Self-supervised learning (for videos)

Computer Vision (SJK02)

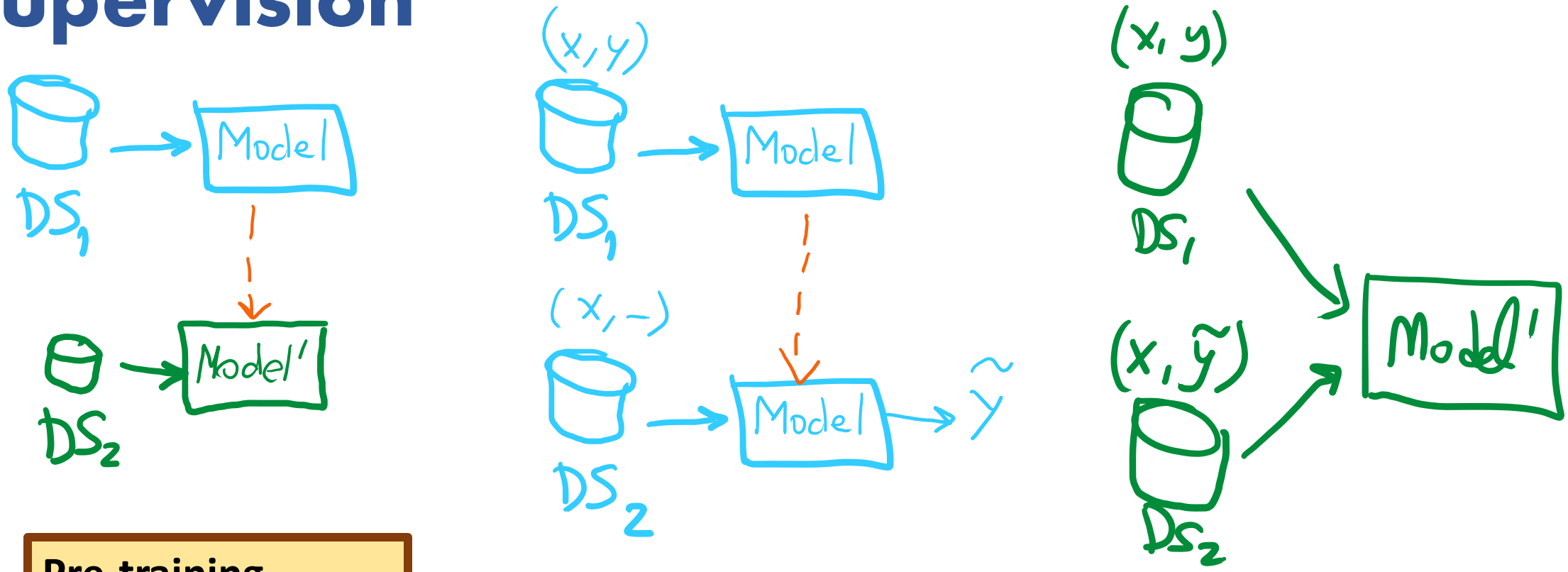
Universitat Jaume I

The issues with labelling

Annotation cost
Annotation bias

(Lack of domain generalisation)
(Lack of robustness)

Pre-training ≠ self-training ≠ self-supervision



Pre-training
Transfer learning
Fine-tuning

Self-training
Labeled + Pseudo-labeled datasets

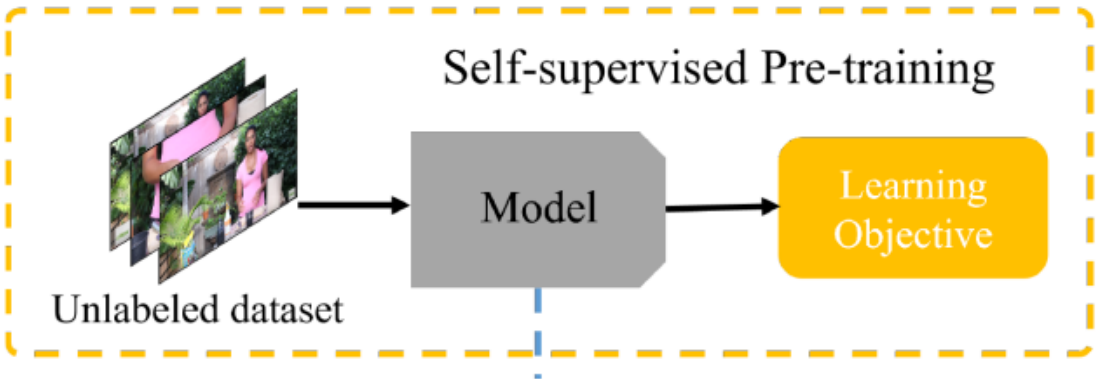
[Rethinking pre-training and self-training]

Self-supervised learning (SSL) in a nutshell

Alternative to pretraining a model

- Large dataset (without labels!)
- Results in higher generalisation

Self-supervised generalization for videos: a survey (2023)


