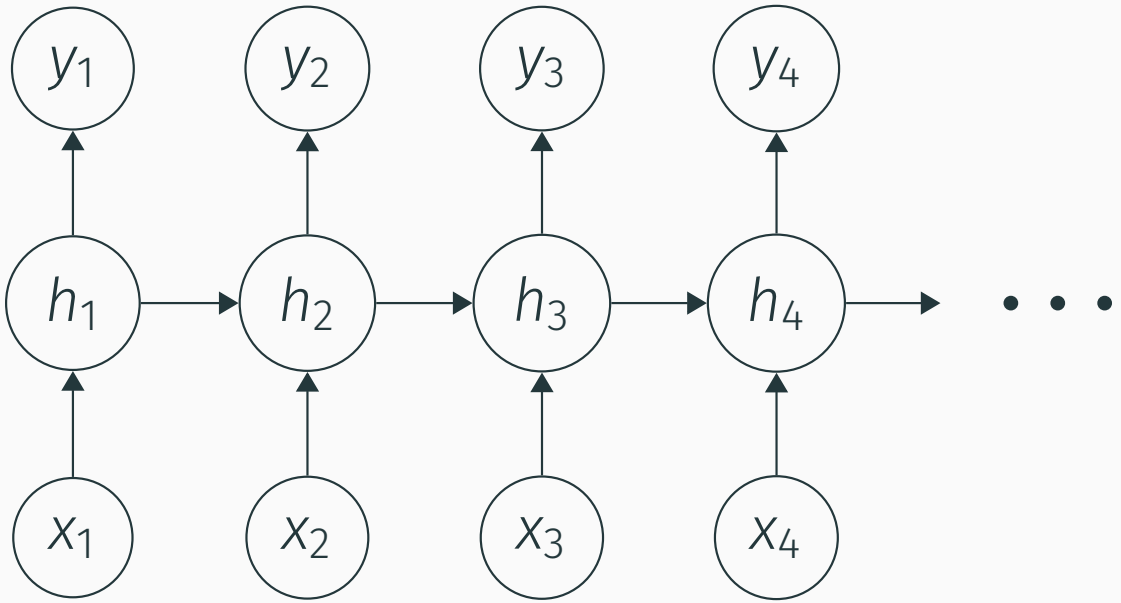
$$h_2 = \sigma(W_2 h_1)$$

$$h_1 = \sigma(W_1 x)$$

Autoencoders

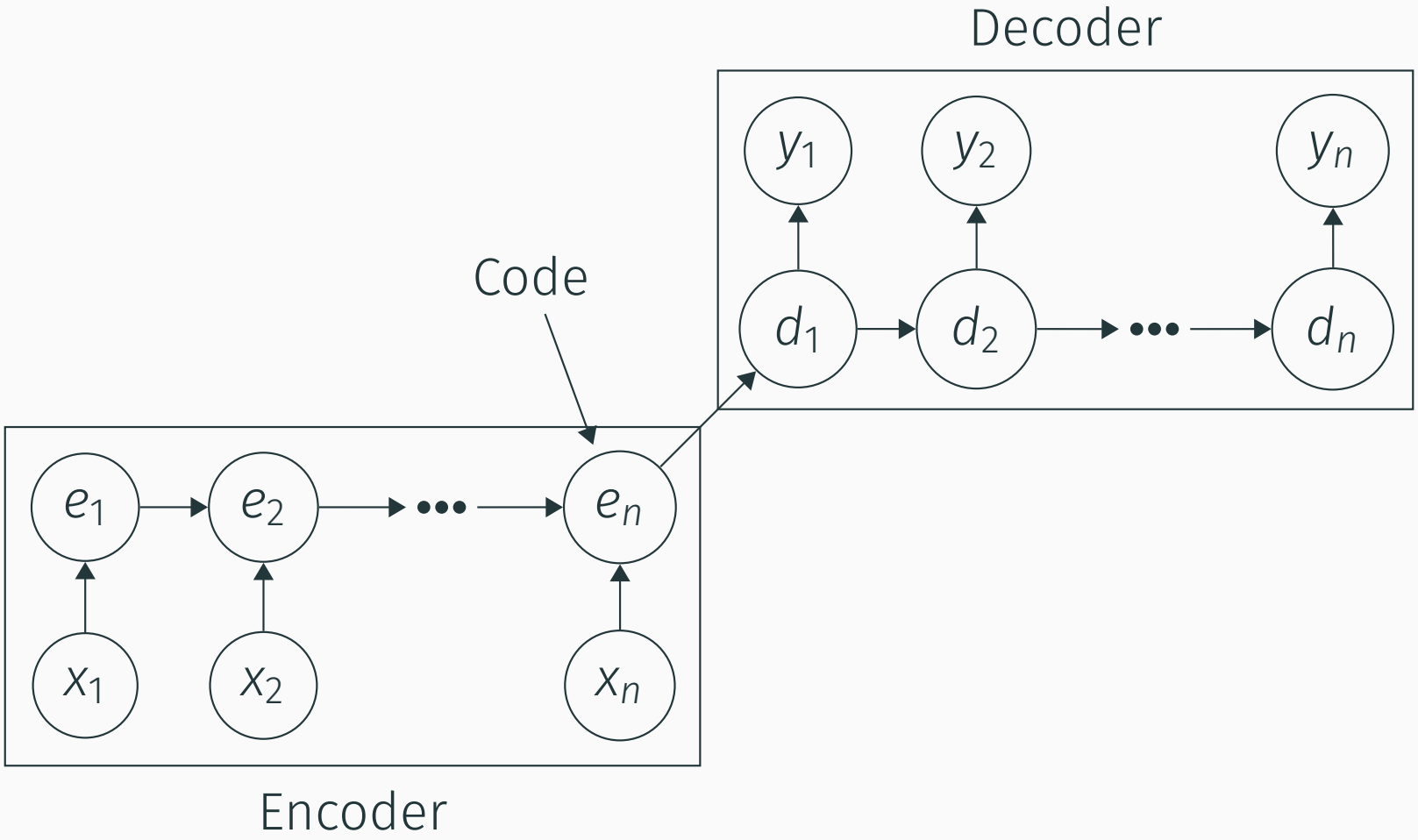


