UNIVERSITY OF AMSTERDAM

VIDEOGRAPH
Recognizing Minutes-Long Human Activities

NOURELDIEN HUSSEIN, EFSTRATIOS GAVVES, ARNOLD SMEULDERS

ICCV 2019
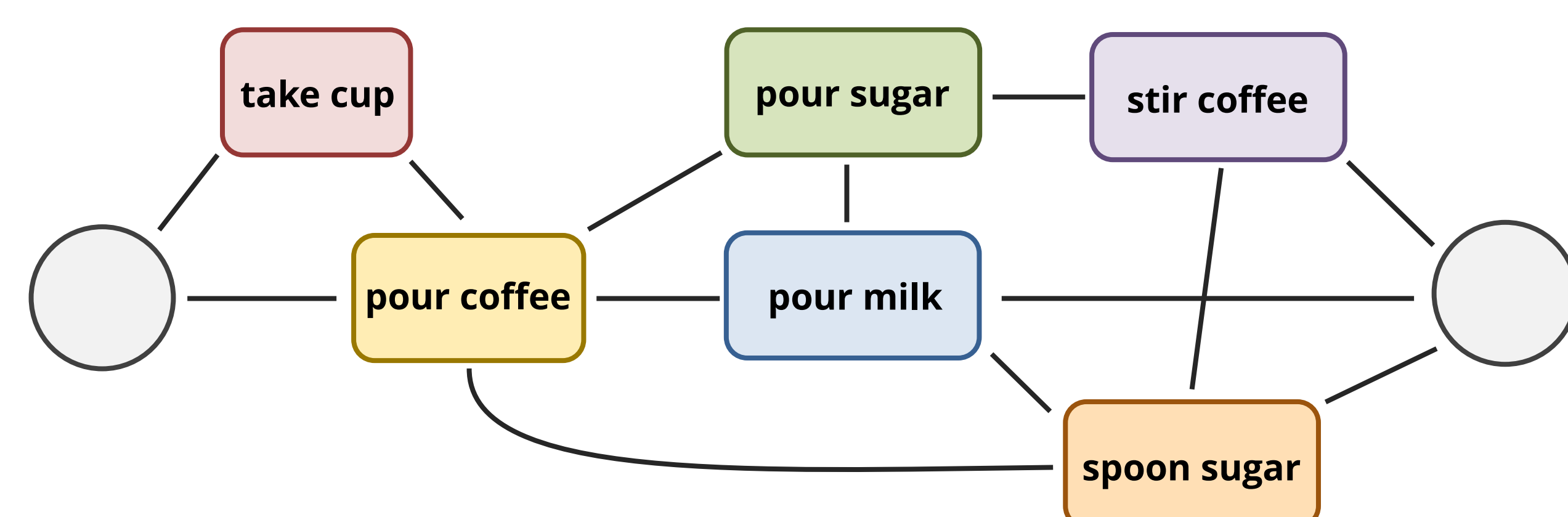Seoul, Korea

PROBLEM

THE GRAPH
Represents the structure of a long-range human activity.

THE NODES
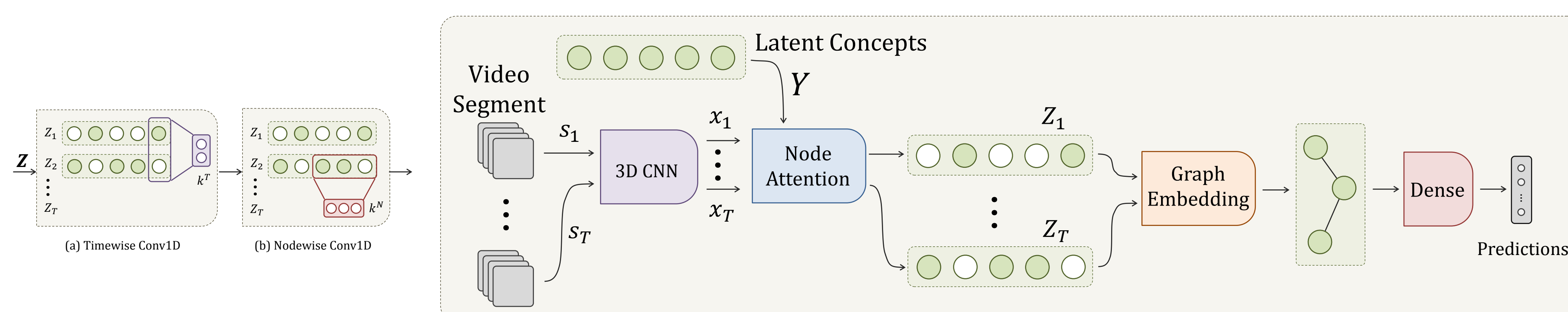Symbolize the core visual concepts that describe the activity.