Self-supervised Pre-training
Model
Learning Objective
Unlabeled dataset

Downstream tasks

Action **recognition**

Temporal action **segmentation**

Temporal Action **Step Localization**

Video **retrieval**

Text-to-Video Retrieval

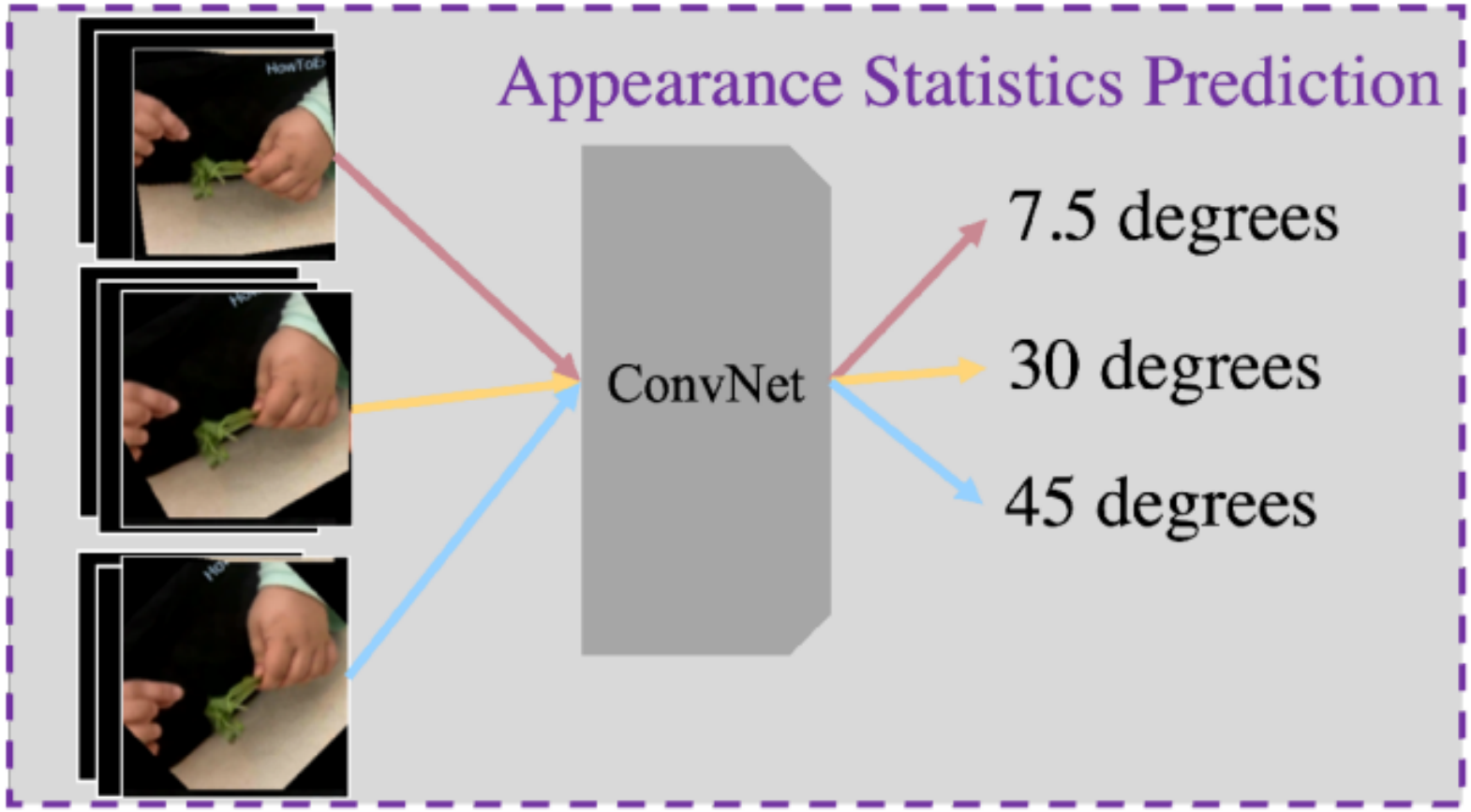
Video **Captioning**

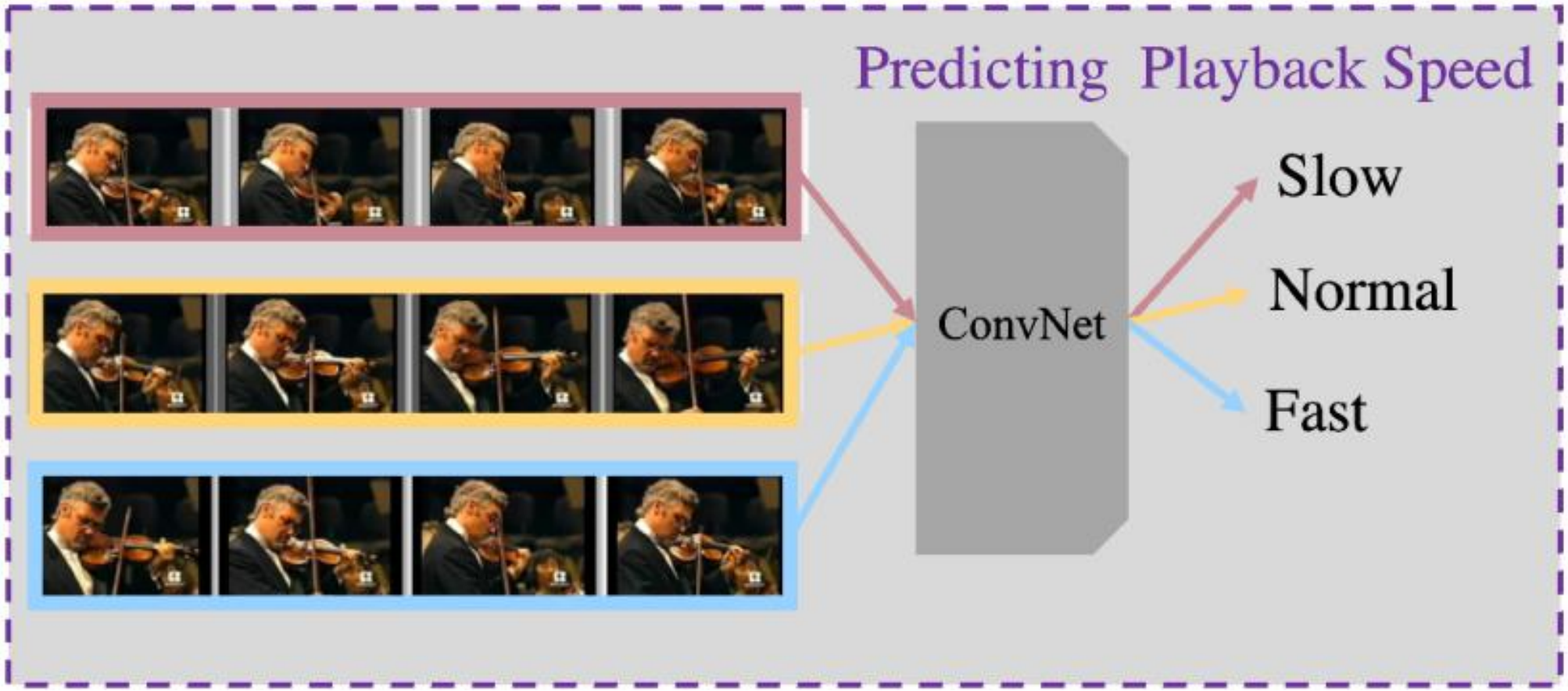
Approaches

Pretext
Generative
Contrastive
Multimodal*

(*) Not seen here

Pretext tasks





Why is this approach more specific to videos?