Recurrent Neural Networks

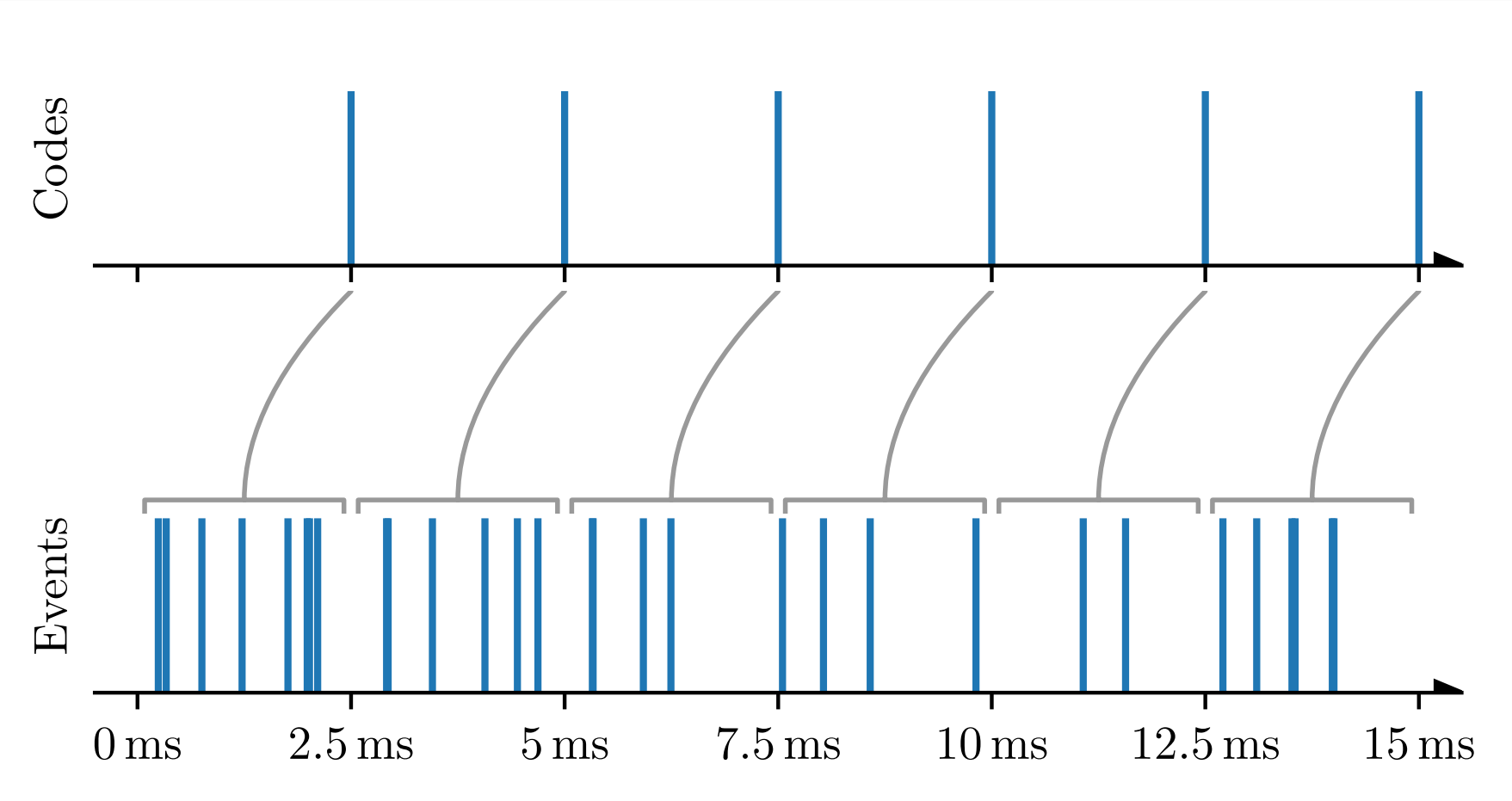


$$h_t = \sigma\left(Wx_t + Uh_{t-1}\right)$$

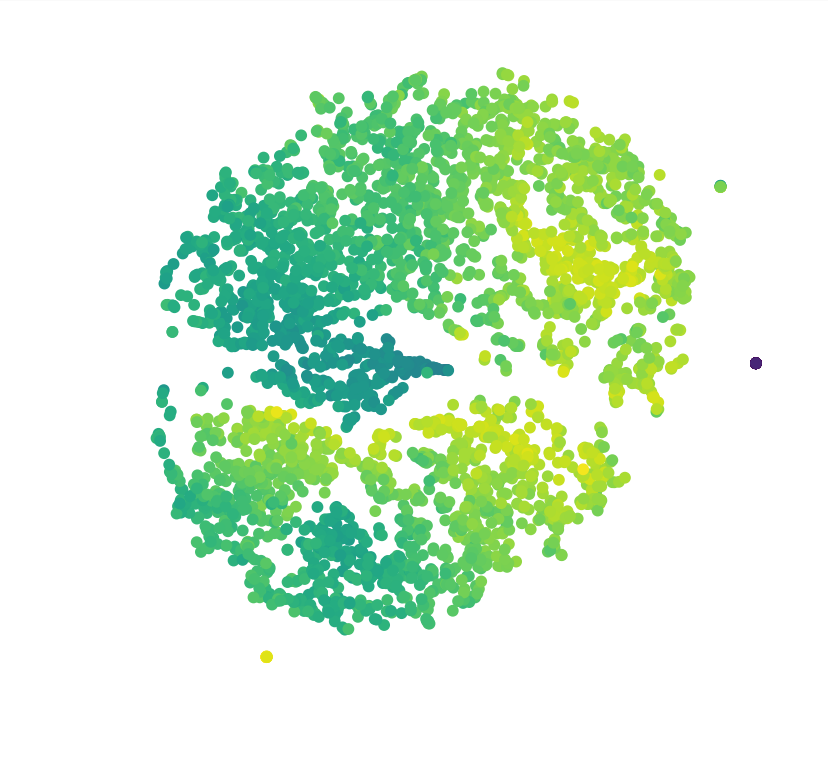
