Permutation Invariant Convolution

For Recognizing Long-range Activities



Noureldien Hussein Efstratios Gavves Arnold Smeulders

University of Amsterdam

Motivation

Motivation

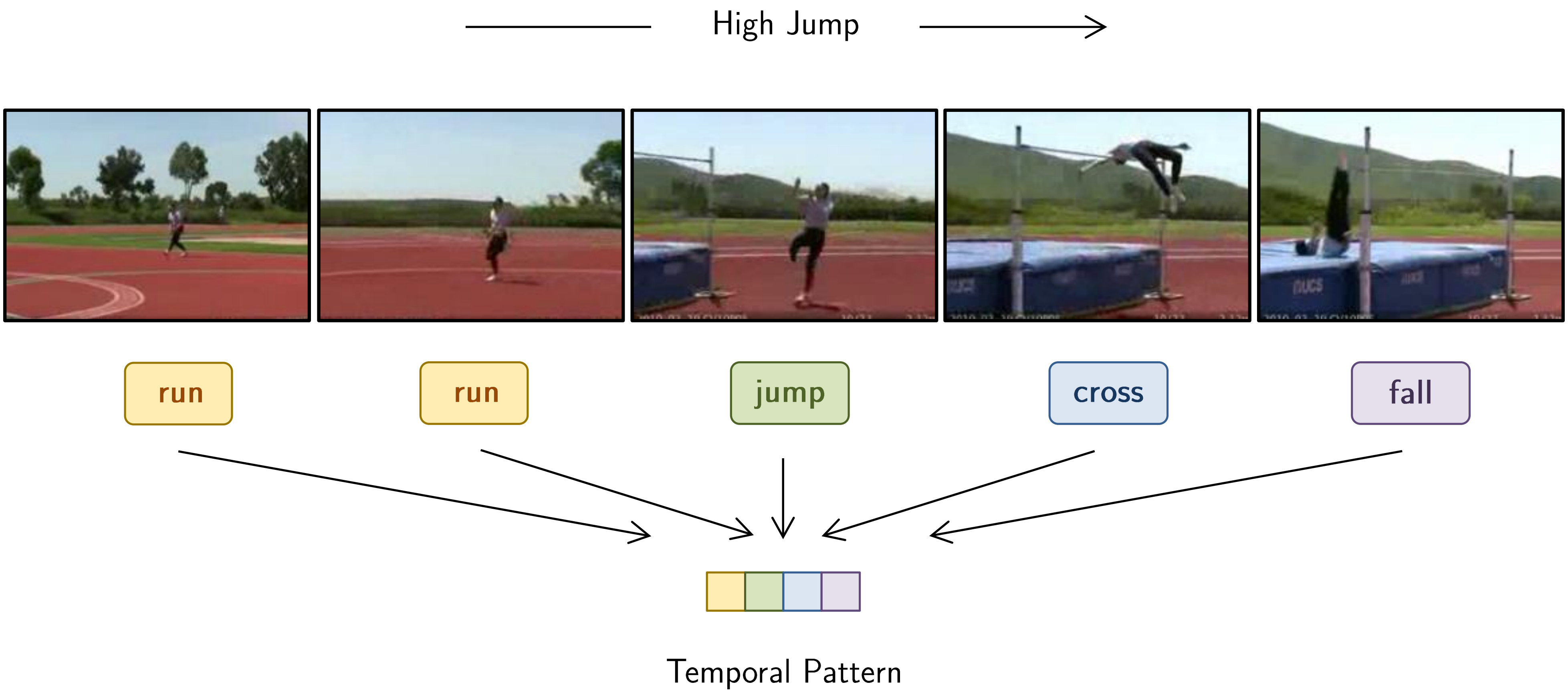
High Jump

Motivation

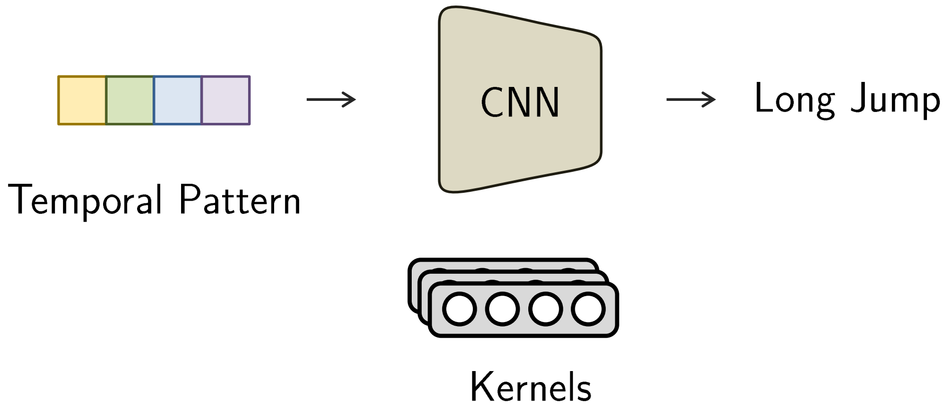
High Jump

run run jump cross fall

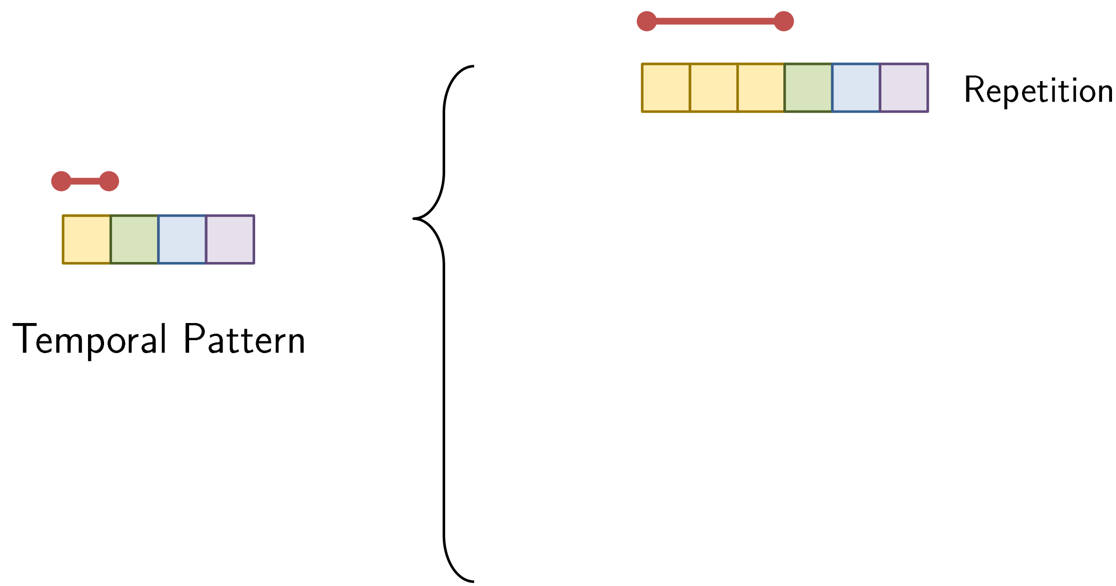
Motivation



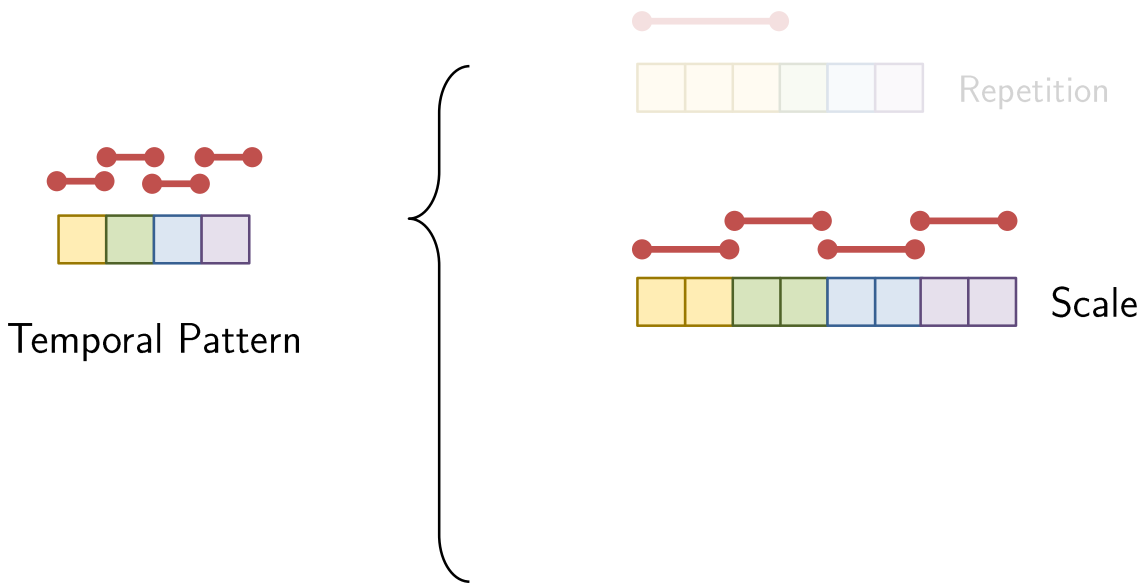
Motivation



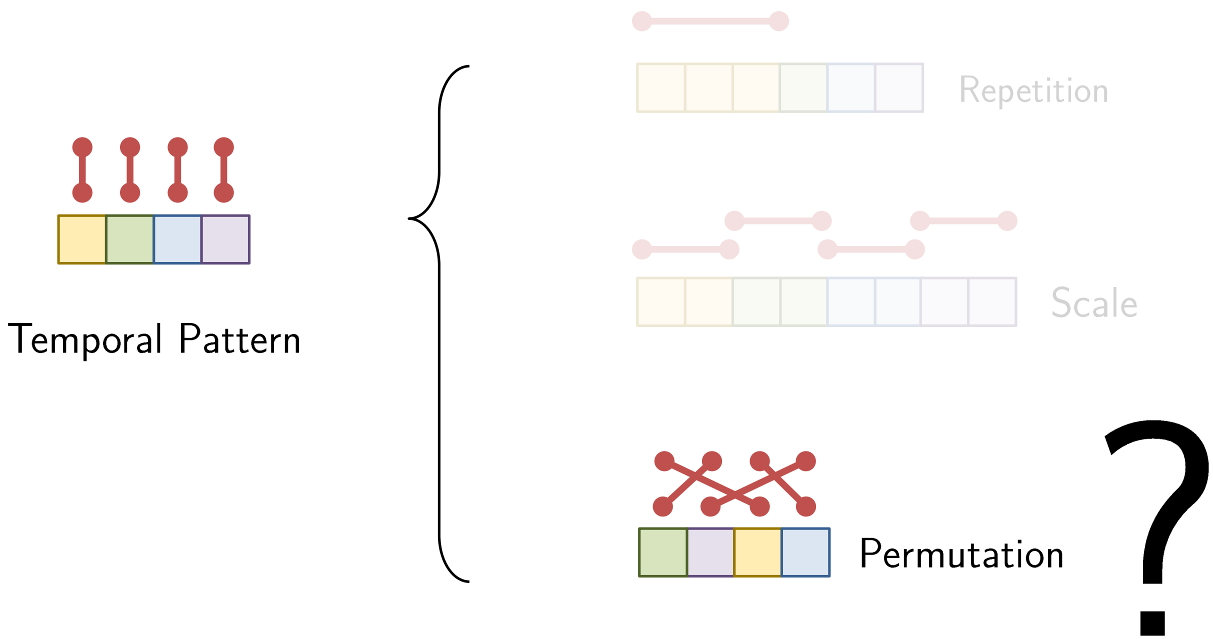
Motivation



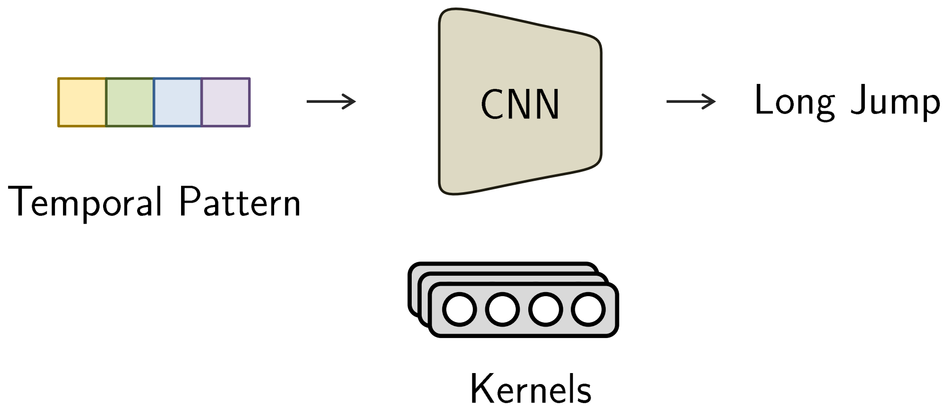