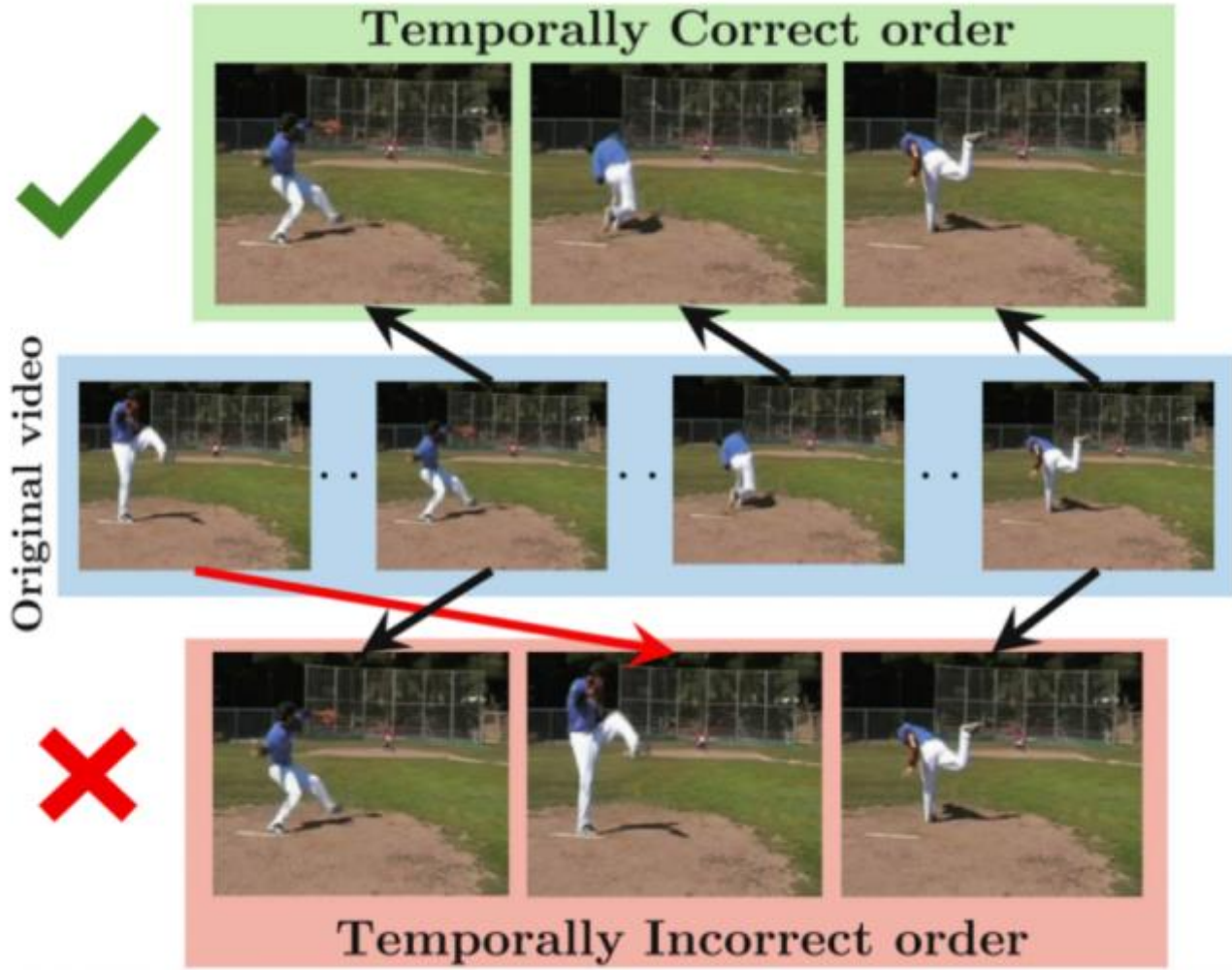
Other pretext task: *your turn*

Think of other pretext task for SSL for videos

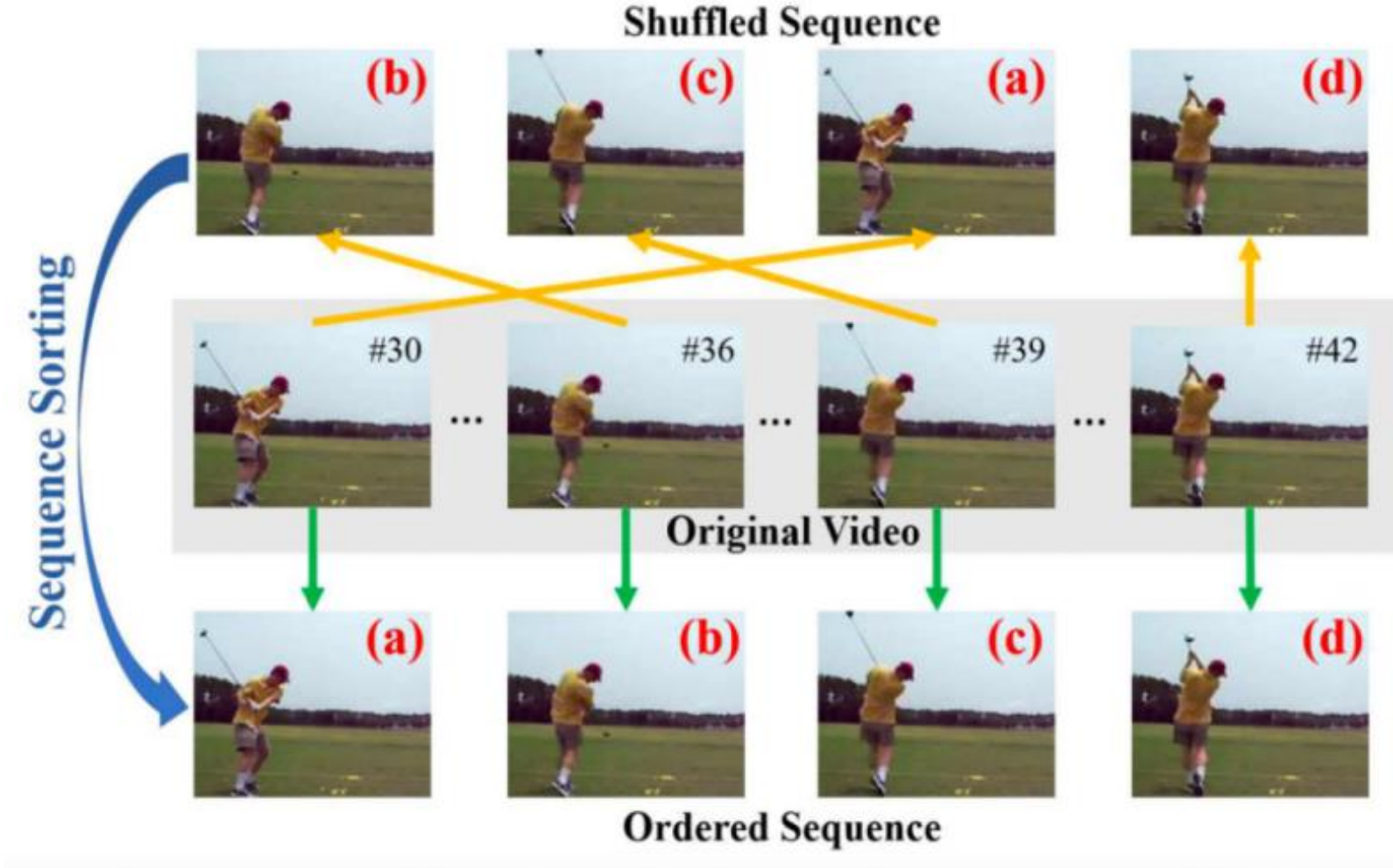
www.socrative.com
Room 219986

Main hypothesis of SSL

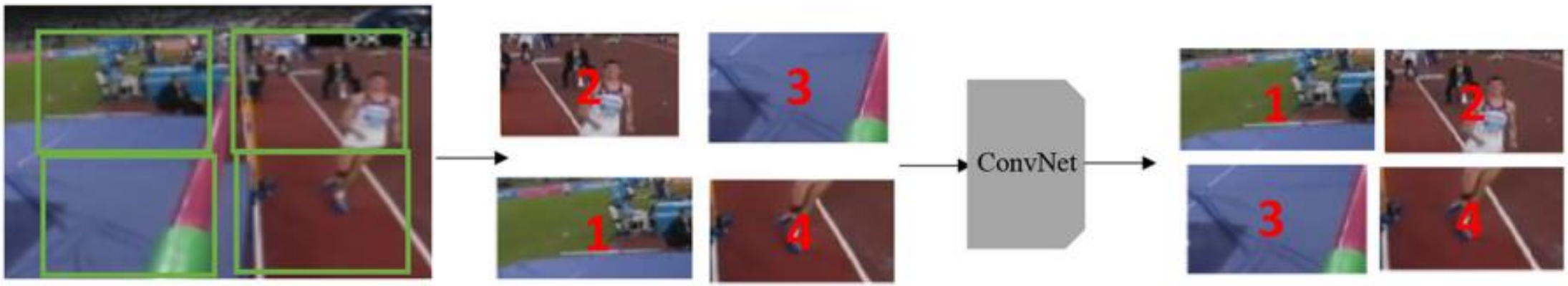
True or false?
The idea of SSL is that if the model can solve a complicated task that requires high-level understanding of the input, then it will learn more generalisable features