Recurrent Autoencoders



Event Sequence Compression



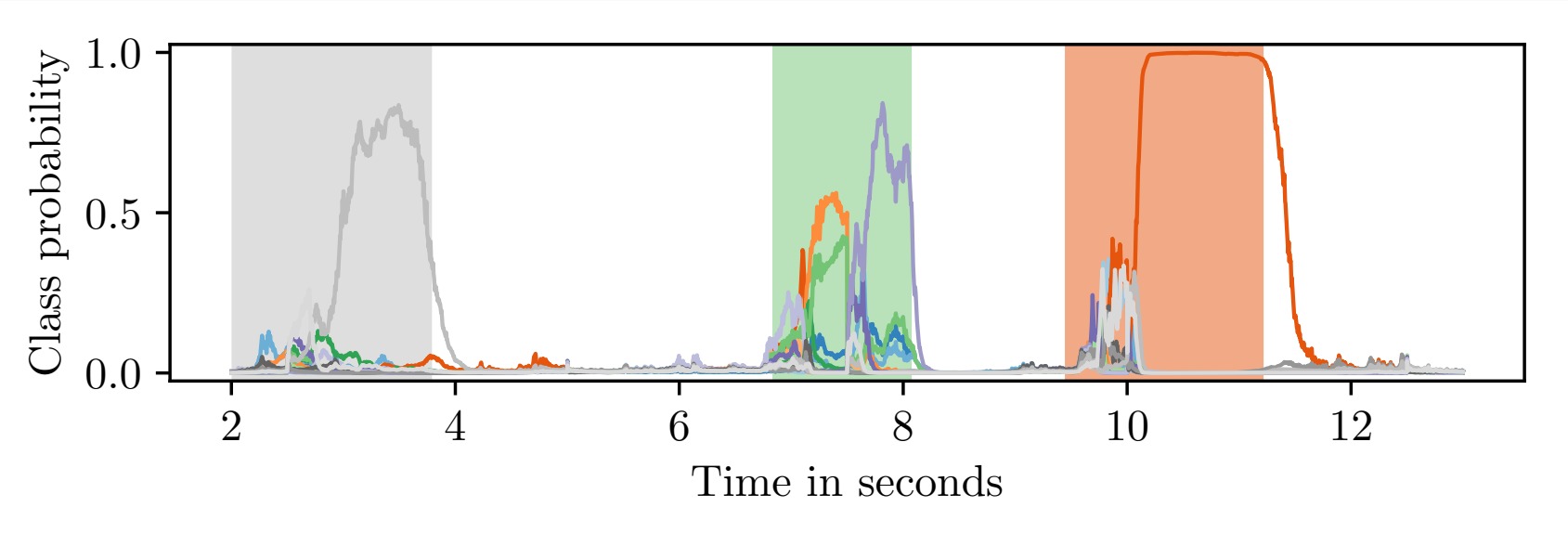
Learned Representations