Motivation



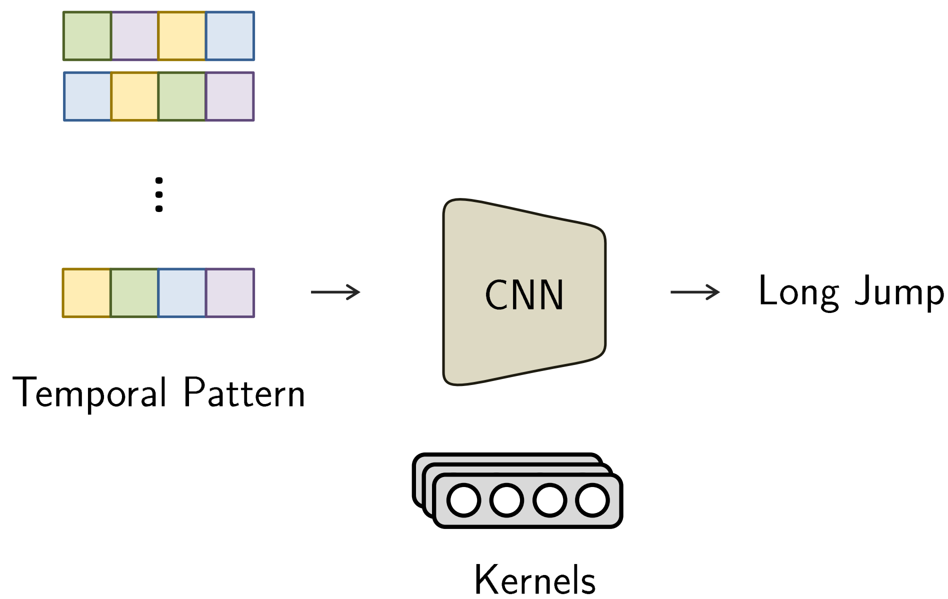
Motivation



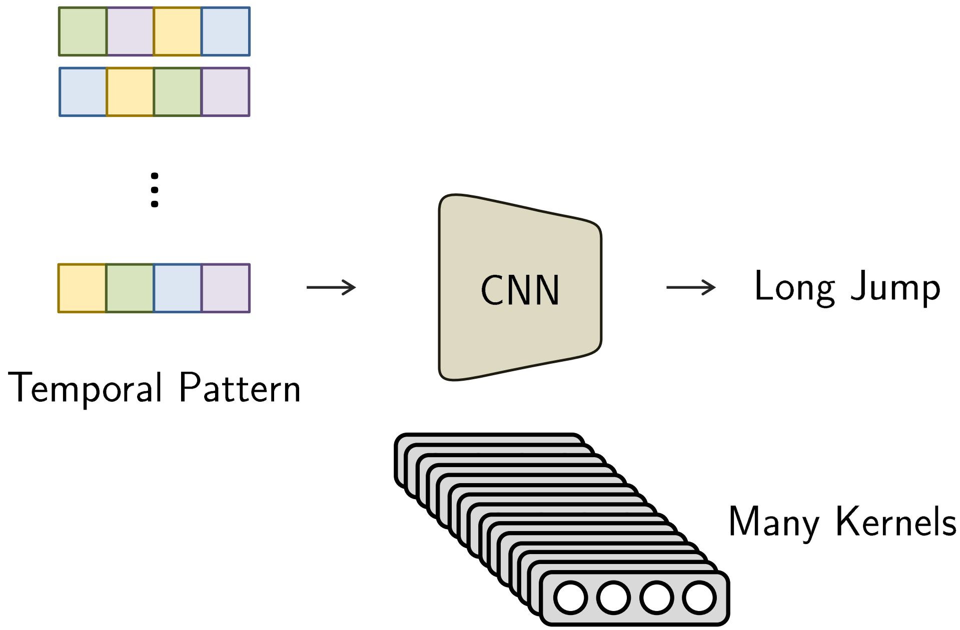
Motivation



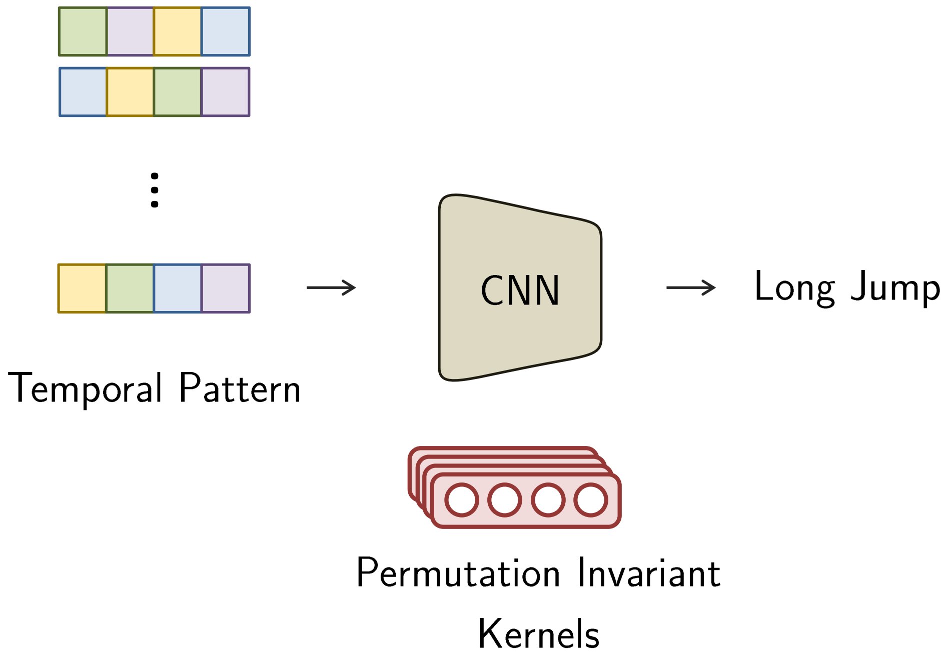
Motivation



Motivation

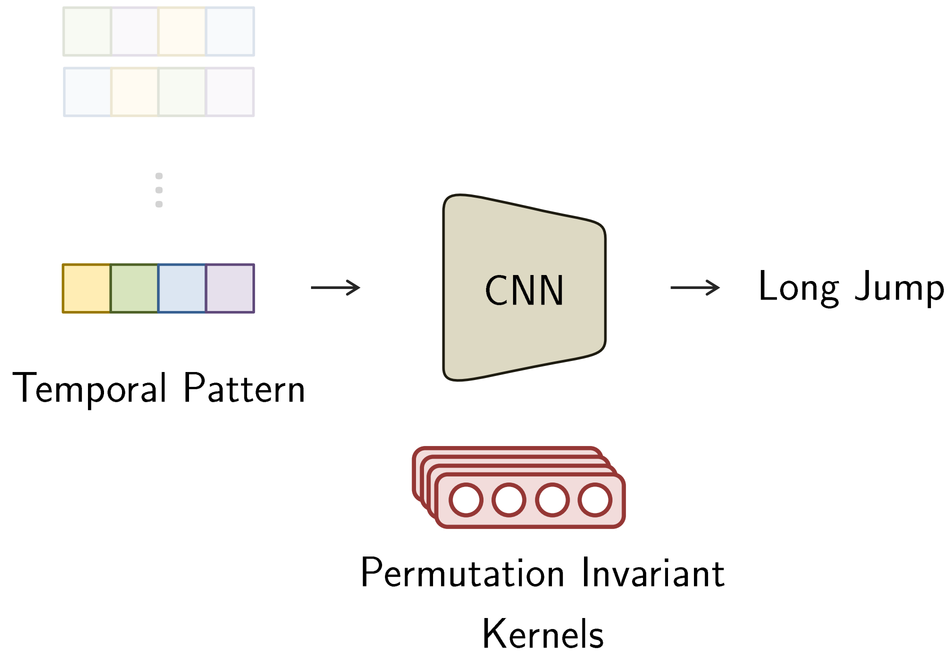


Motivation



CNN
Long Jump
Temporal Pattern
Permutation Invariant
Kernels

Motivation



CNN
Long Jump
Temporal Pattern
Permutation Invariant
Kernels

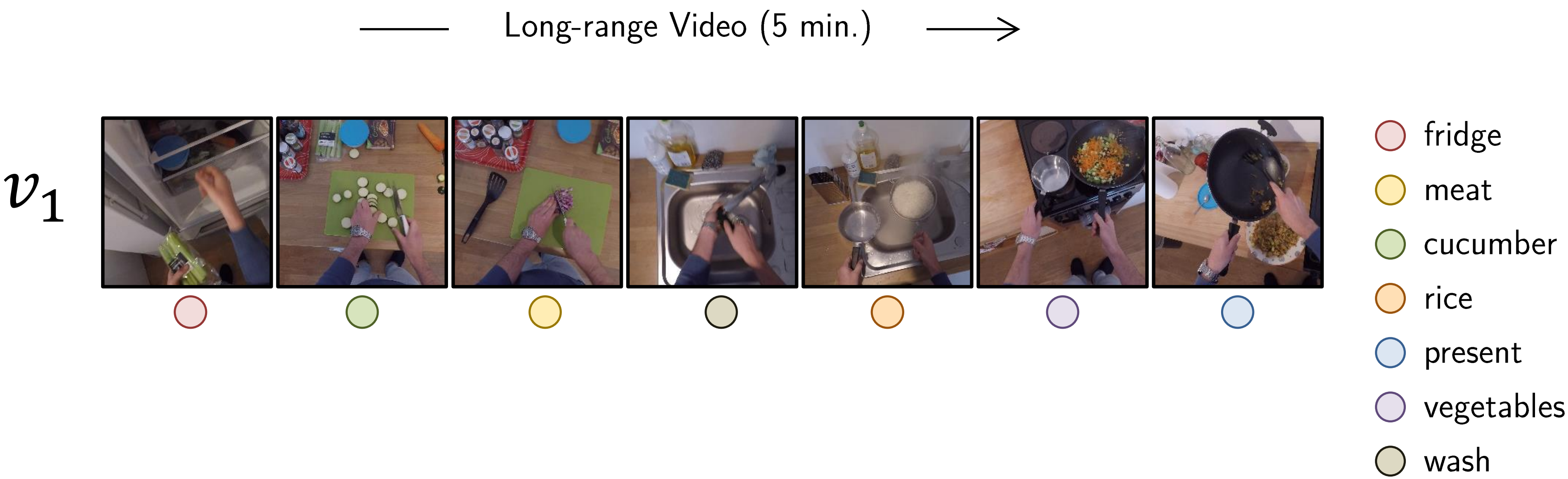
Permutation?

Motivation

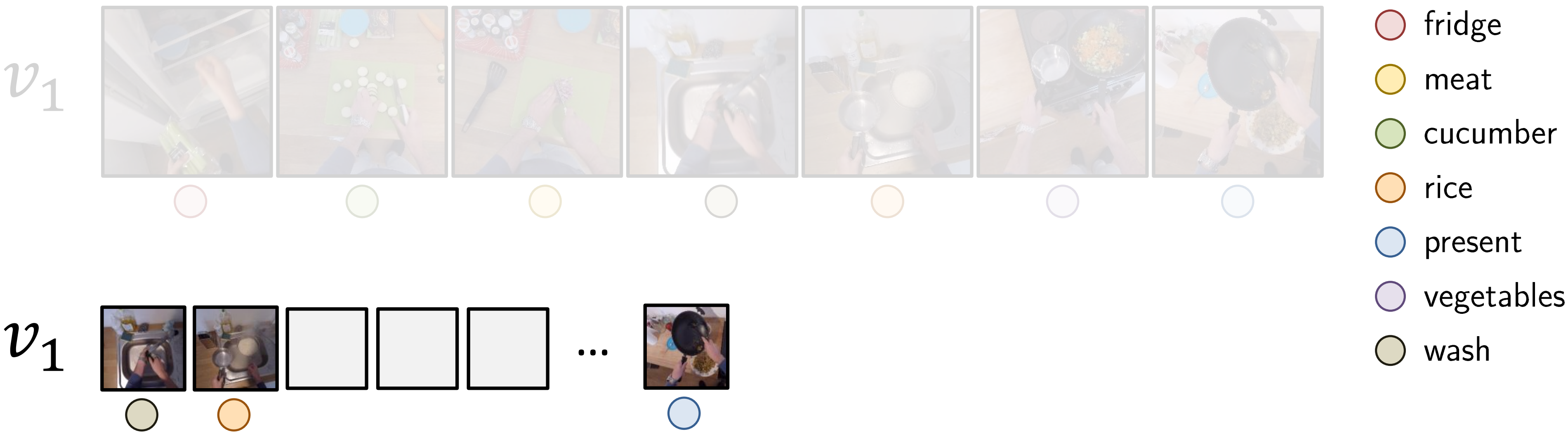
Short-range Video (10 sec.)