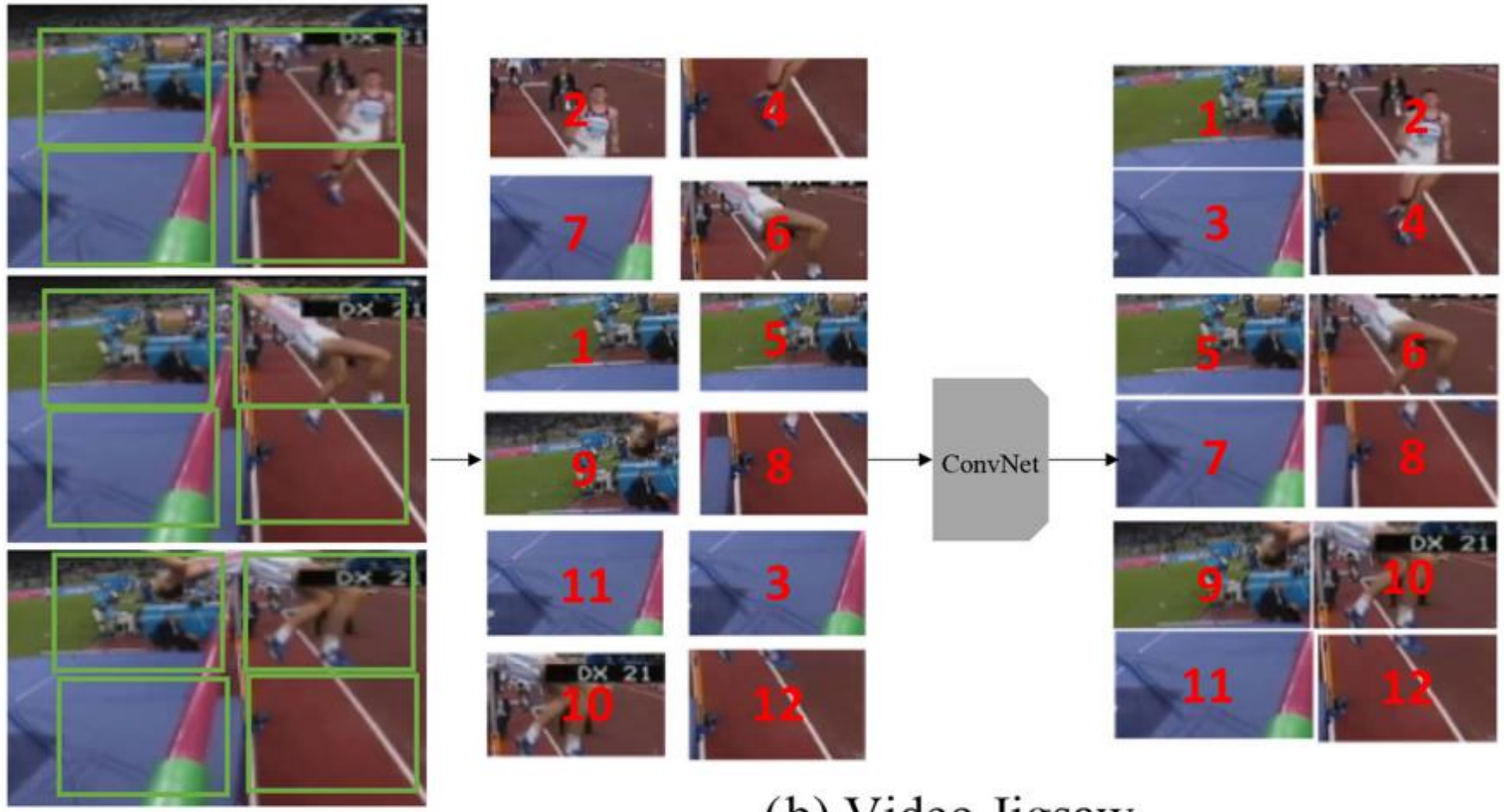


(a) Binary Classification Task



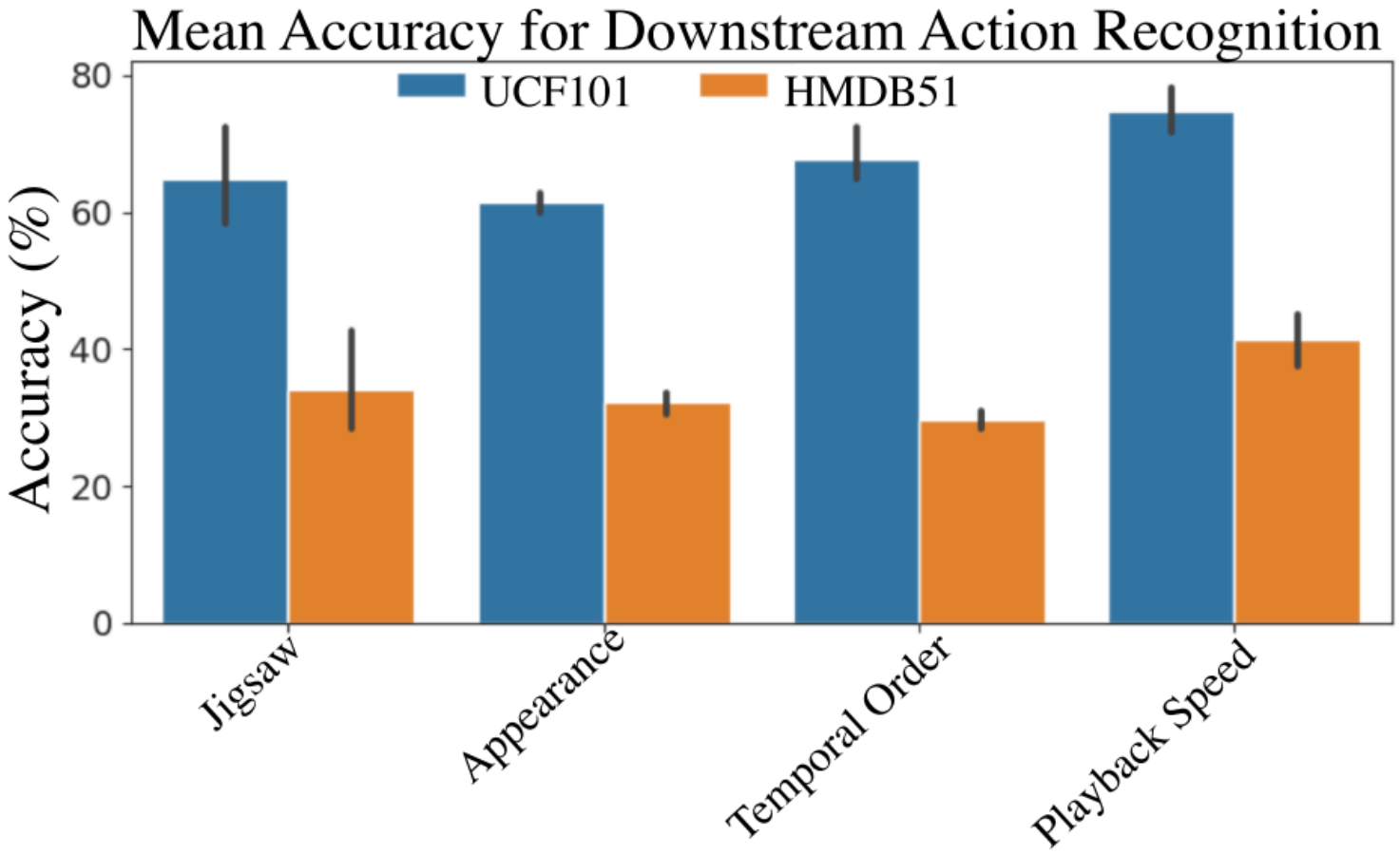
(b) Sequence Sorting


2
3
1
4
ConvNet
1
2
3
4



(b) Video Jigsaw

Do different pretext tasks make any difference?