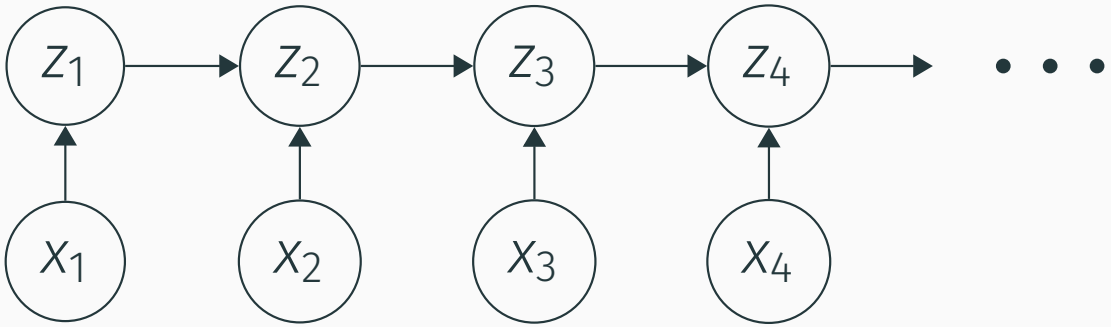
(a) Push Hand Up

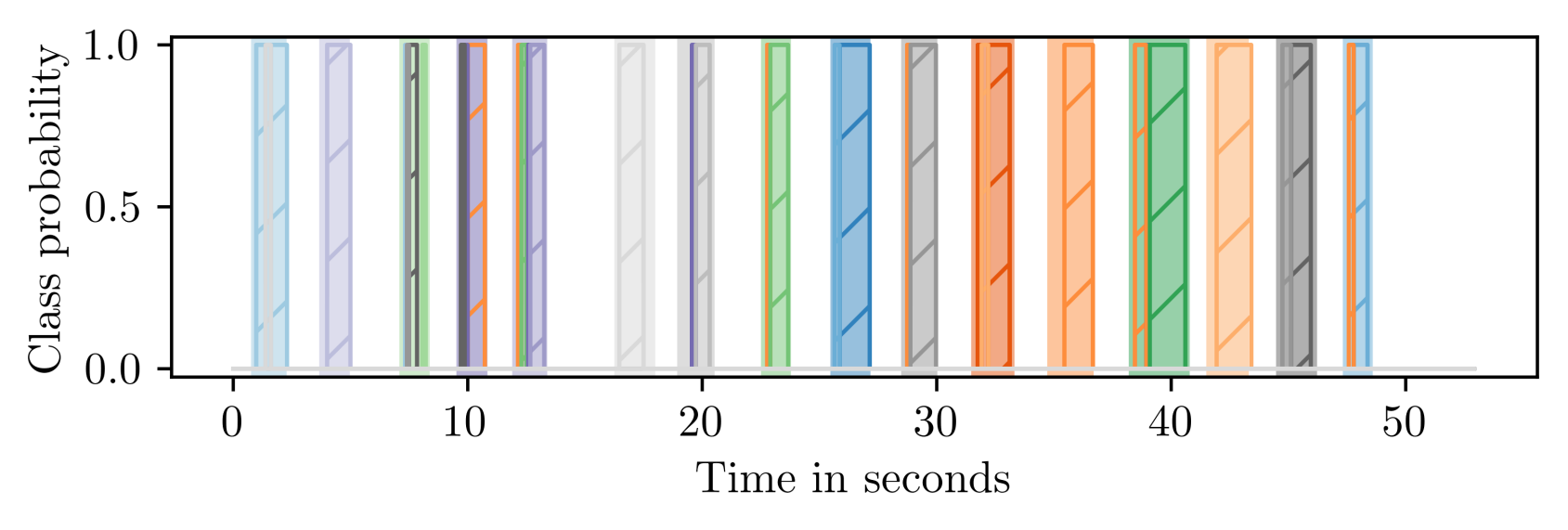
(b) Swipe Right

Framewise Classification