run jump cross fall

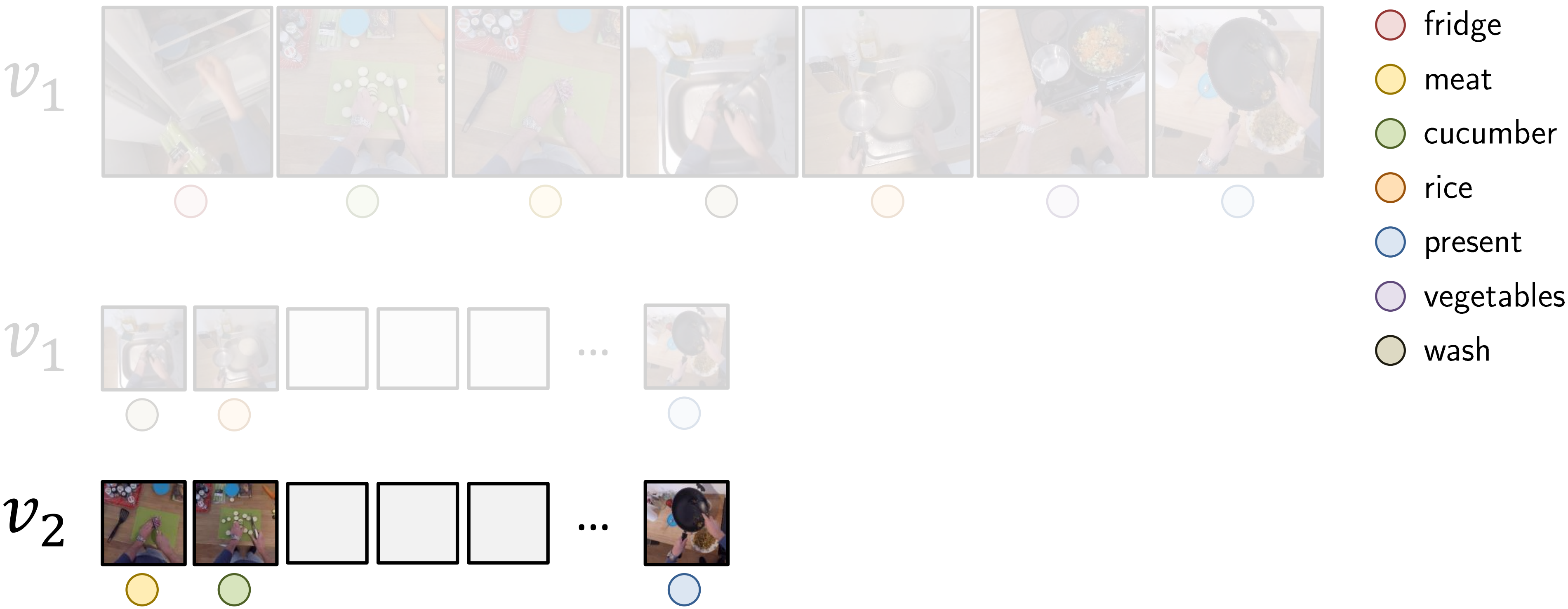
Motivation


Long-range Video (5 min.)
v_1
fridge
meat
cucumber
rice
present
vegetables
wash

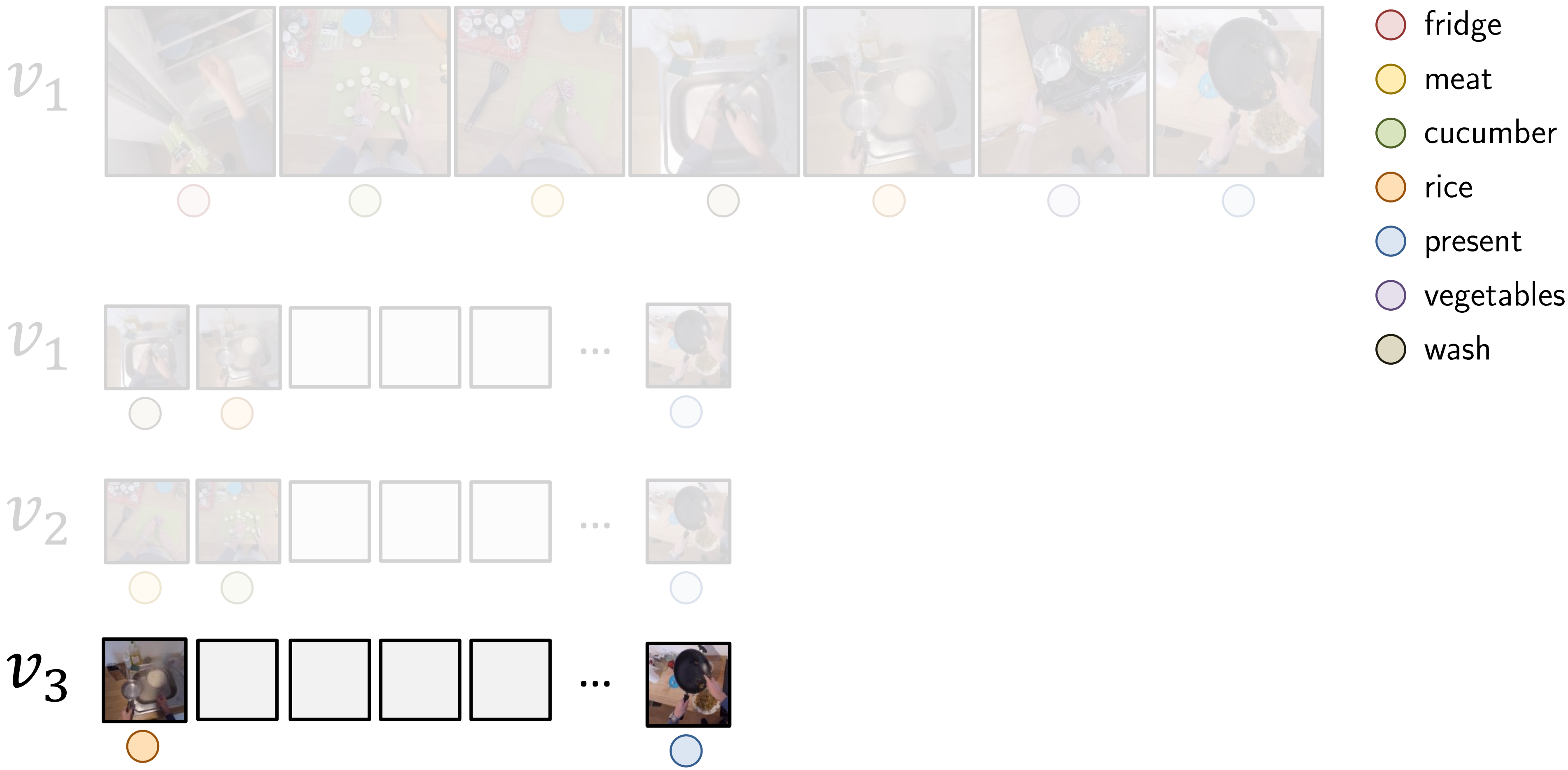
Motivation



Motivation

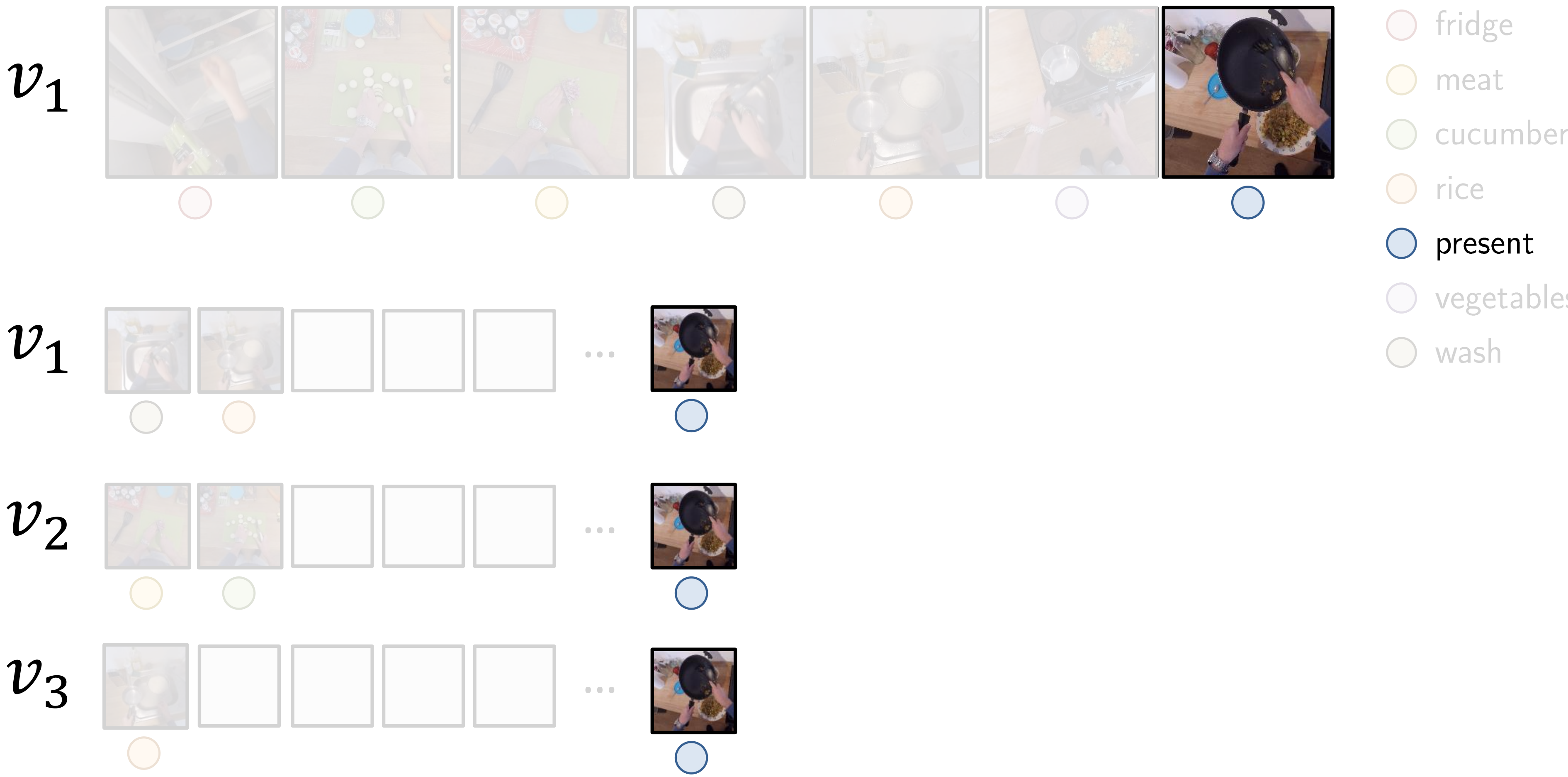


Motivation



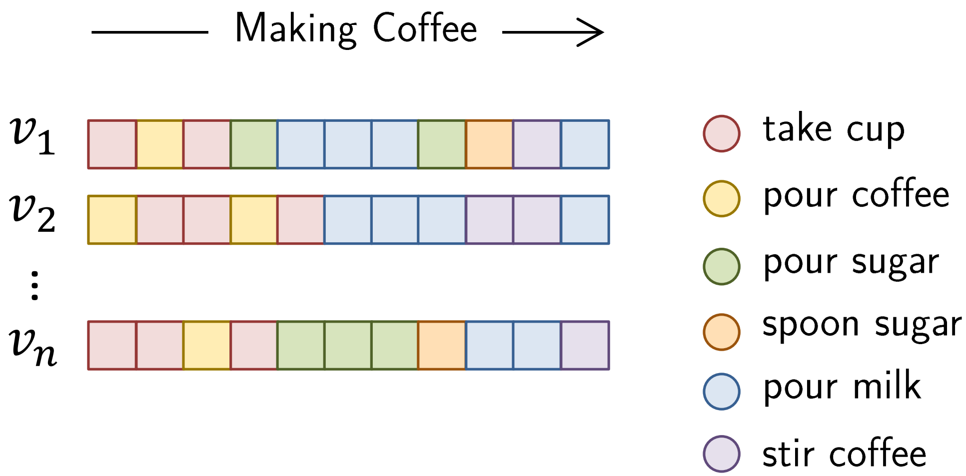
v_1
v_1
v_2
v_3
...
fridge
meat
cucumber
rice
present
vegetables
wash

Motivation


v_1
v_1
v_2
v_3
fridge
meat
cucumber
rice
present
vegetables
wash

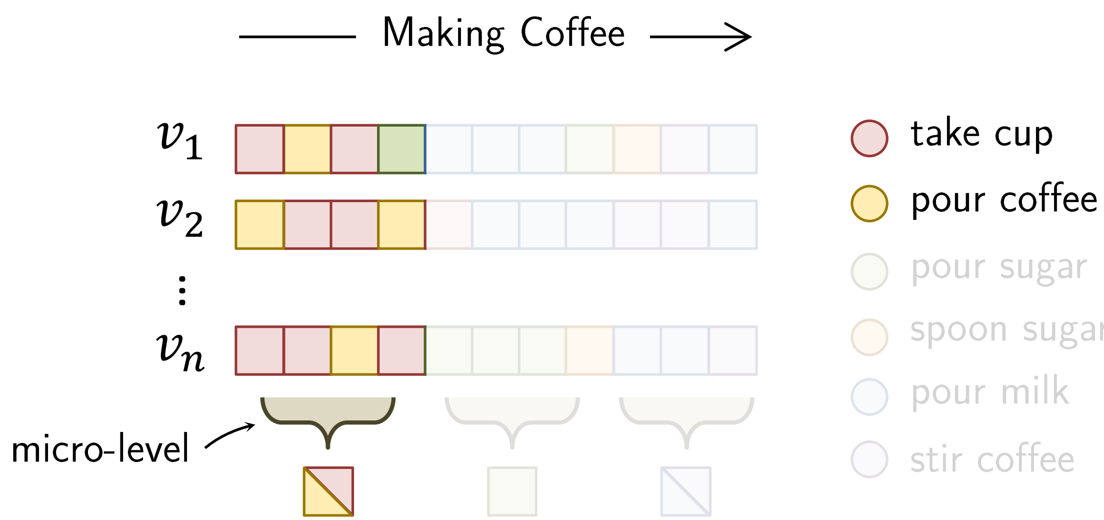
Method

Solution

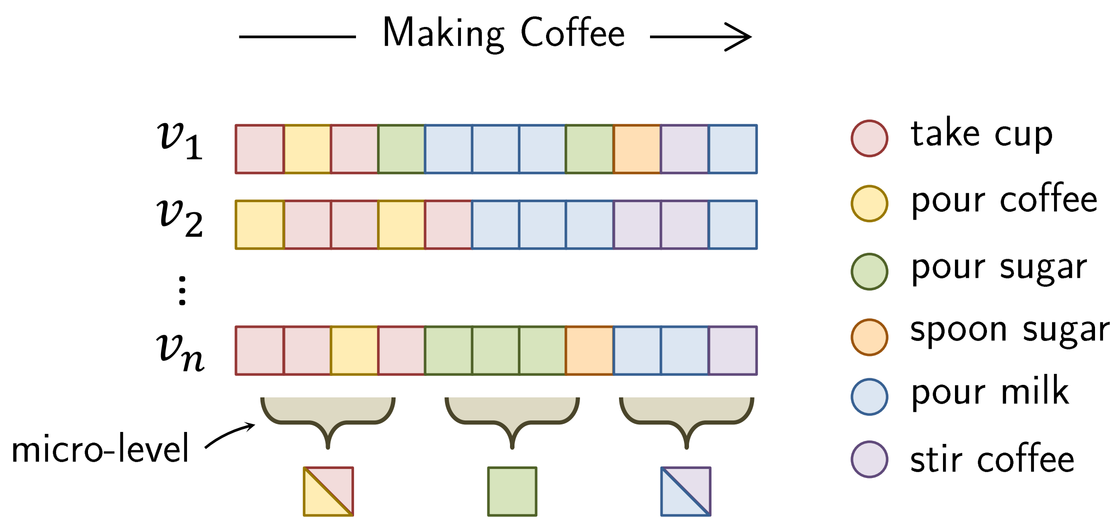


Making Coffee
v_1
v_2
⋮
v_n
take cup
pour coffee
pour sugar
spoon sugar
pour milk
stir coffee