THE EDGES
Stand for the relationships between these visual concepts.

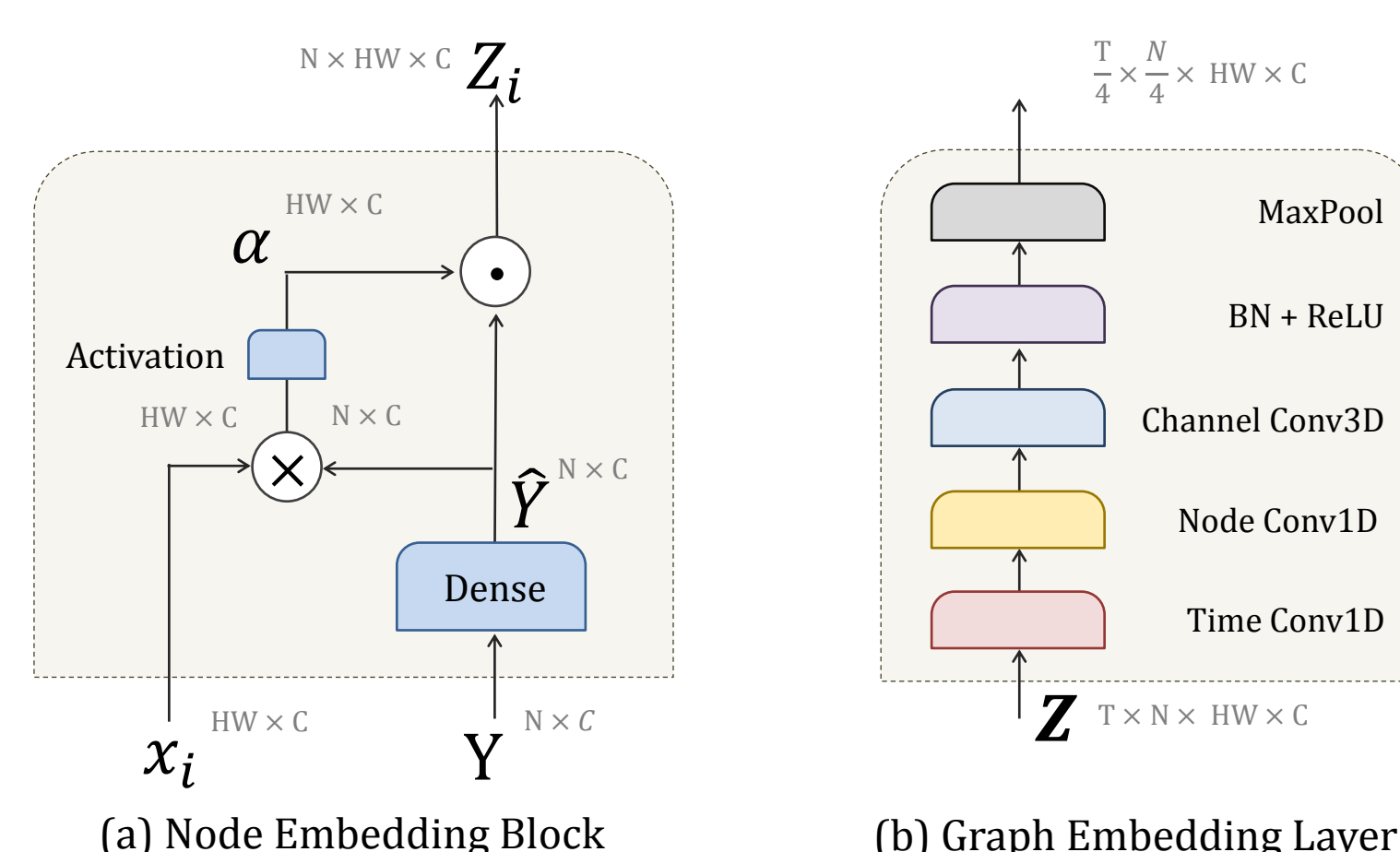Long-range Activity "Preparing Coffee"