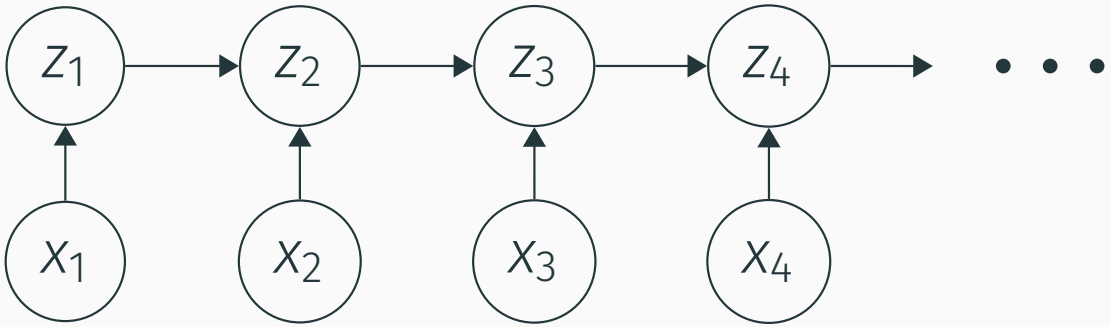


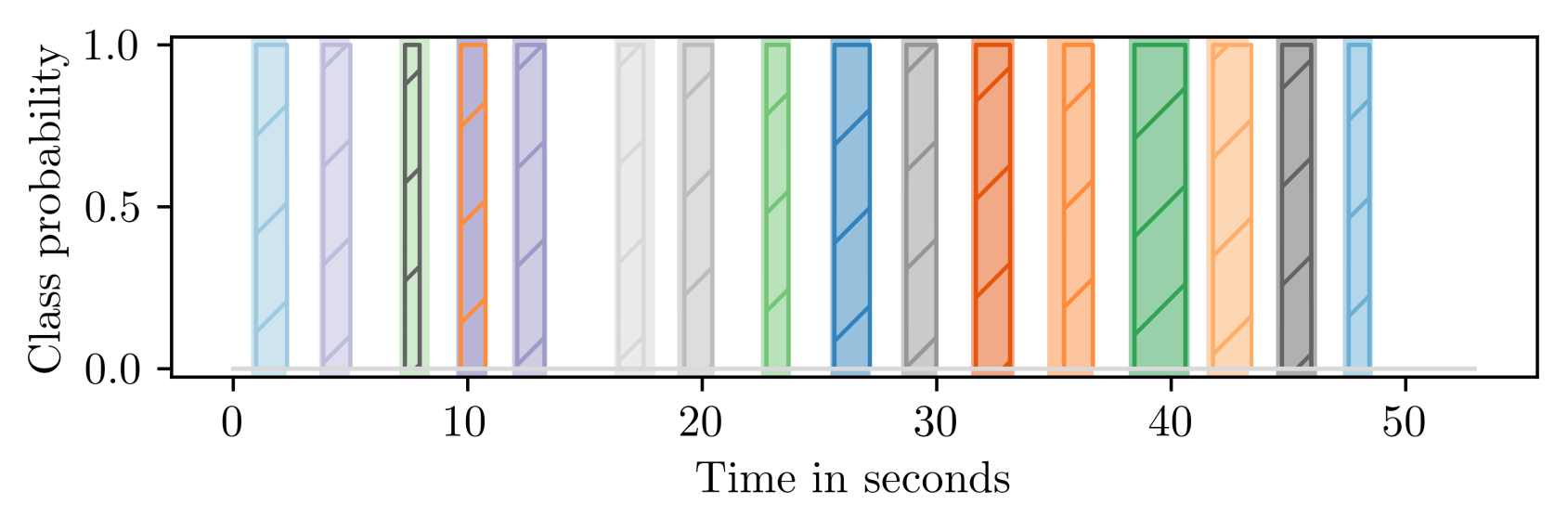
HMM Decoding



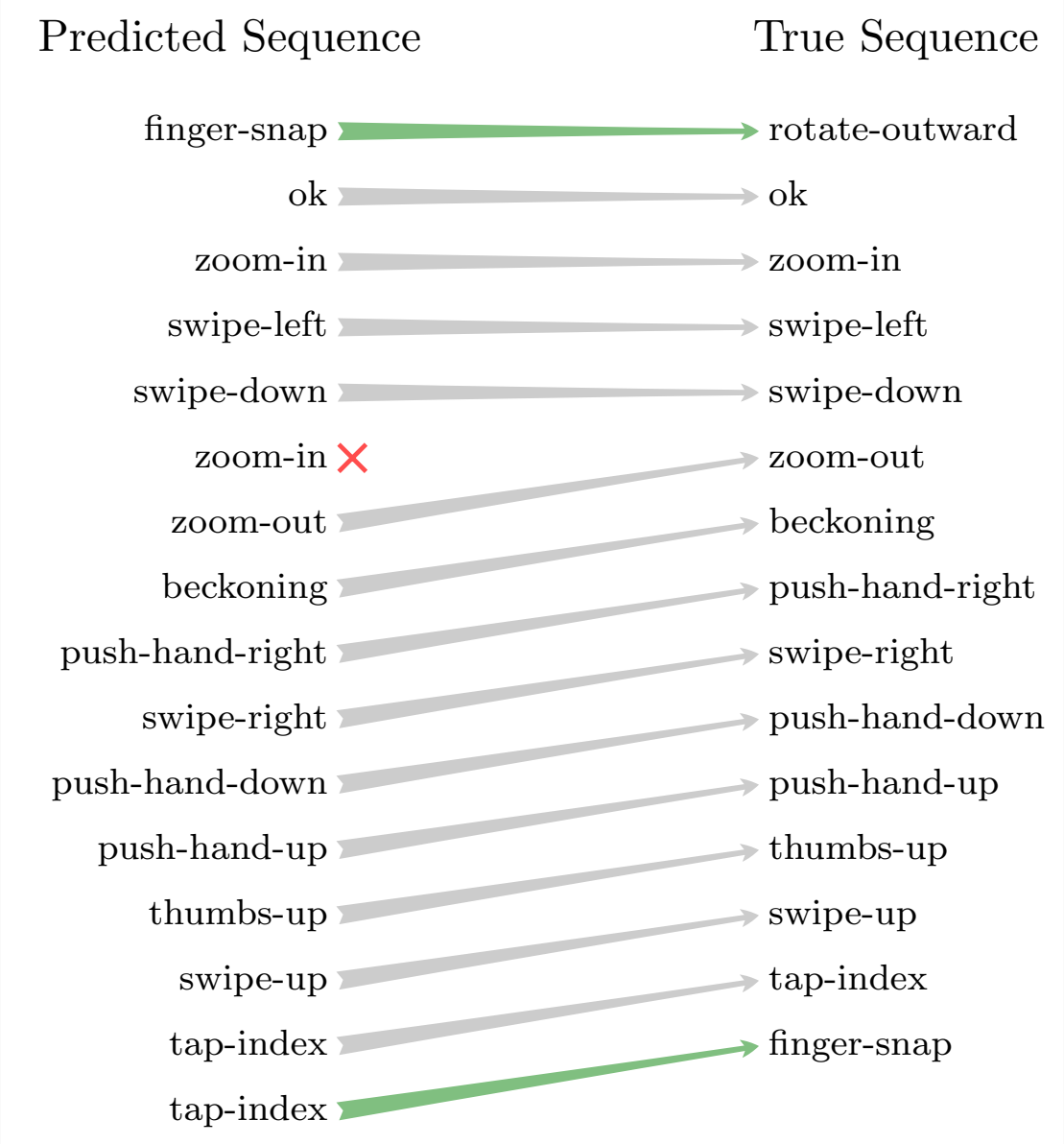


HMM Segmentation