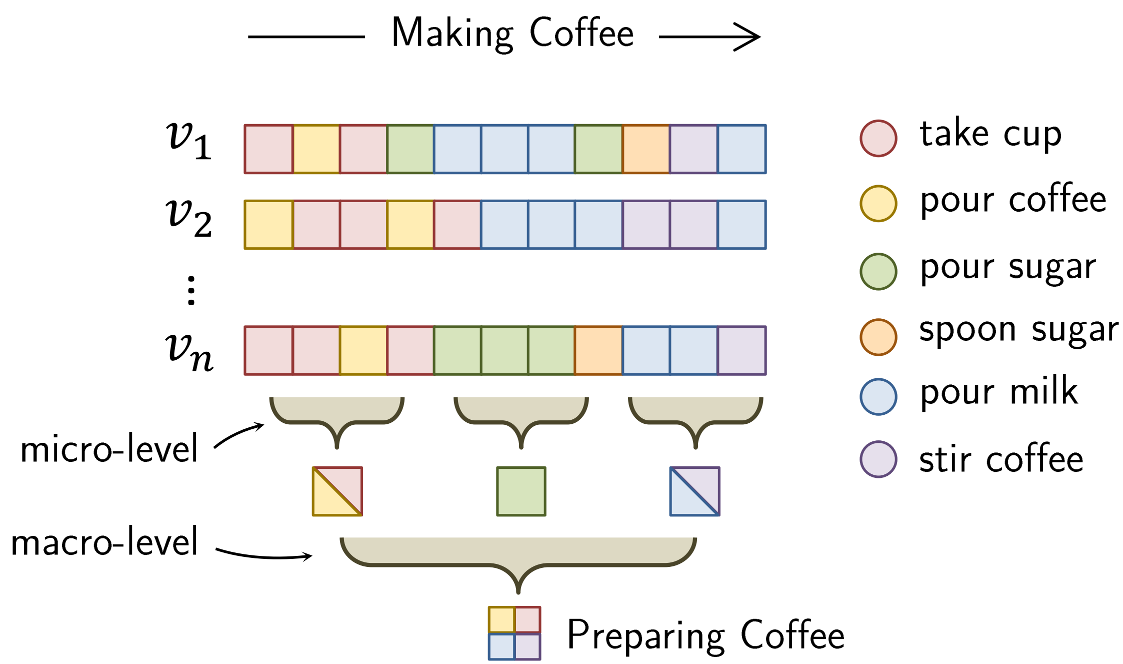
Solution


Making Coffee
v_1
v_2
$\vdots$
v_n
micro-level
take cup
pour coffee
pour sugar
spoon sugar
pour milk
stir coffee

Solution

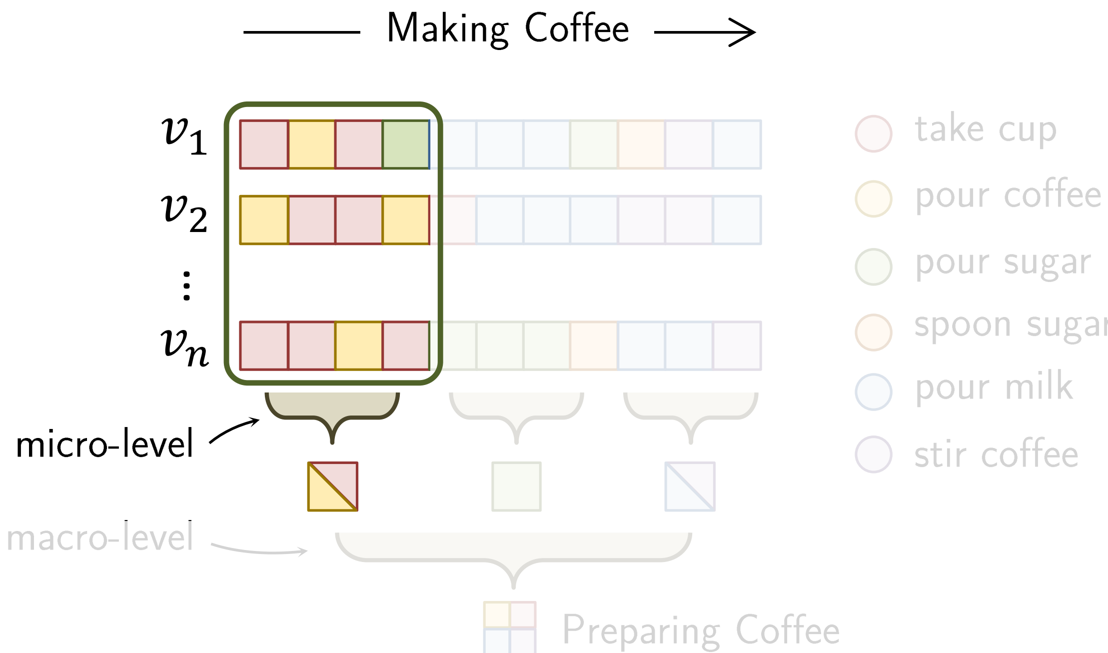

Making Coffee
v_1
v_2
v_n
micro-level
take cup
pour coffee
pour sugar
spoon sugar
pour milk
stir coffee

Solution

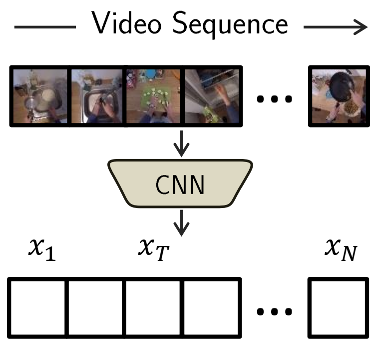


Making Coffee
v_1
v_2
v_n
take cup
pour coffee
pour sugar
spoon sugar
pour milk
stir coffee
micro-level
macro-level
Preparing Coffee

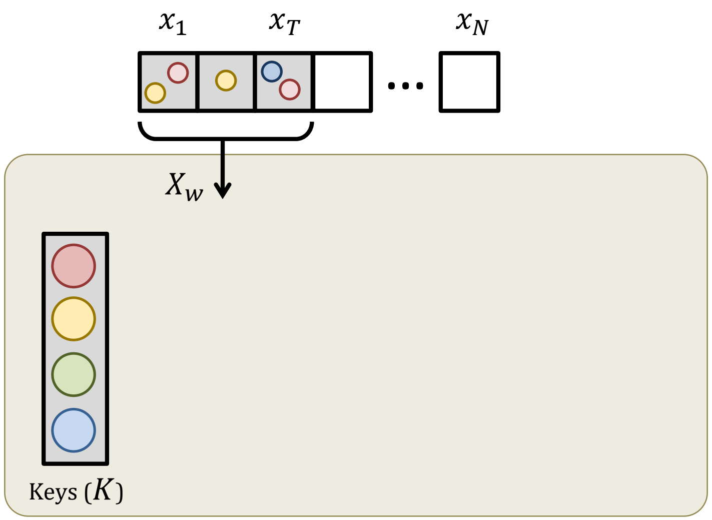
Solution


Making Coffee
v_1
v_2
v_n
micro-level
macro-level
Preparing Coffee
take cup
pour coffee
pour sugar
spoon sugar
pour milk
stir coffee