METHOD

GRAPH NODES
Node attention block learns the node representations by measuring their similarity between the input features and the latent concepts.

NODE RELASHIONSHIPS
Nodewise 1D convolution learns the correlation between neighboring nodes.
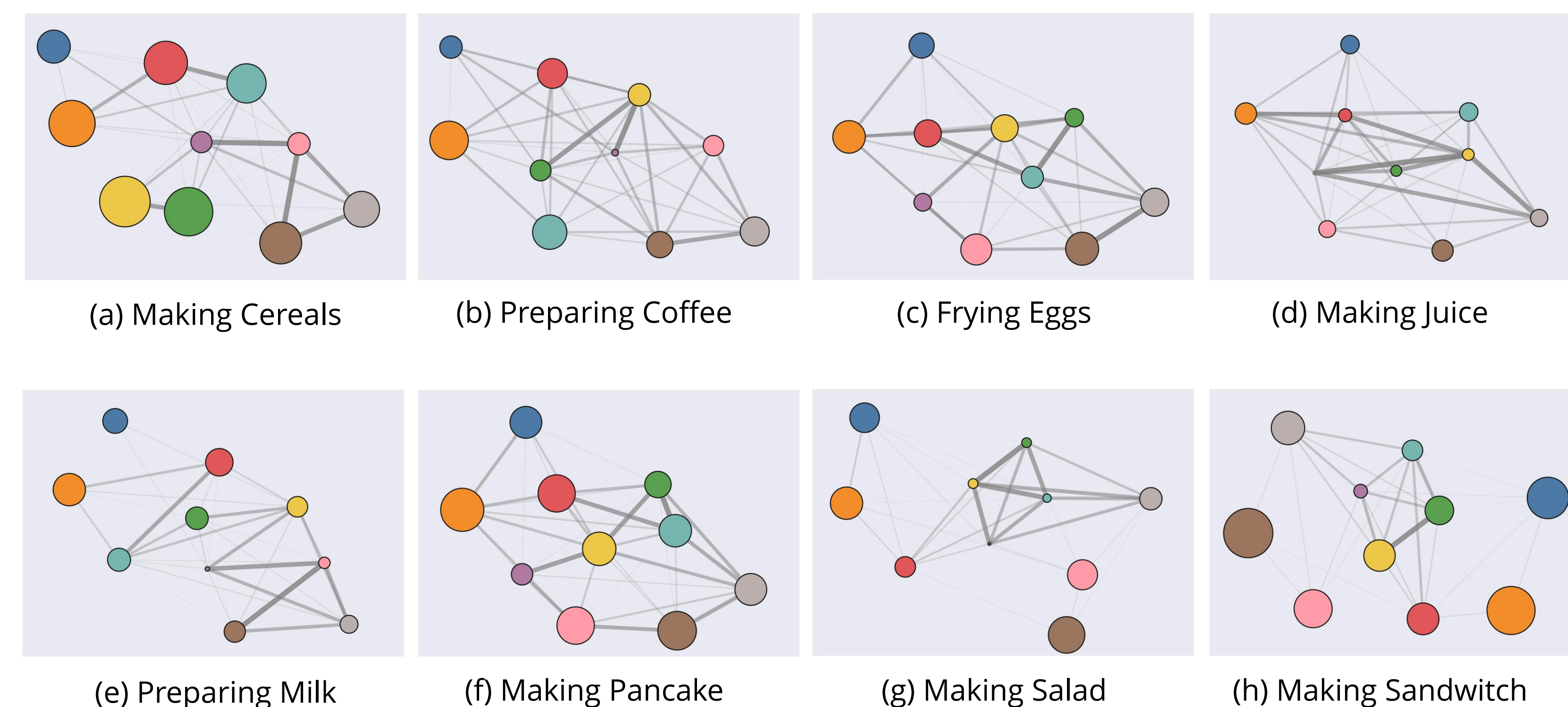
TIME RELASHIONSHIPS
Timewise 1D convolution learns how the graph nodes transition over times.