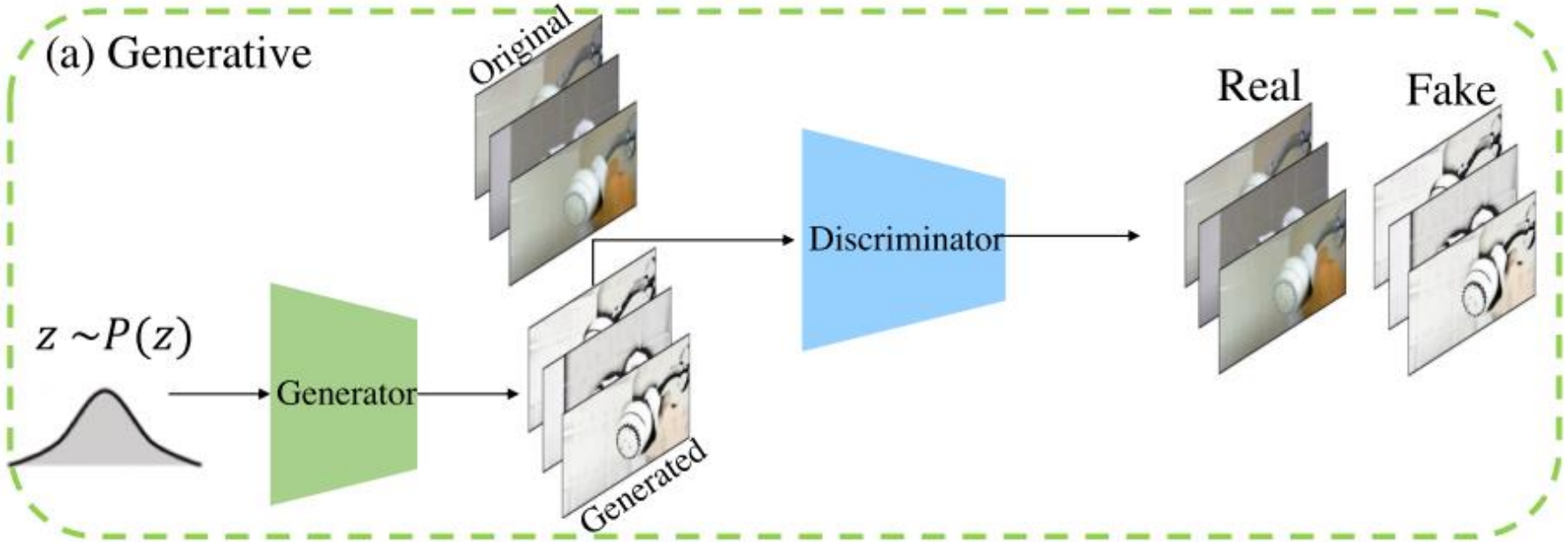


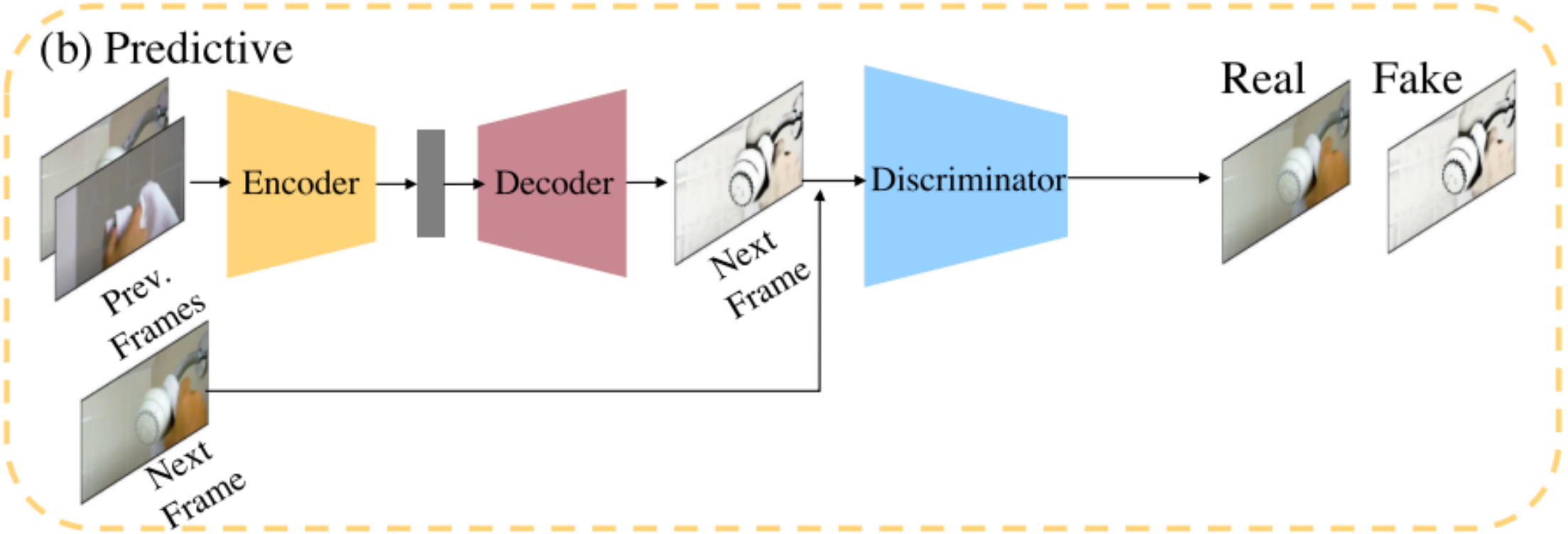
Are time-related pretext tasks more helpful?

Generative approaches

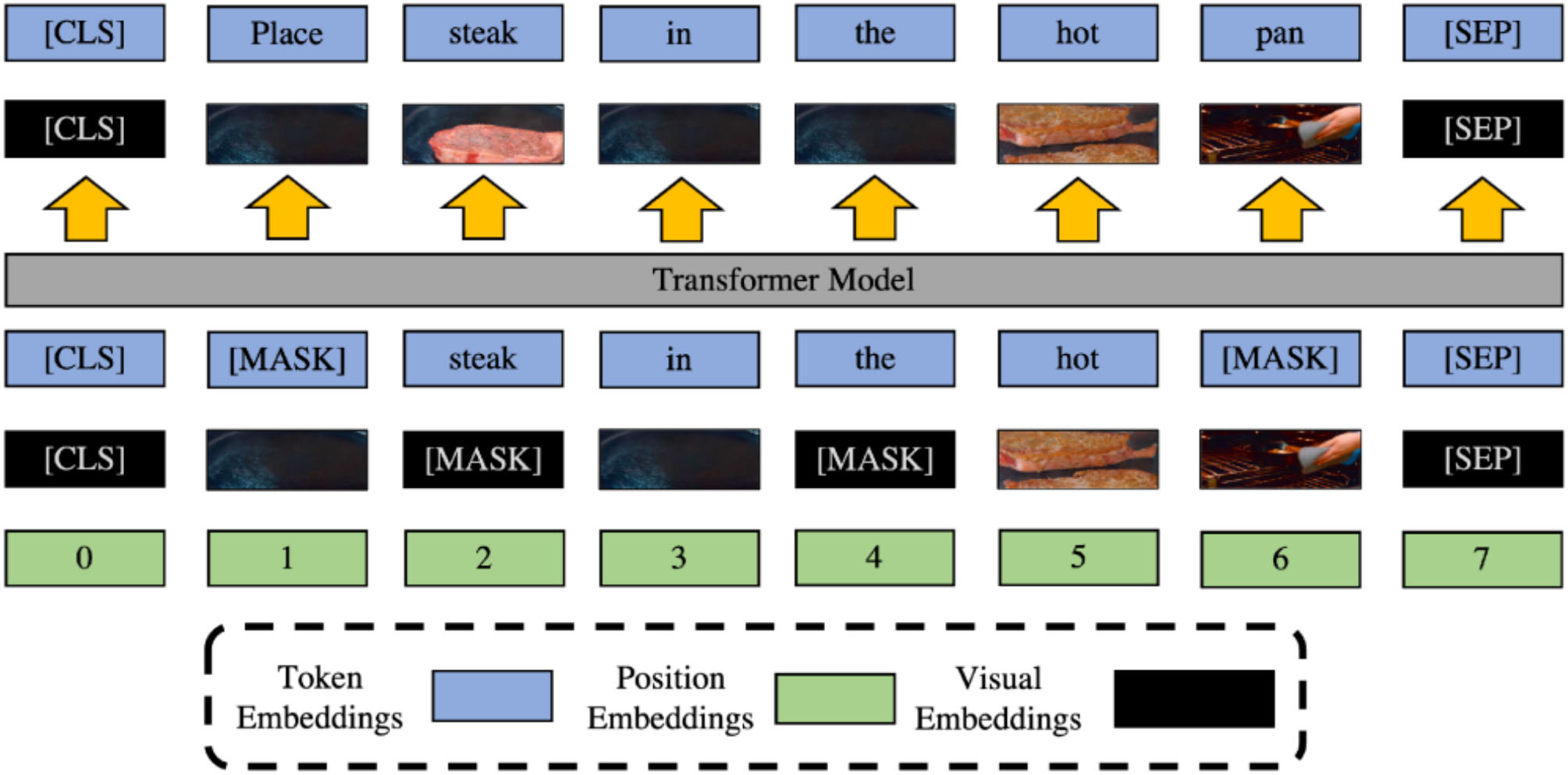
GANs (Generative Adversarial Networks)
Masked Autoencoders (MAEs)