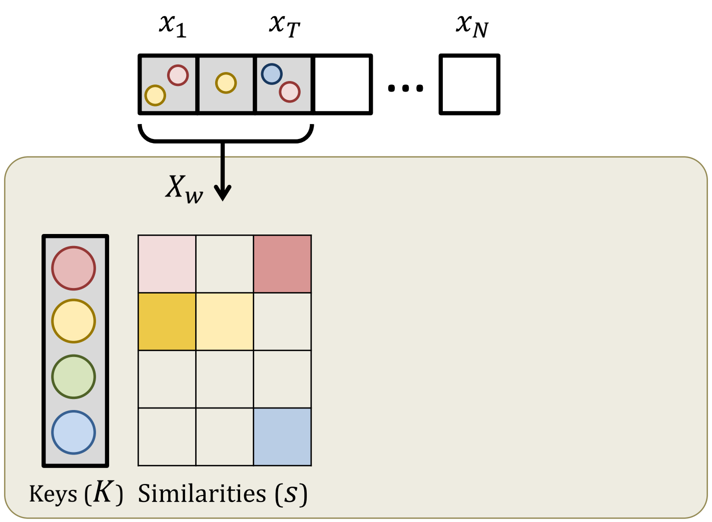
Solution



Solution

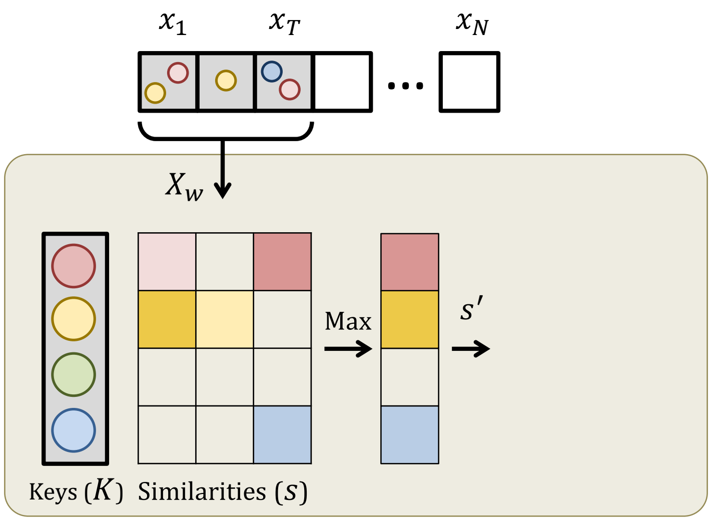


Solution



$$s = K \otimes X_w^T \qquad (1)$$

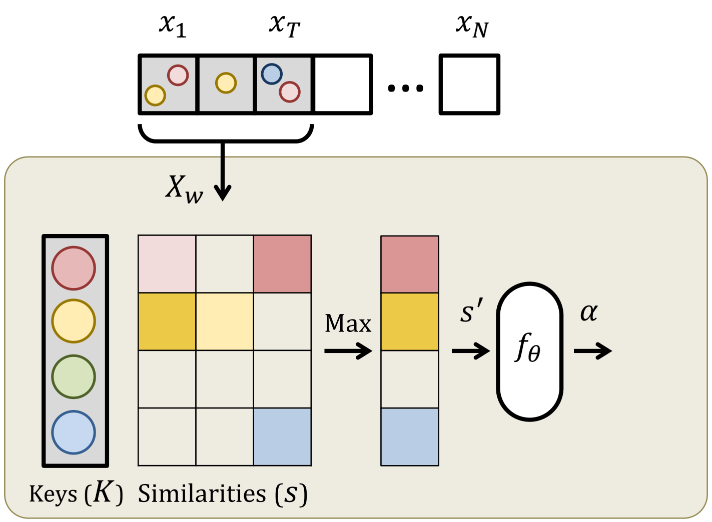
Solution



$$s = K \otimes X_w^T \tag{1}$$

$$s' = \max_{\text{row}}(s) \tag{2}$$

Solution

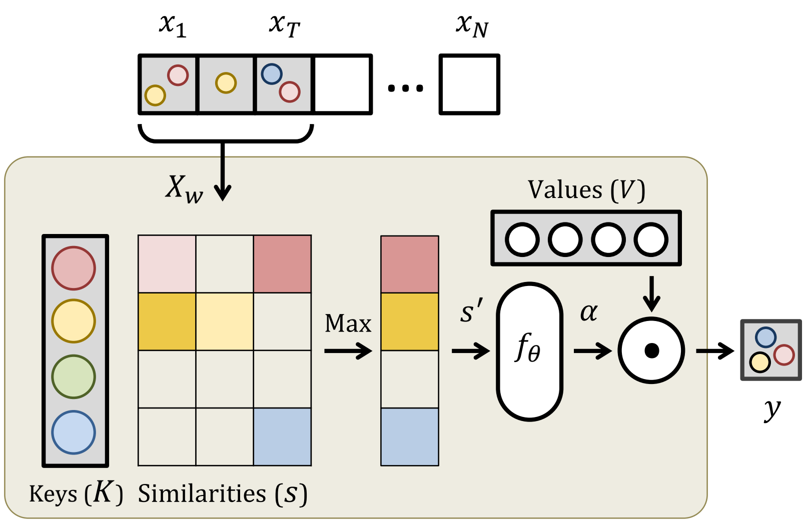


$$s = K \otimes X_w^T \qquad (1)$$

$$s' = \max_{\text{row}}(s) \qquad (2)$$

$$\alpha = \sigma\left[\, f_\theta(s') \,\right] \qquad (3)$$

Solution



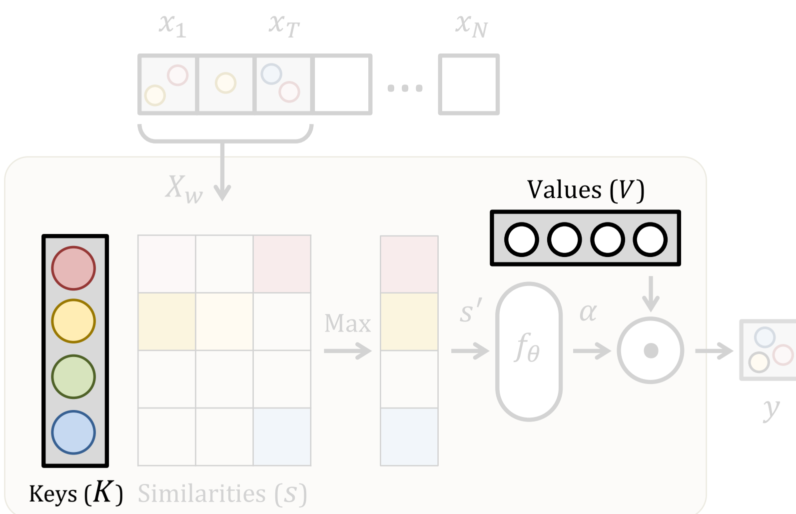
$$s = K \otimes X_w^T \tag{1}$$

$$s' = \max_{\text{row}}(s) \tag{2}$$

$$\alpha = \sigma\,[\, f_\theta(s')\,] \tag{3}$$

$$y = \alpha \odot V \tag{4}$$

Solution



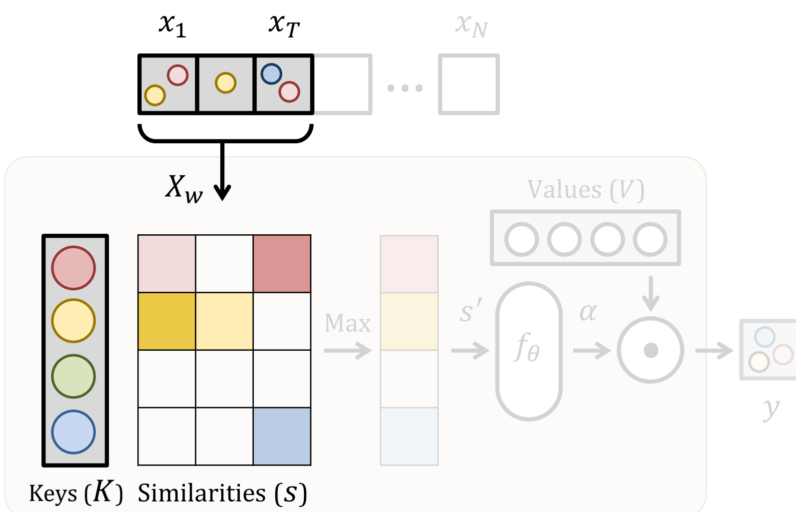
$$s = K \otimes X_w^T \tag{1}$$

$$s' = \max_{\text{row}}(s) \tag{2}$$

$$\alpha = \sigma\,[\,f_\theta(s')\,] \tag{3}$$

$$y = \alpha \odot V \tag{4}$$

Solution



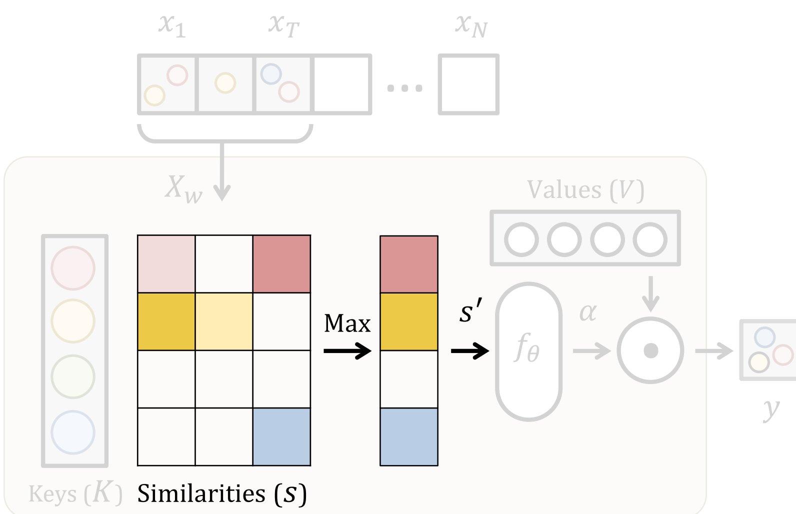
$$s = K \otimes X_w^T \tag{1}$$

$$s' = \max_{row}(s) \tag{2}$$

$$\alpha = \sigma\,[\,f_\theta(s')\,] \tag{3}$$

$$y = \alpha \odot V \tag{4}$$

Solution



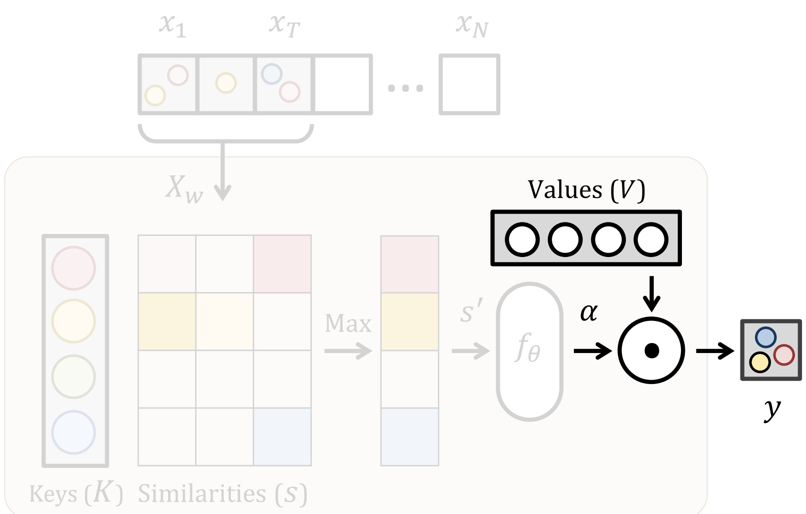
$$s = K \otimes X_w^T \tag{1}$$

$$s' = \max_{\text{row}}(s) \tag{2}$$

$$\alpha = \sigma\,[\, f_\theta(s')\,] \tag{3}$$

$$y = \alpha \odot V \tag{4}$$

Solution



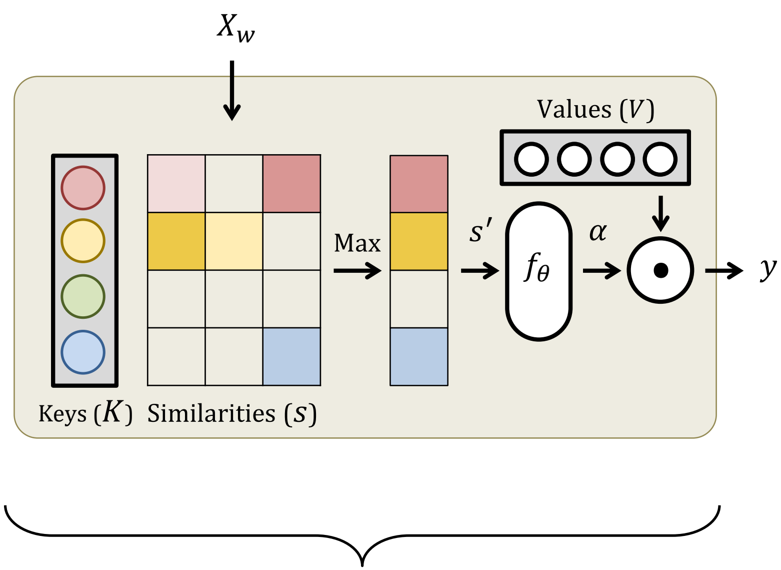
$$s = K \otimes X_w^T \tag{1}$$

$$s' = \max_{\text{row}}(s) \tag{2}$$

$$\alpha = \sigma\,[\,f_\theta(s')\,] \tag{3}$$

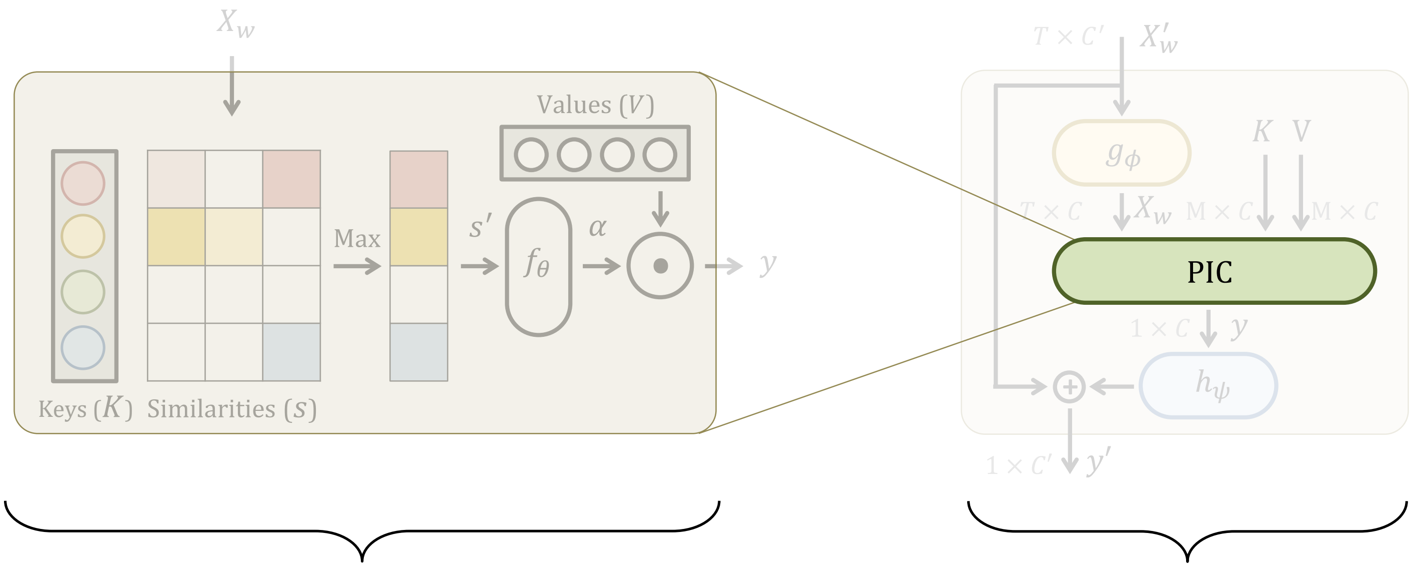
$$y = \alpha \odot V \tag{4}$$

Solution



PIC Operation

Solution

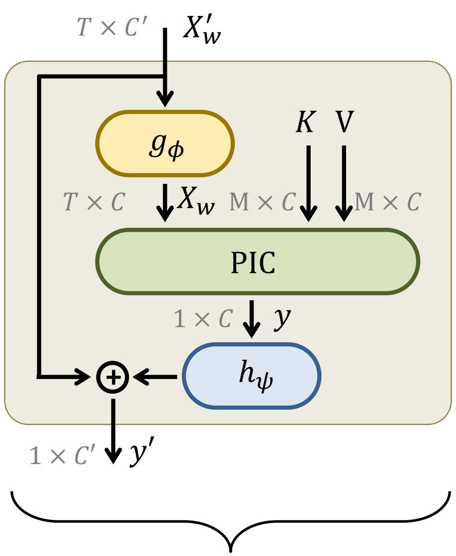


PIC Operation

PIC Layer

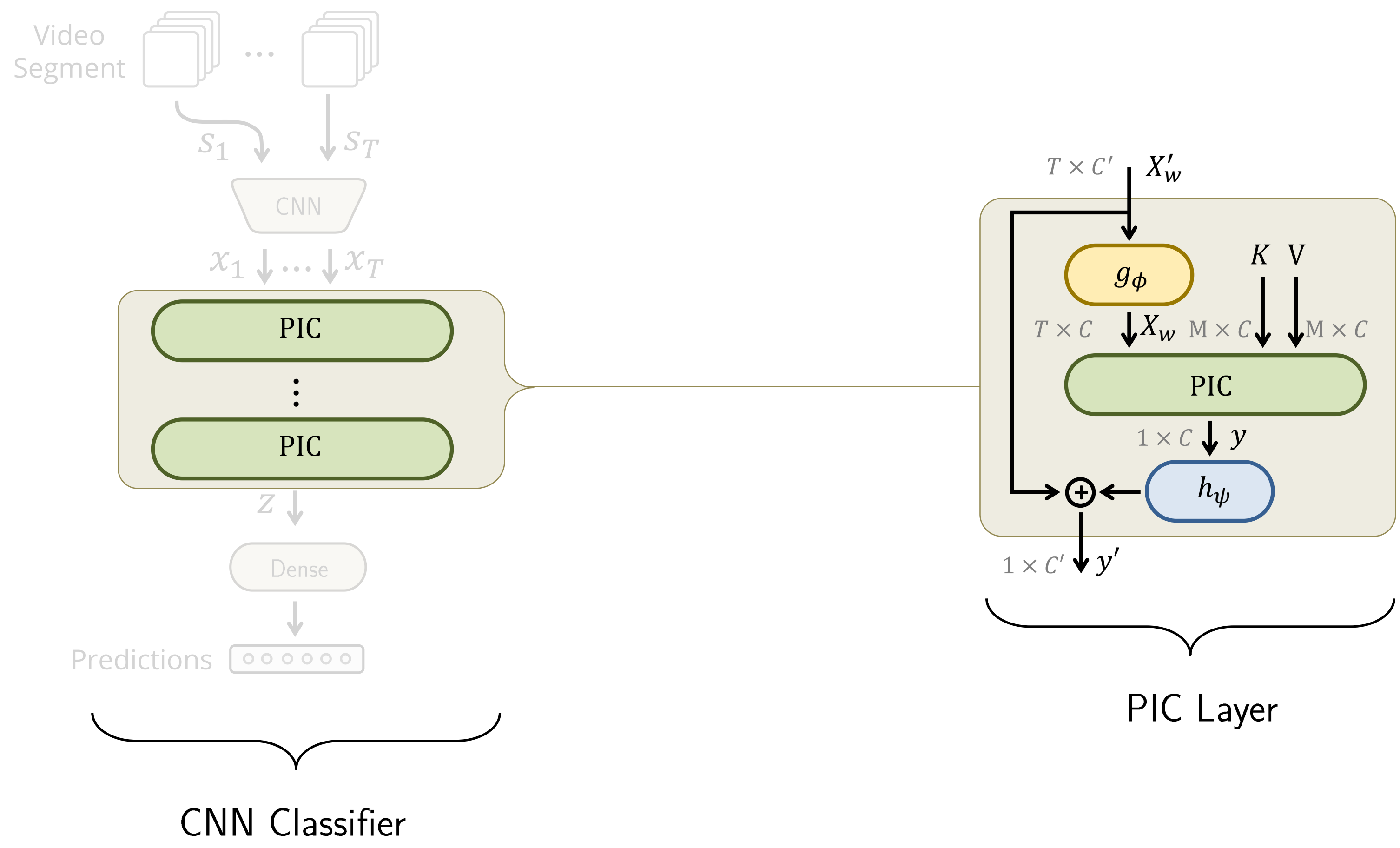
Solution

PIC Operation



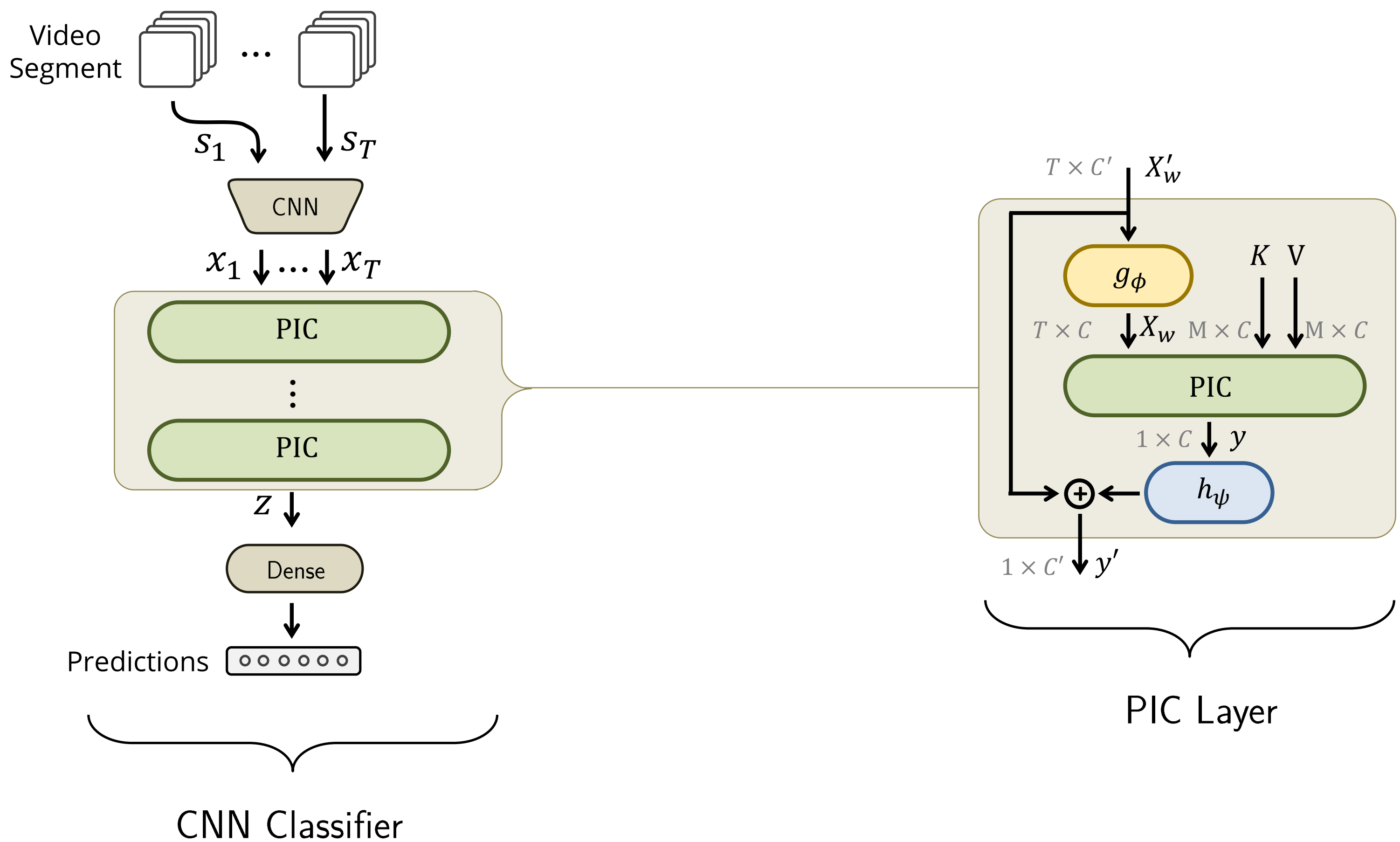
PIC Layer

Solution



Video
Segment
s_1
s_T
CNN
x_1 ... x_T
PIC
PIC
Z
Dense
Predictions
CNN Classifier
$T \times C'$
X'_w
g_ϕ
K
V
$T \times C$
X_w
$\mathrm{M} \times C$
$\mathrm{M} \times C$
PIC
$1 \times C$
y
h_ψ
$1 \times C'$
y'
PIC Layer

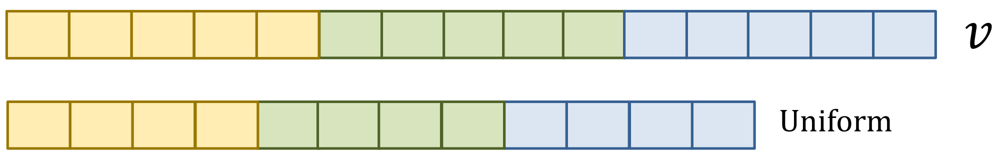
Solution



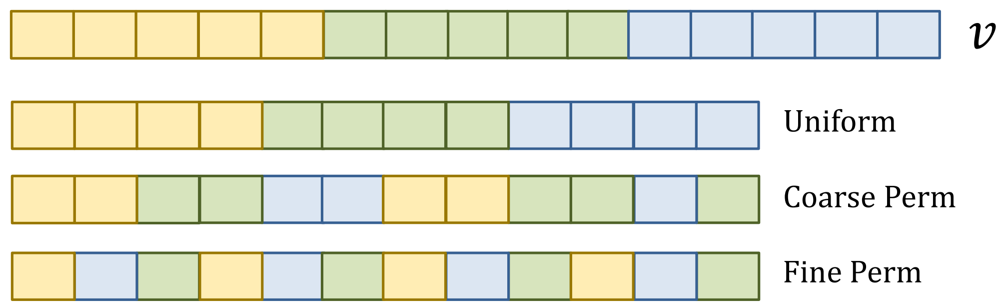
Video
Segment
s_1
s_T
CNN
x_1
x_T
PIC
PIC
Z
Dense
Predictions
CNN Classifier
$T \times C'$
X'_w
g_ϕ
K
V
$T \times C$
X_w
$\mathrm{M} \times C$
$\mathrm{M} \times C$
PIC
$1 \times C$
y
h_ψ
$1 \times C'$
y'
PIC Layer

Experiments

Permutation Invariance



Permutation Invariance