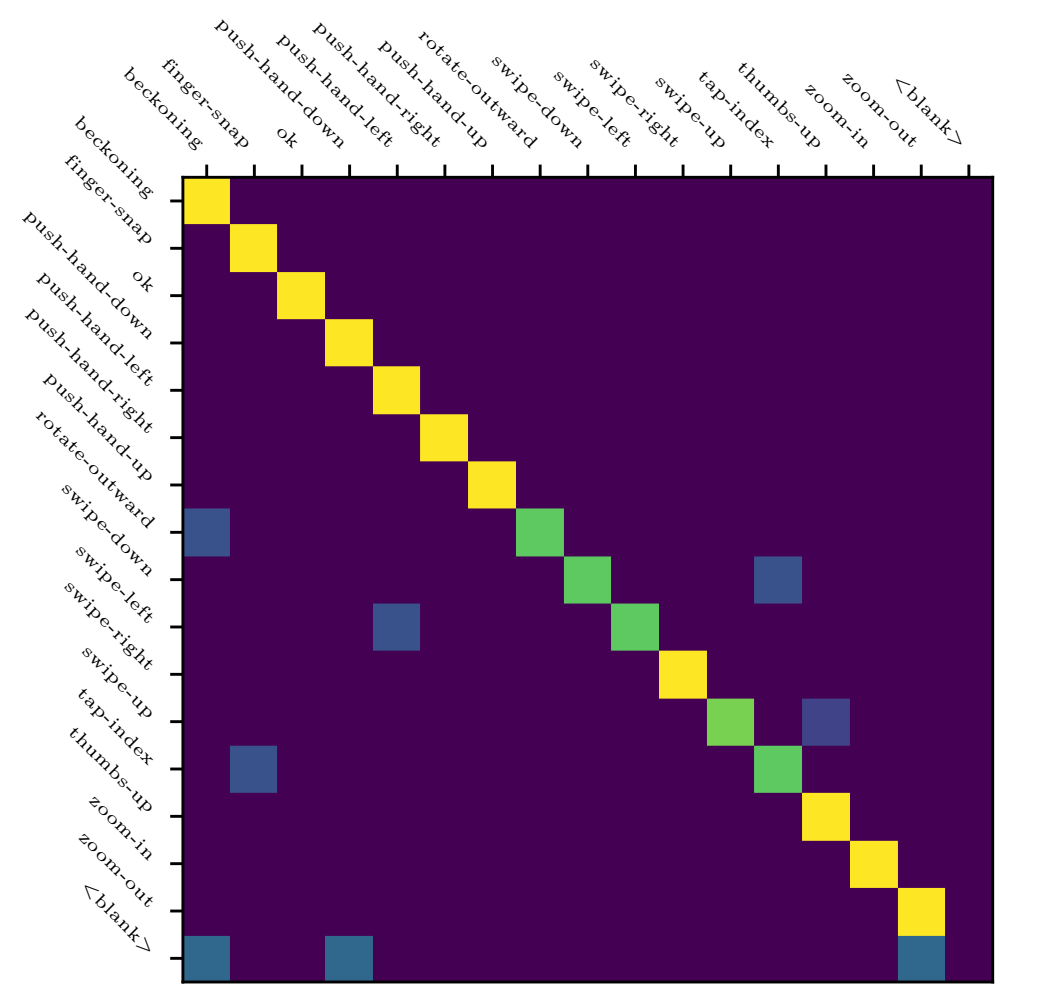




Predictions



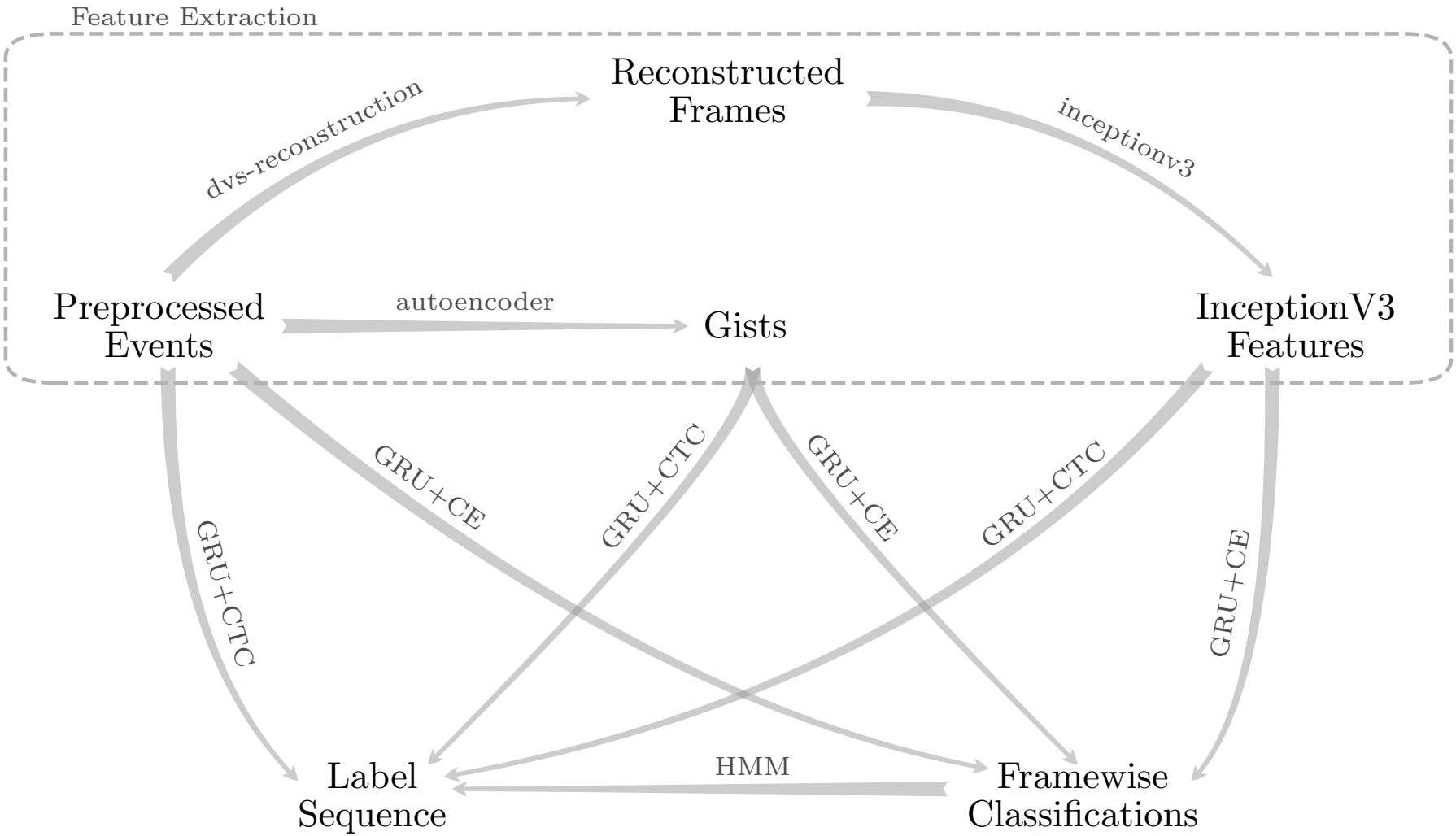
Results



Mean Levenshtein Distance 2.0, Label Error Rate 0.125

Thank You

Methods


Feature Extraction
Reconstructed Frames
dvs-reconstruction
inceptionv3
Preprocessed Events
autoencoder
Gists
InceptionV3 Features
GRU+CE
GRU+CTC
GRU+CE
GRU+CTC
GRU+CTC
GRU+CE
GRU+CE
Label Sequence
HMM
Framewise Classifications

Frame Reconstructions