Predicting next frame



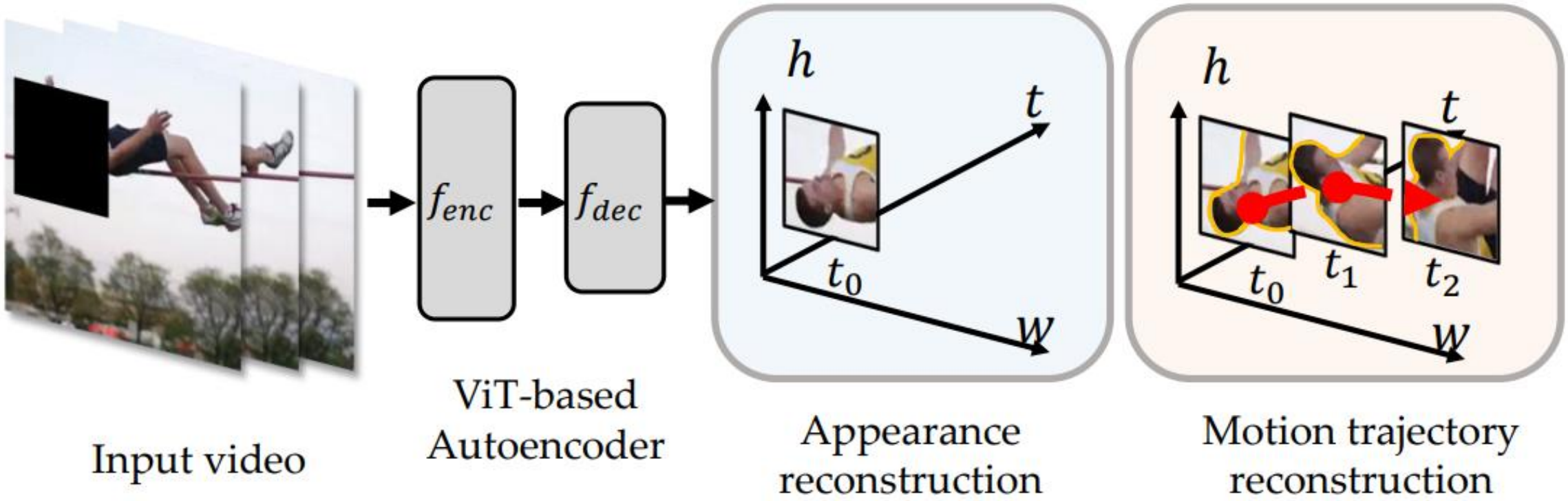


Masked modelling



(a) Two key factors to recognize a *high jump* action.

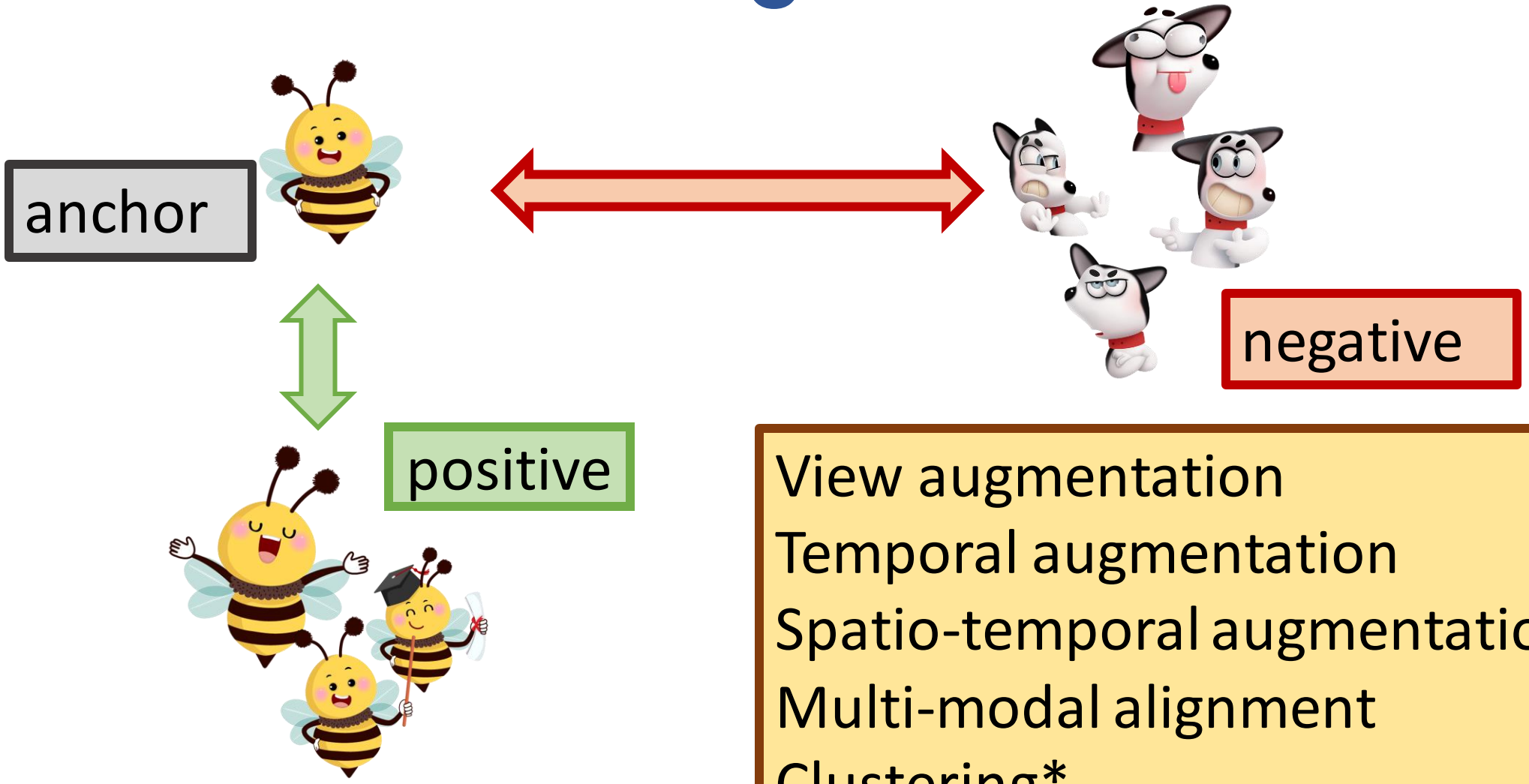
[Masked Motion Encoding for Self-Supervised Video Representation Learning](#) (CVPR 2023)



(b) Appearance reconstruction vs. motion trajectory reconstruction.