Permutation Invariance

Baseline	Coarse	Uniform	Drop ↓	Fine	Uniform	Drop ↓
Timeception	84.6	82.2	2.4	84.6	81.9	2.7
PIC-Ordered	80.2	77.6	2.6	80.2	76.3	3.9
PIC	87.5	87.0	**0.5**	87.5	86.7	**0.8**

Drop in Accuracy in Permuted Video Segments

Permutation Invariance

Baseline	Coarse	Uniform	Drop ↓	Fine	Uniform	Drop ↓
Timeception	84.6	82.2	2.4	84.6	81.9	2.7
PIC-Ordered	80.2	77.6	2.6	80.2	76.3	3.9
PIC	87.5	87.0	**0.5**	87.5	86.7	**0.8**

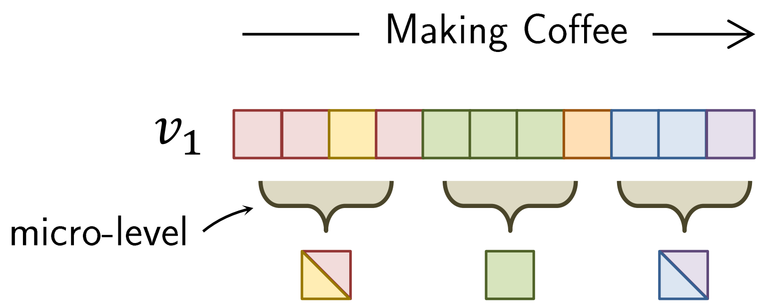
Drop in Accuracy in Permuted Video Segments

Permutation Invariance

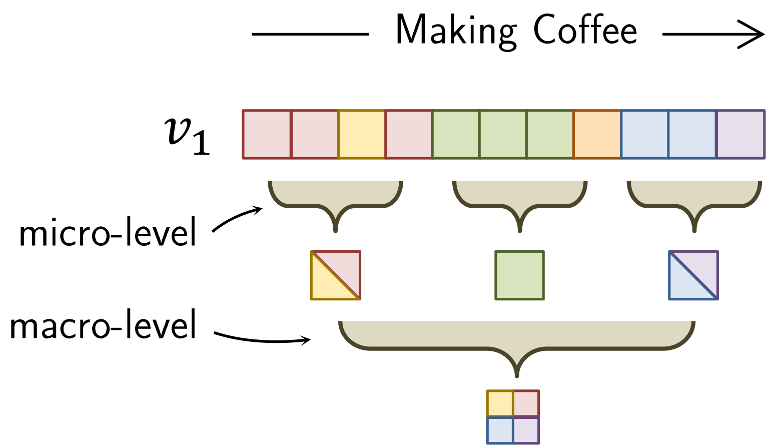
Baseline	Coarse	Uniform	Drop ↓	Fine	Uniform	Drop ↓
Timeception	84.6	82.2	2.4	84.6	81.9	2.7
PIC-Ordered	80.2	77.6	2.6	80.2	76.3	3.9
PIC	87.5	87.0	**0.5**	87.5	86.7	**0.8**

Drop in Accuracy in Permuted Video Segments

Cascaded Layers



Cascaded Layers



Cascaded Layers

Baseline	Accuracy (%) @ Layer			
	1	2	3	4
ActionVlad	83.07	----	----	----
Non-local	82.29	83.33	83.07	83.29
PIC-Global	86.76	85.68	85.68	85.42
Timeception	83.85	84.90	85.30	86.93
PIC	86.25	87.72	88.02	**89.84**

Increase in Accuracy with Cascaded Layers

Cascaded Layers

Baseline	Accuracy (%) @ Layer			
	1	2	3	4
ActionVlad	83.07	----	----	----
Non-local	82.29	83.33	83.07	83.29
PIC-Global	86.76	85.68	85.68	85.42
Timeception	83.85	84.90	85.30	86.93
PIC	86.25	87.72	88.02	**89.84**

Increase in Accuracy with Cascaded Layers

Cascaded Layers

Baseline	Accuracy (%) @ Layer			
	1	2	3	4
ActionVlad	83.07	----	----	----
Non-local	82.29	83.33	83.07	83.29
PIC-Global	86.76	85.68	85.68	85.42
Timeception	83.85	84.90	85.30	86.93
PIC	86.25	87.72	88.02	**89.84**

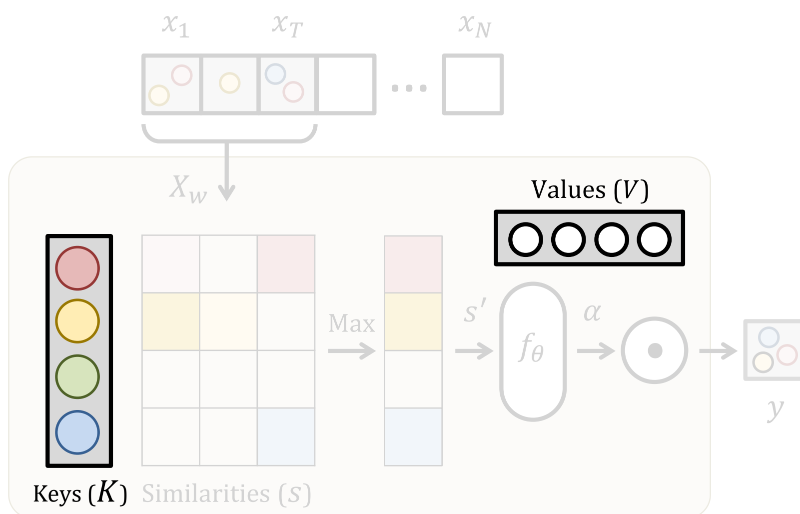
Increase in Accuracy with Cascaded Layers

Cascaded Layers

Baseline	Accuracy (%) @ Layer			
	1	2	3	4
ActionVlad	83.07	----	----	----
Non-local	82.29	83.33	83.07	83.29
PIC-Global	86.76	85.68	85.68	85.42
Timeception	83.85	84.90	85.30	86.93
PIC	86.25	87.72	88.02	**89.84**

Increase in Accuracy with Cascaded Layers

Key-Value Pair Kernels



Cascaded Layers

Baseline	Accuracy (%) @ Layer			
	1	2	3	4
Non-local	82.29	83.33	83.07	83.29
PIC-Inferred	82.55	83.85	84.90	84.64
PIC	86.27	87.72	88.02	**89.84**

Benefit of Shared v.s. Inferred Key-Value Kernels

Cascaded Layers

Baseline	Accuracy (%) @ Layer			
	1	2	3	4
Non-local	82.29	83.33	83.07	83.29
PIC-Inferred	82.55	83.85	84.90	84.64
PIC	86.27	87.72	88.02	**89.84**