(a) Node Embedding Block   (b) Graph Embedding Layer

RESULTS

GRAPH VISUALIZATION

● Foodbox ● Kettle ● Eggs ● Cereal ● Sandwich ● Pan

(a) Making Cereals   (b) Preparing Coffee   (c) Frying Eggs   (d) Making Juice
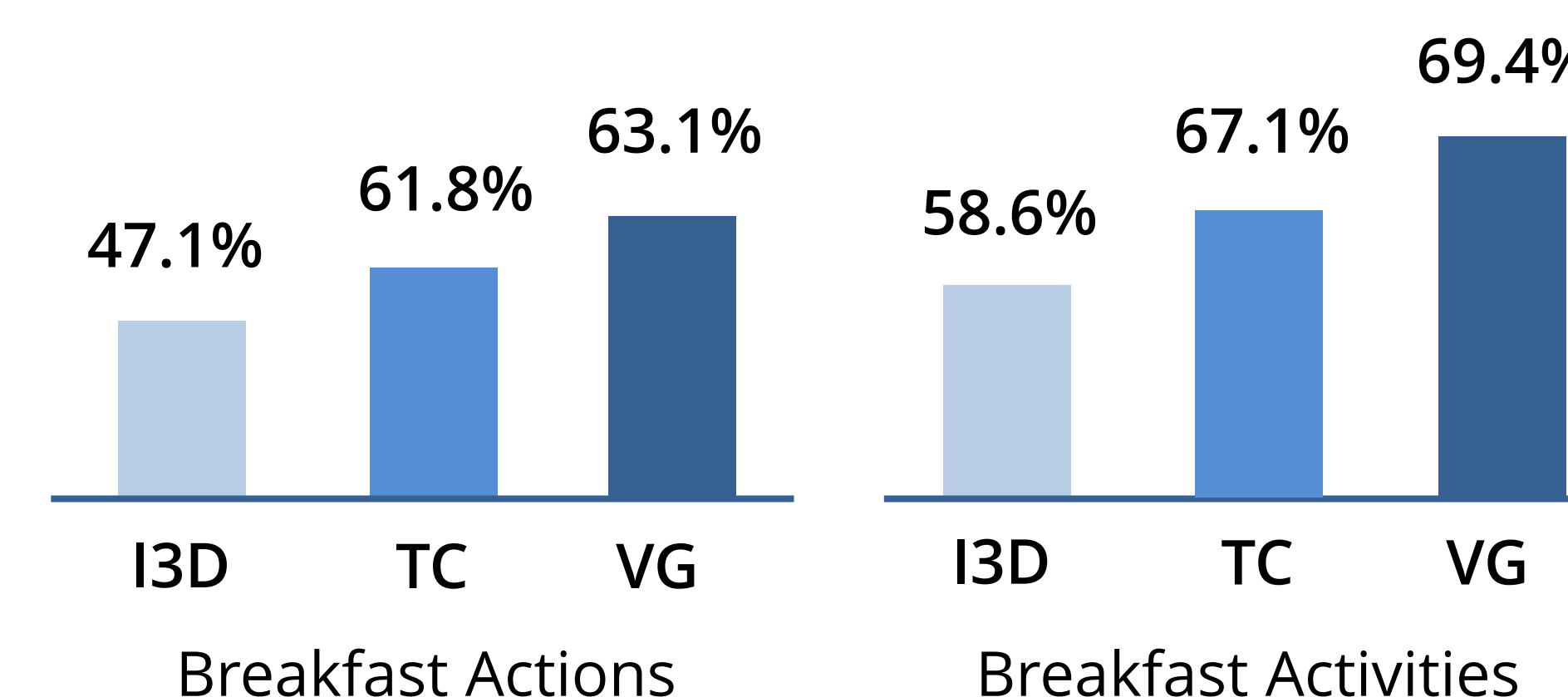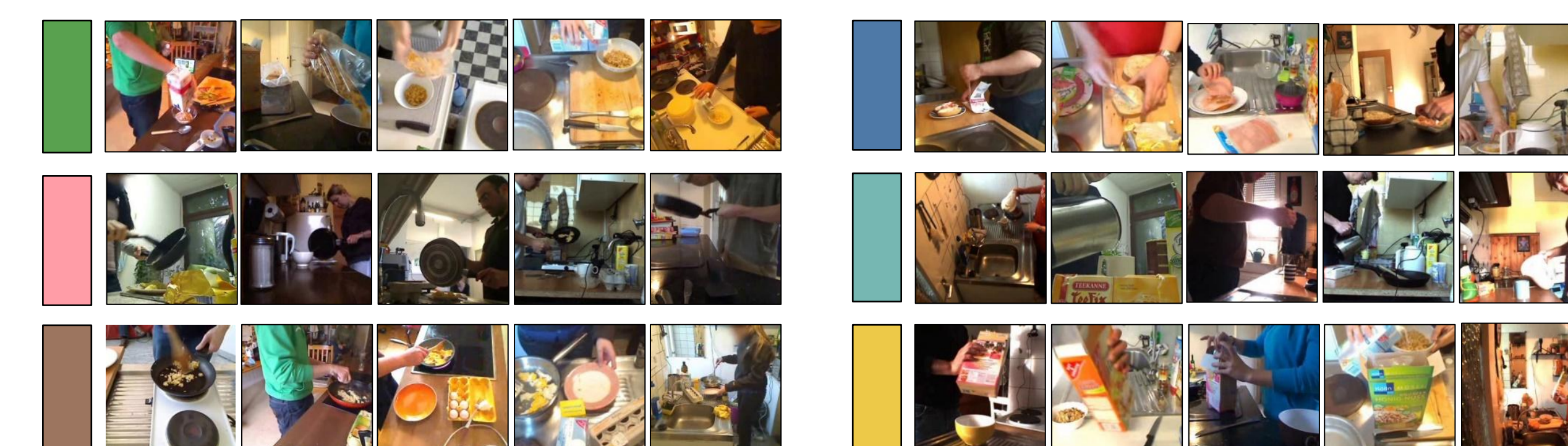
(e) Preparing Milk   (f) Making Pancake   (g) Making Salad   (h) Making Sandwitch

LEARNED NODES

Semantics of the discovered nodes by the node attention block.

The nodes represent the most discriminant visual evidences.

47.1%   61.8%   63.1%
I3D   TC   VG
Breakfast Actions

58.6%   67.1%   69.4%
I3D   TC   VG
Breakfast Activities

EFFECTIVENES

Videograph (VG) improves over related methods Timeception (TC) and I3D.

The same result is confirmed for single- and multi-label classification of Breakfast.

CHARADES RESULTS

VideoGraph improves on related works, when classifying the activities of Charades.

VG   37.8%
TC   37.2%
I3D   31.8%