Masked Motion Encoding for Self-Supervised Video Representation Learning (CVPR 2023)

Contrastive learning



View augmentation
Temporal augmentation
Spatio-temporal augmentation
Multi-modal alignment
Clustering*

(*) Not seen here

Contrastive loss

$$\mathbf{L} = (\mathbf{1} - \mathbf{Y}) * ||\mathbf{x_i} - \mathbf{x_j}||^{\mathbf{2}} + \mathbf{Y} * \mathbf{max}(\mathbf{0}, \mathbf{m} - ||\mathbf{x_i} - \mathbf{x_j}||^{\mathbf{2}})$$

Y=0 if x_i and x_j have the same labels (and 1 otherwise)

Triplet loss

$$\mathbf{L} = \mathbf{max}(\mathbf{0}, ||\mathbf{x} - \mathbf{x}^{+}||^{\mathbf{2}} - ||\mathbf{x} - \mathbf{x}^{-}||^{\mathbf{2}} + \mathbf{m})$$

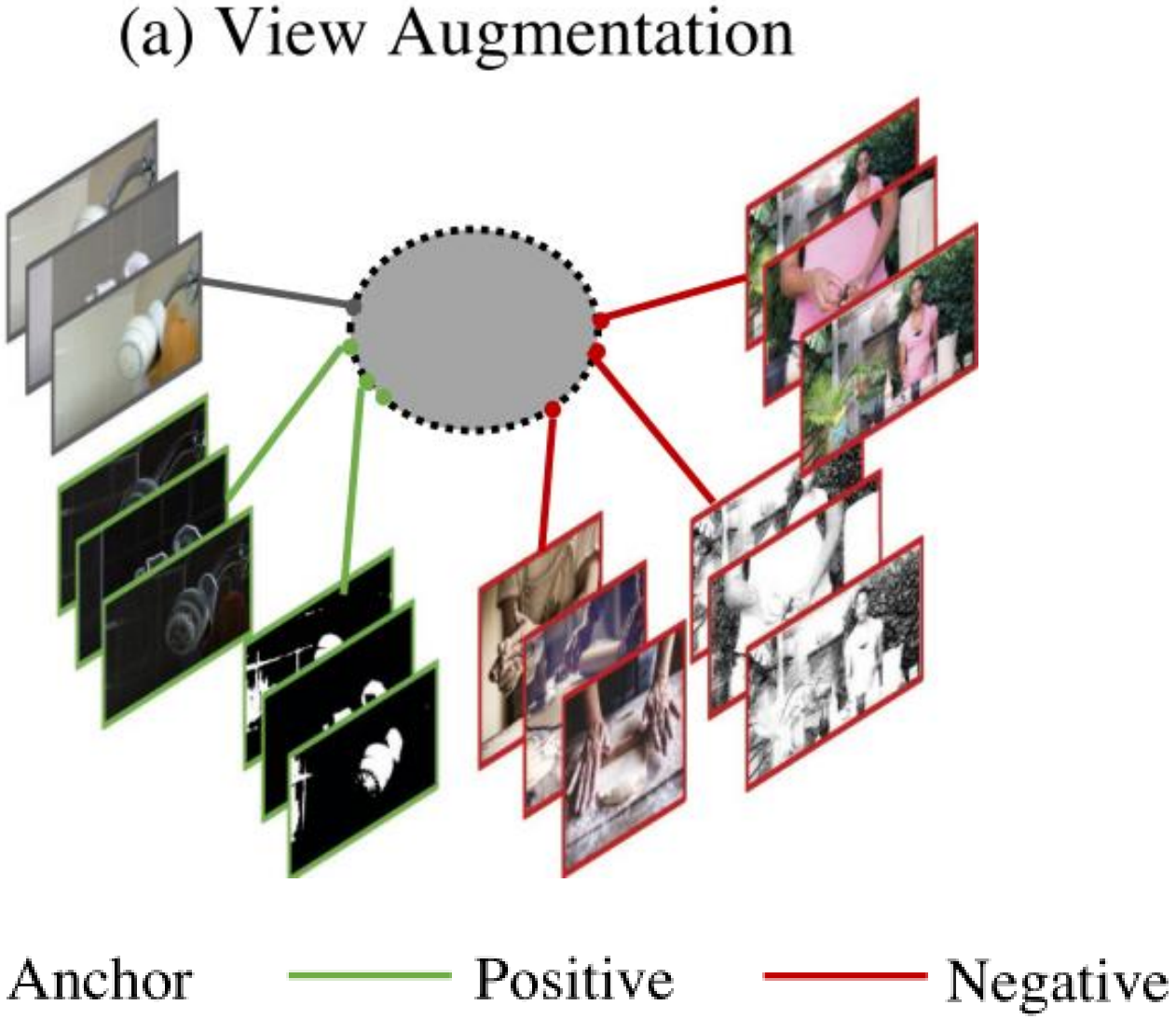
Do the contrastive loss and the triplet loss require the same data at the same time?

Noise Contrastive Estimation Loss (NCE)

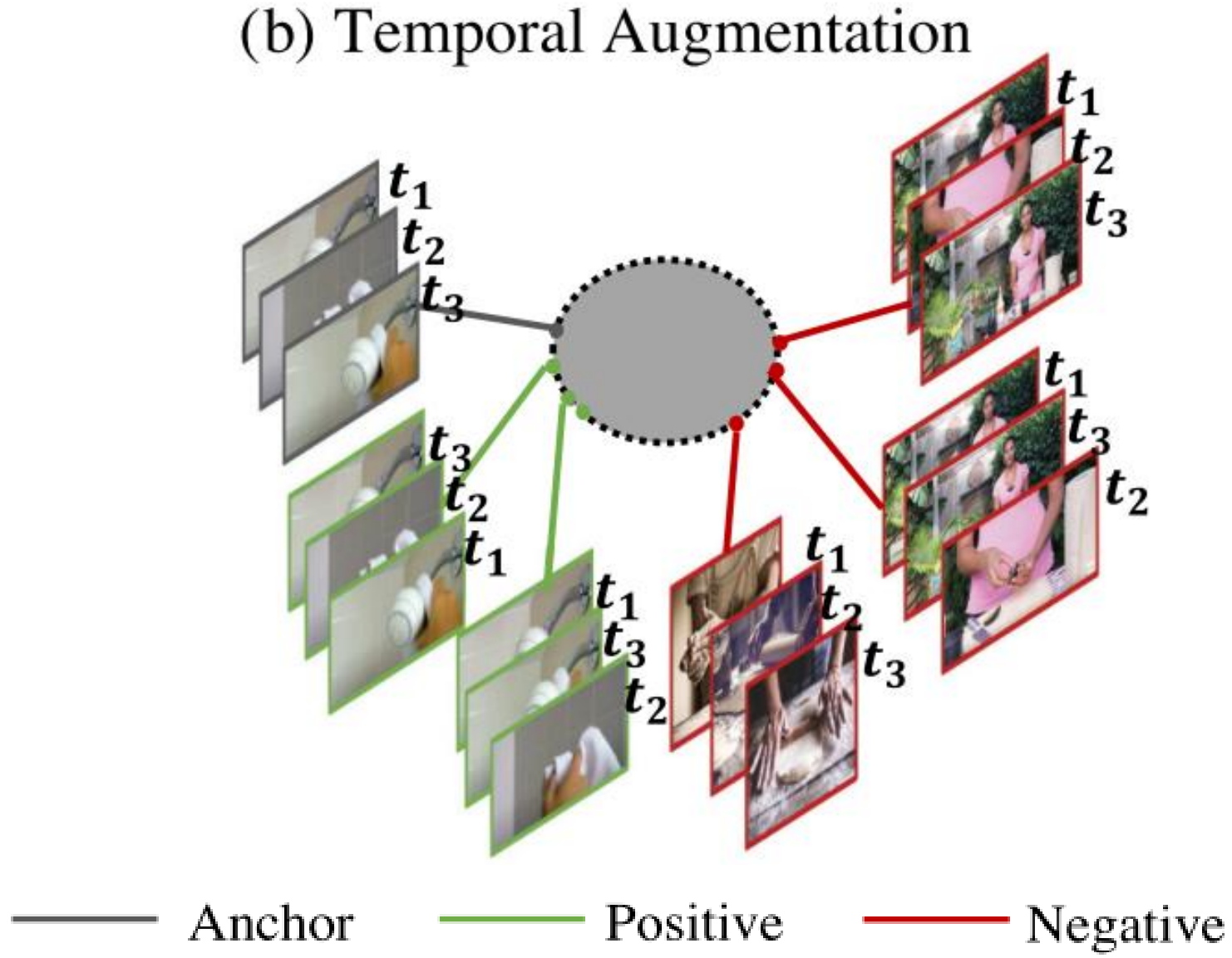
Uses a pair of positive and a set of negative examples