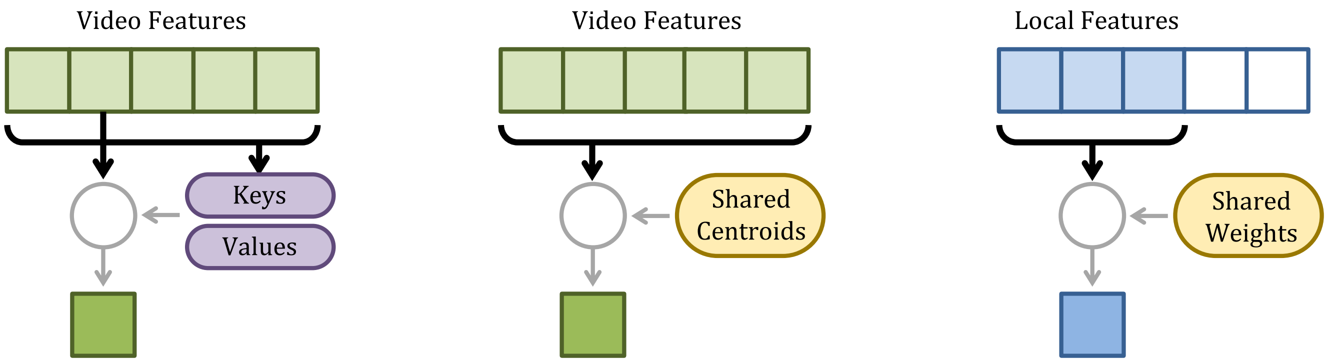
Benefit of Shared v.s. Inferred Key-Value Kernels

Cascaded Layers

Baseline	Accuracy (%) @ Layer			
	1	2	3	4
Non-local	82.29	83.33	83.07	83.29
PIC-Inferred	82.55	83.85	84.90	84.64
PIC	86.27	87.72	88.02	**89.84**

Benefit of Shared v.s. Inferred Key-Value Kernels

Related Works



(a) Self-Attention (b) Vector Aggregation (c) Convolution

Related Works

Method	Backbone	mAP (%)
SlowFast [47]	—	42.1
SlowFast-NL* [47]	—	42.5
3D CNN [6]	R101	35.5
3D CNN + TC [2]	R101	41.1
3D CNN + **PIC**	R101	**42.7**
3D CNN [7]	R101-NL	41.0
3D CNN + FB [7]	R101-NL	42.5
3D CNN + **PIC**	R101-NL	**43.8**

Charades

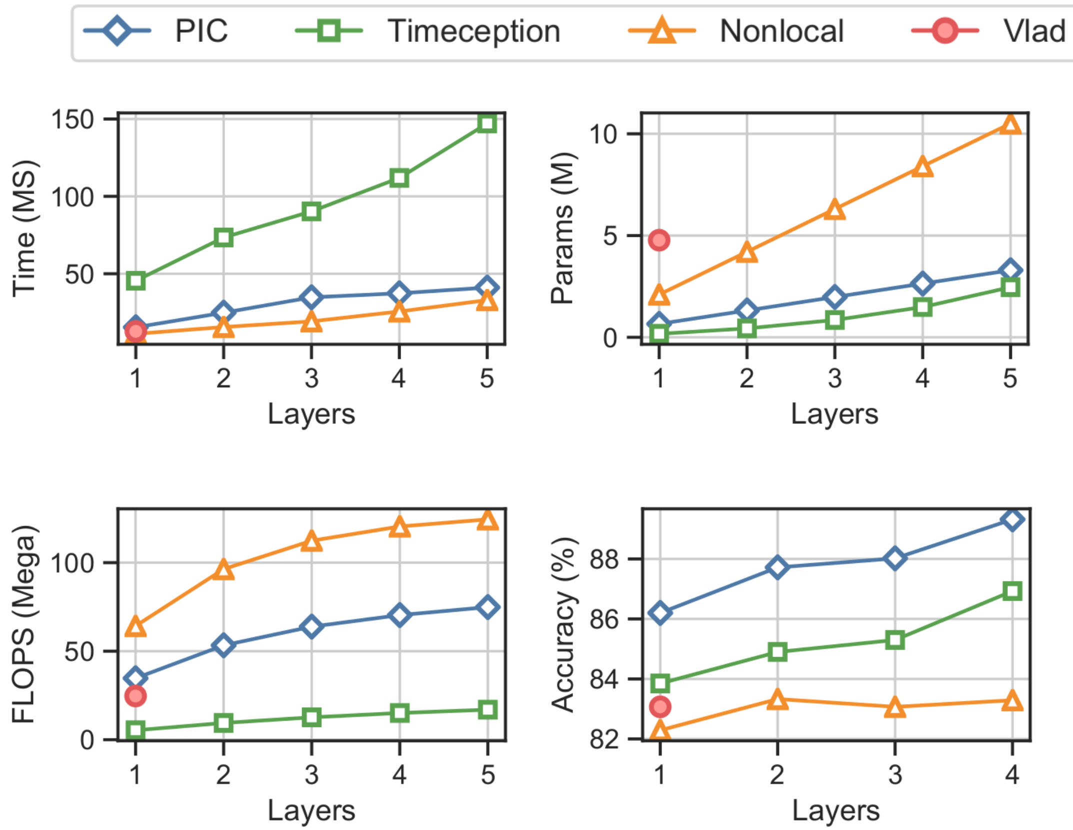
Method	Backbone	Accuracy (%)
3D CNN	I3D	80.64
3D CNN + Vlad [46]	I3D	82.67
3D CNN + Nonlocal [6]	I3D	83.79
3D CNN + Timeception [2]	I3D	86.93
3D CNN + **PIC**	I3D	**89.84**

Breakfast

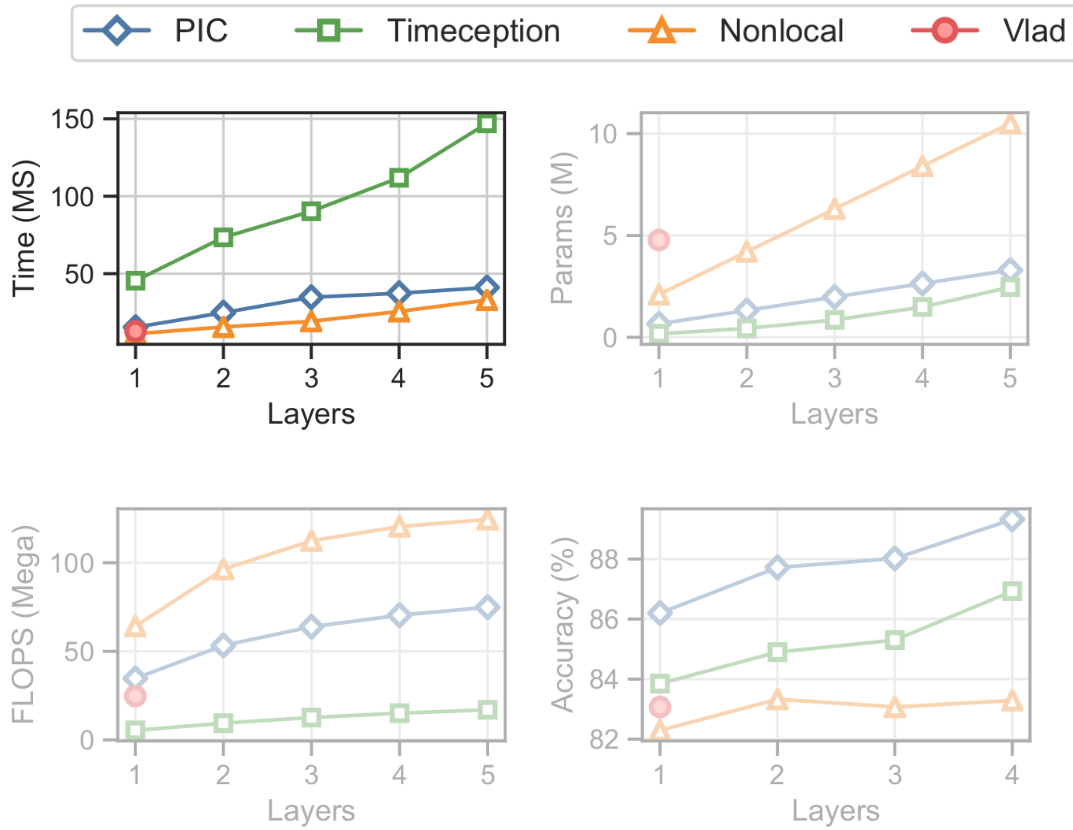
Method	Backbone	mAP (%)
3D CNN	I3D	72.43
3D CNN + Timeception [2]	I3D	74.79
3D CNN + **PIC**	I3D	**78.31**

Multi-Thumos

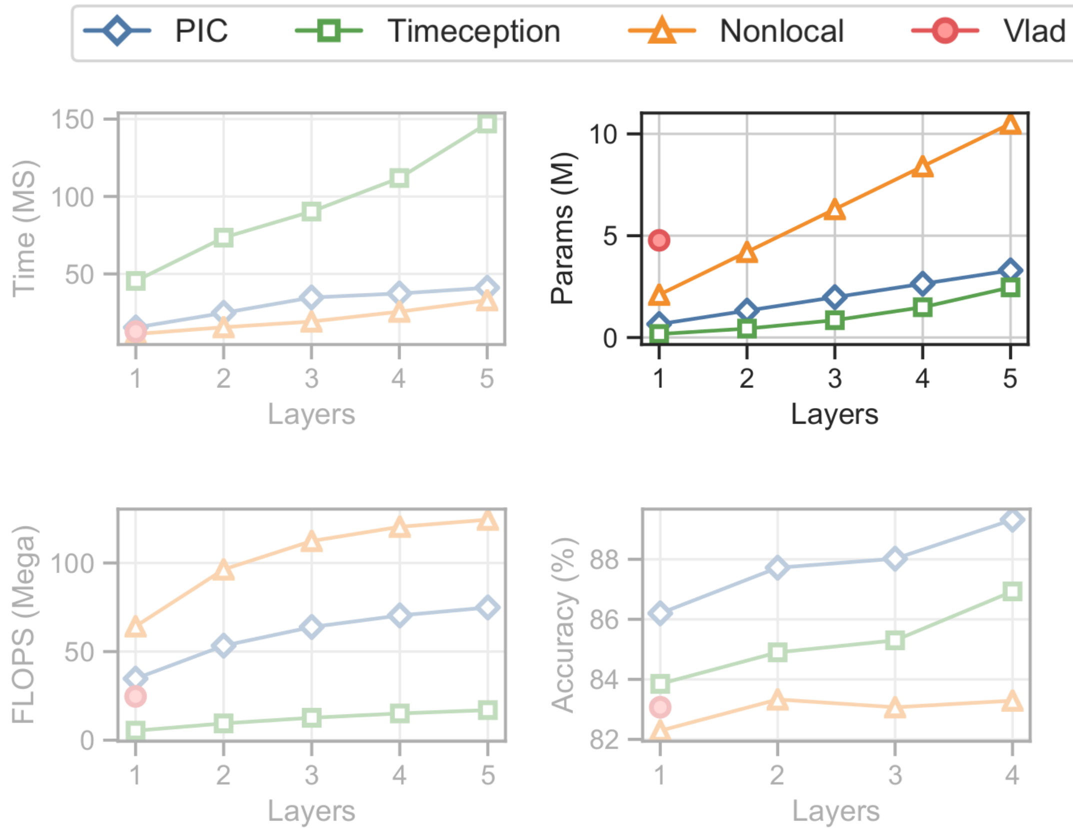
Performance Analysis



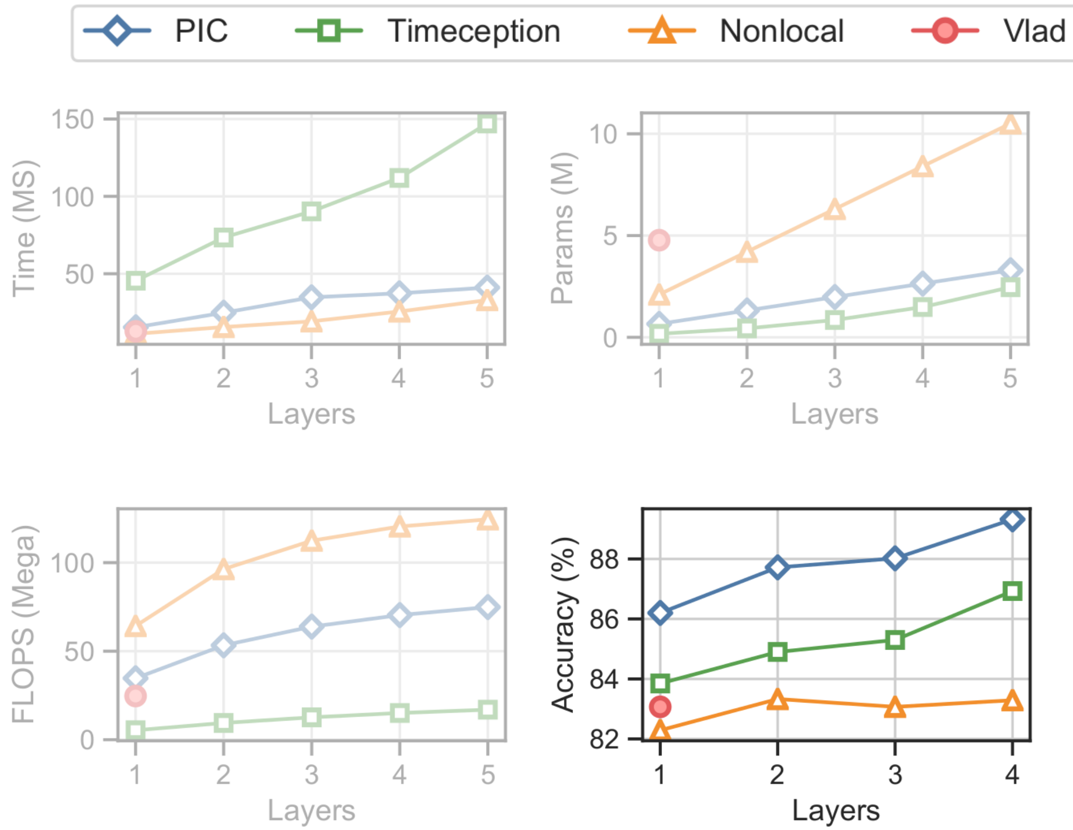
Performance Analysis



Performance Analysis



Performance Analysis



Learned Concepts

A05
A01

Food Box (A05 Milk, A01 Cereals)

A02
A10

Pouring (A02 Coffee, A10 Tea)

A07
A08

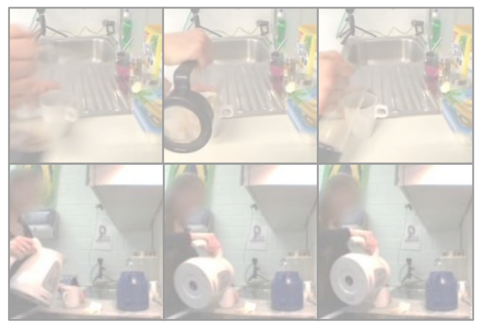
Cutting (A07 Salad, A08 Sandwich)

Learned Concepts

A05
A01

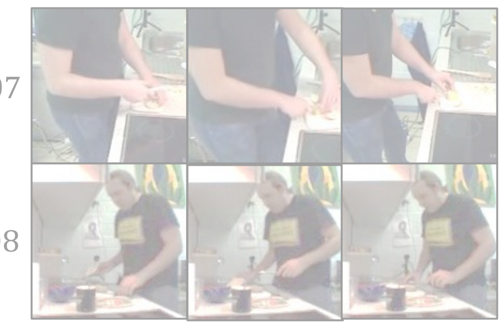
Food Box (A05 Milk, A01 Cereals)

A02
A10



Pouring (A02 Coffee, A10 Tea)

A07
A08



Cutting (A07 Salad, A08 Sandwich)

Object-centric

Learned Concepts

A05
A01

Food Box (A05 Milk, A01 Cereals)

A02
A10

Pouring (A02 Coffee, A10 Tea)

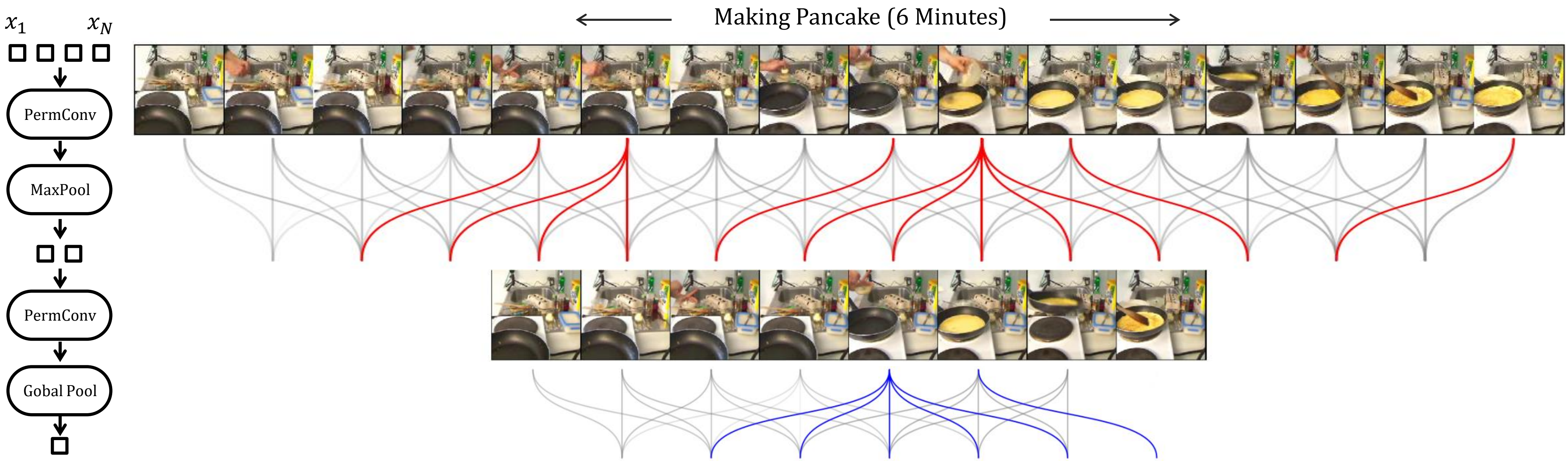
A07
A08

Cutting (A07 Salad, A08 Sandwich)

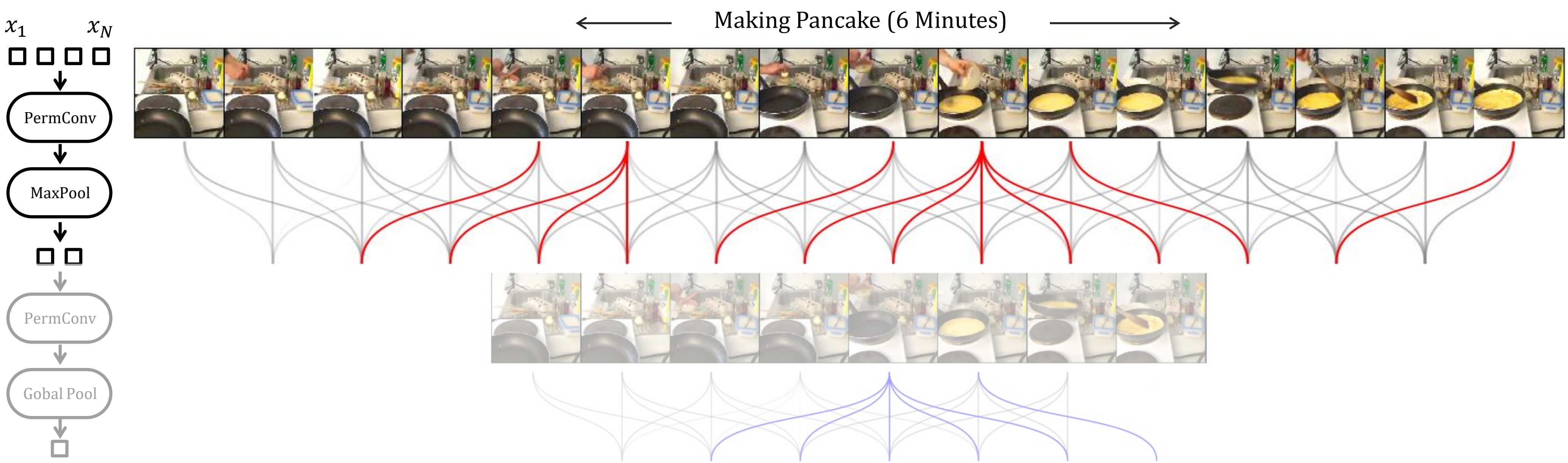
Object-centric

Action-centric

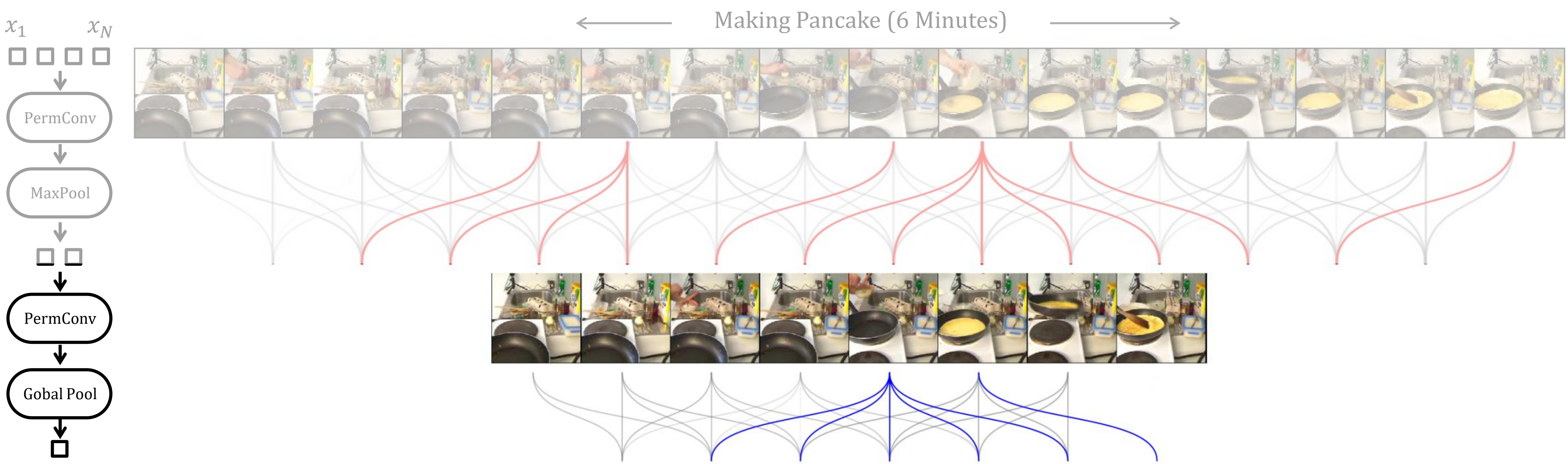
Learned Attention



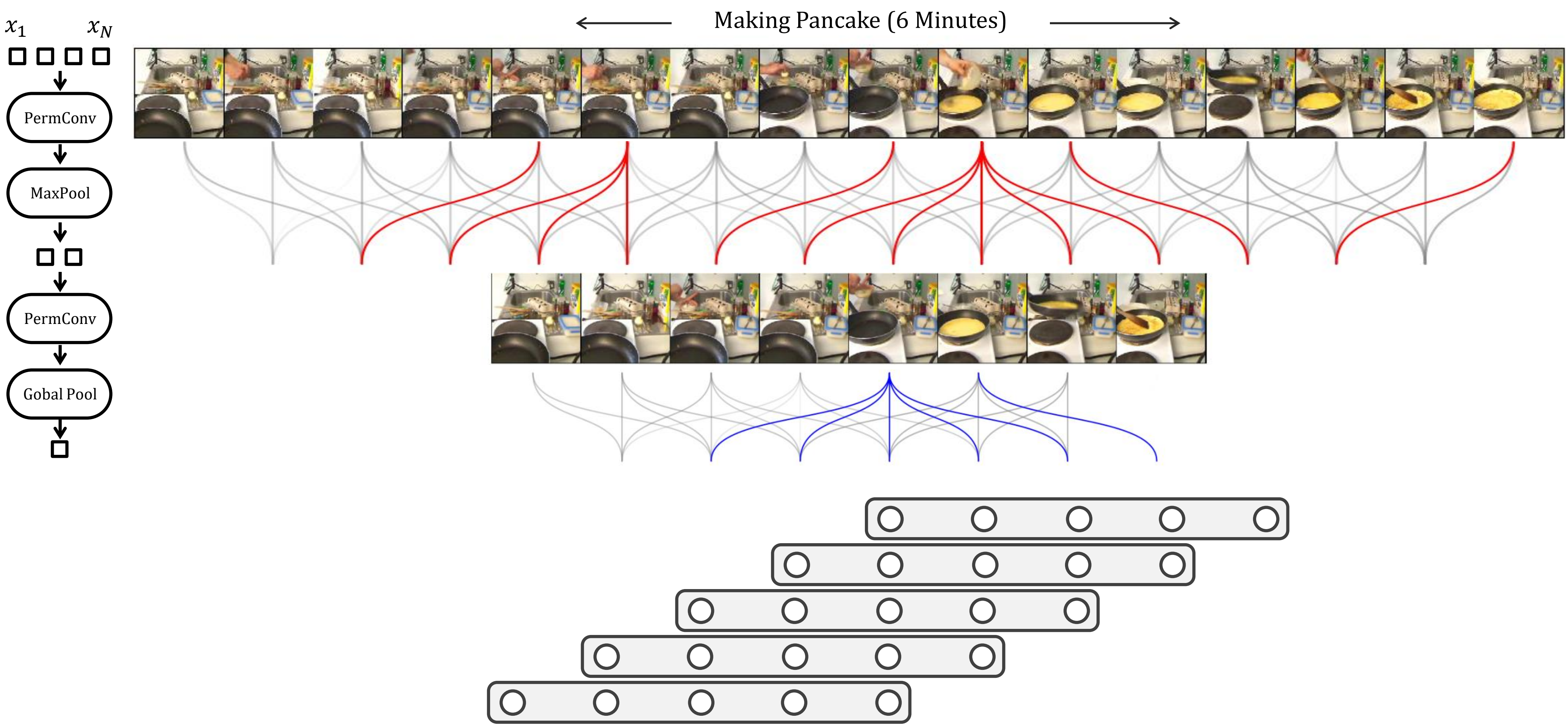
Learned Attention



Learned Attention

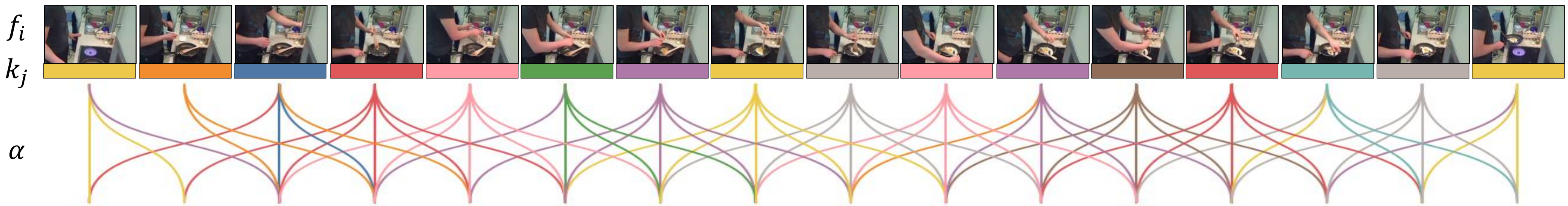


Learned Attention



Learned Attention

Fried Egg



Thank You!

questions