https://www.baeldung.com/cs/contrastive-learning

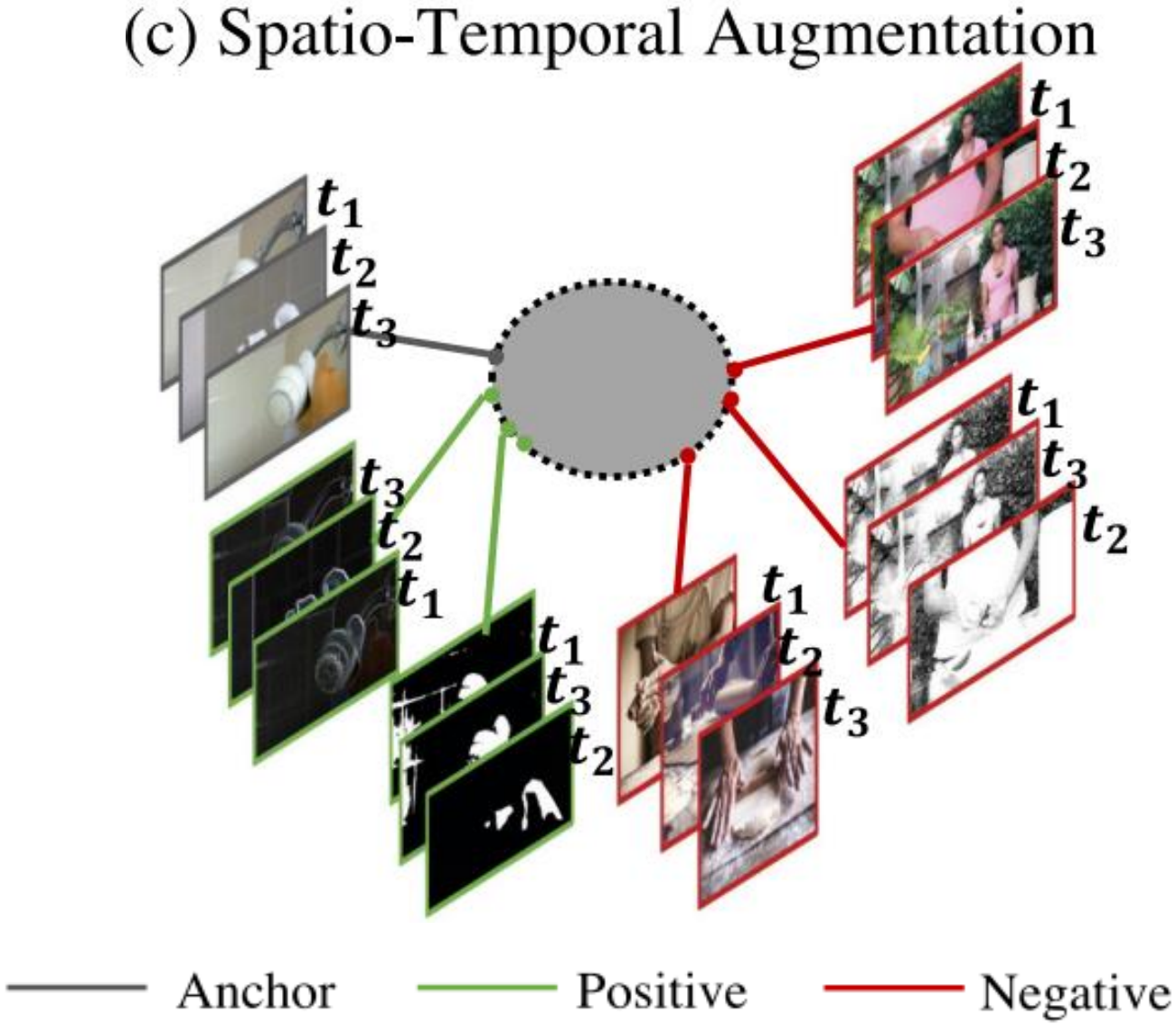
View augmentation



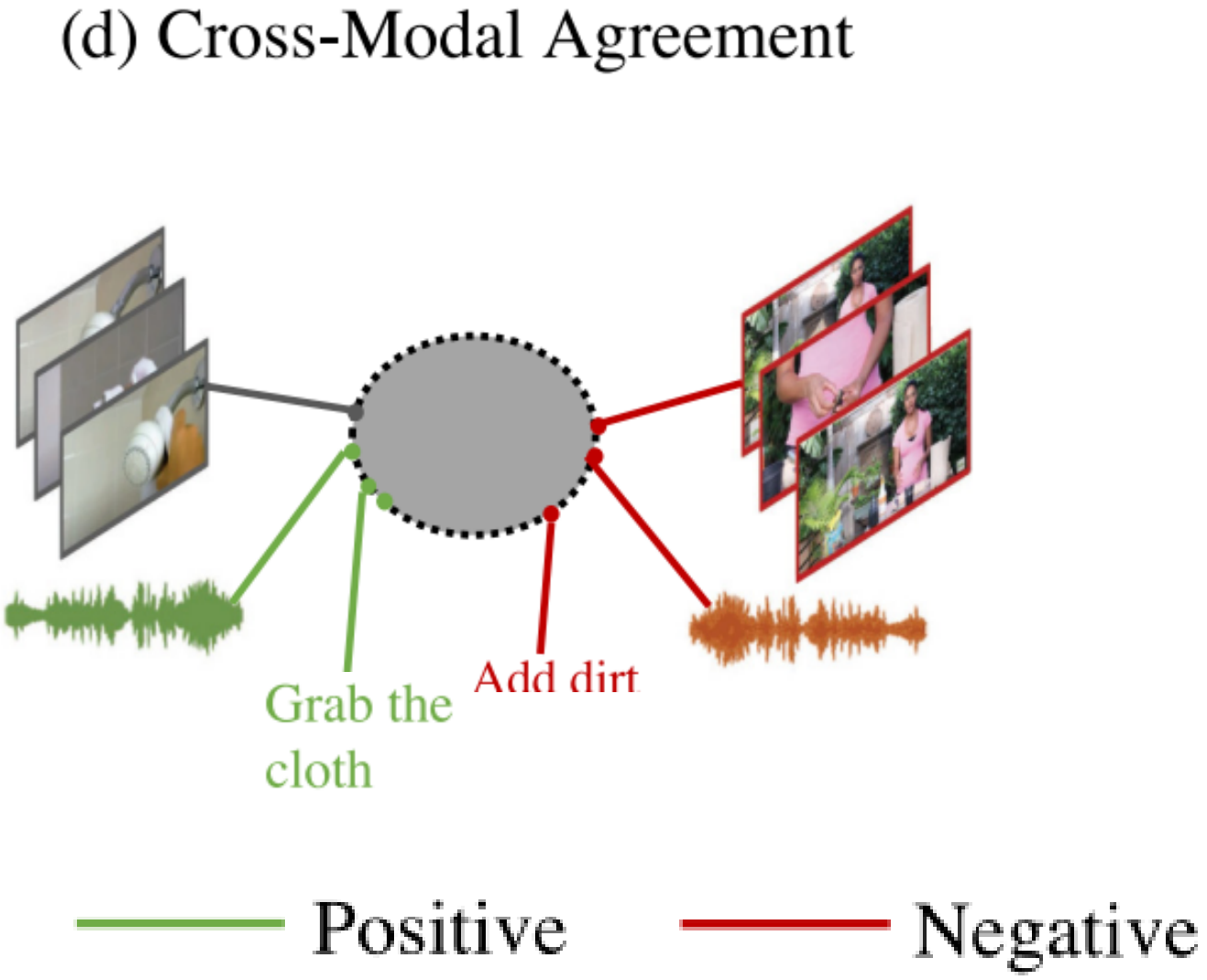
Temporal augmentation



Spatio-temporal augmentation



Multimodal


(d) Cross-Modal Agreement
Grab the cloth
Add dirt
Anchor
Positive
Negative

Comparison of constrastive approaches