tenor.com

# Test of Time: Instilling Video-Language Models with a Sense of Time

Piyush Bagad

University of Amsterdam

Dr. Makarand Tapaswi

IIIT Hyderabad

Prof. Cees Snoek

University of Amsterdam

bpiyush.github.io/testoftime-website/

Paper tag: TUE-AM-239

UNIVERSITY OF AMSTERDAM

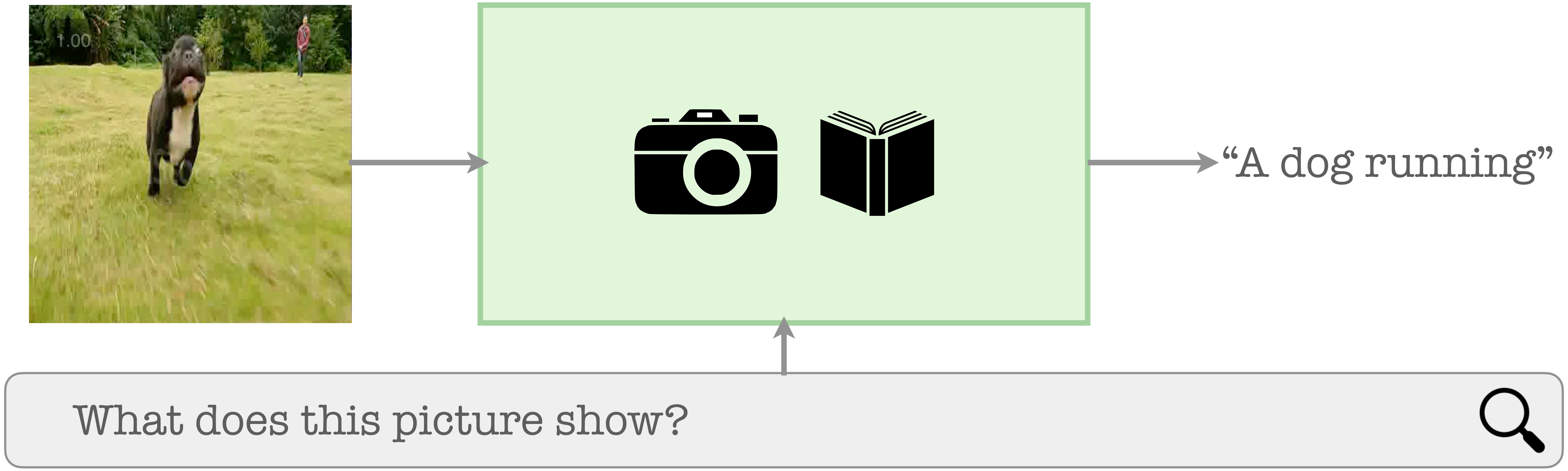
VIS LAB VIDEO & IMAGE SENSE LAB

ellis unit | AMSTERDAM

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

# Quick Preview: The problem

- Foundation models: Language interface + a few (or no) training samples

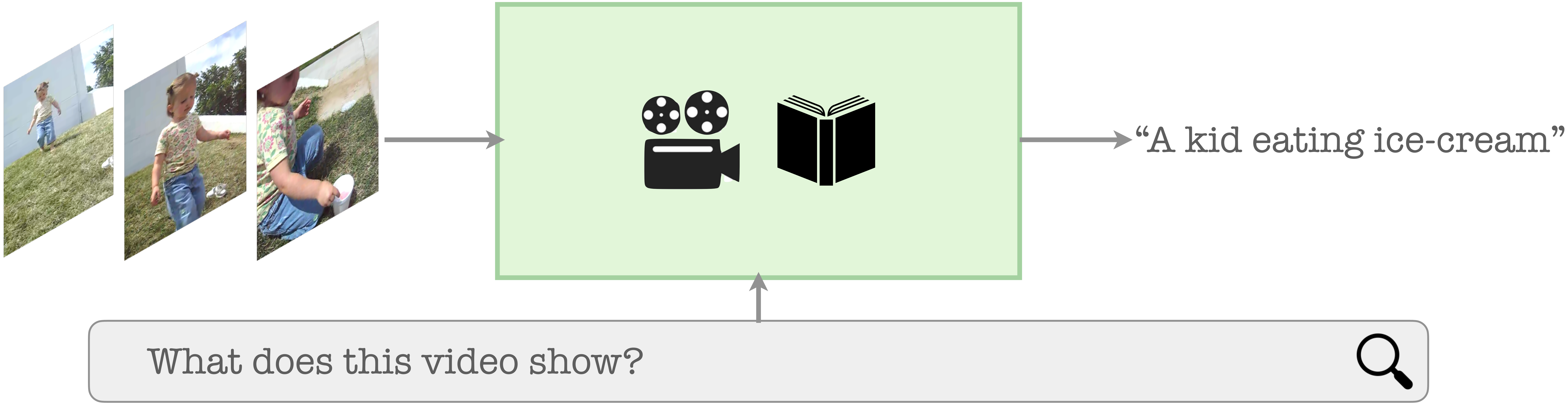


“A dog running”

What does this picture show?

# Quick Preview: The problem

- Foundation models: Language interface + a few (or no) training samples
- Particularly attractive for videos given high cost

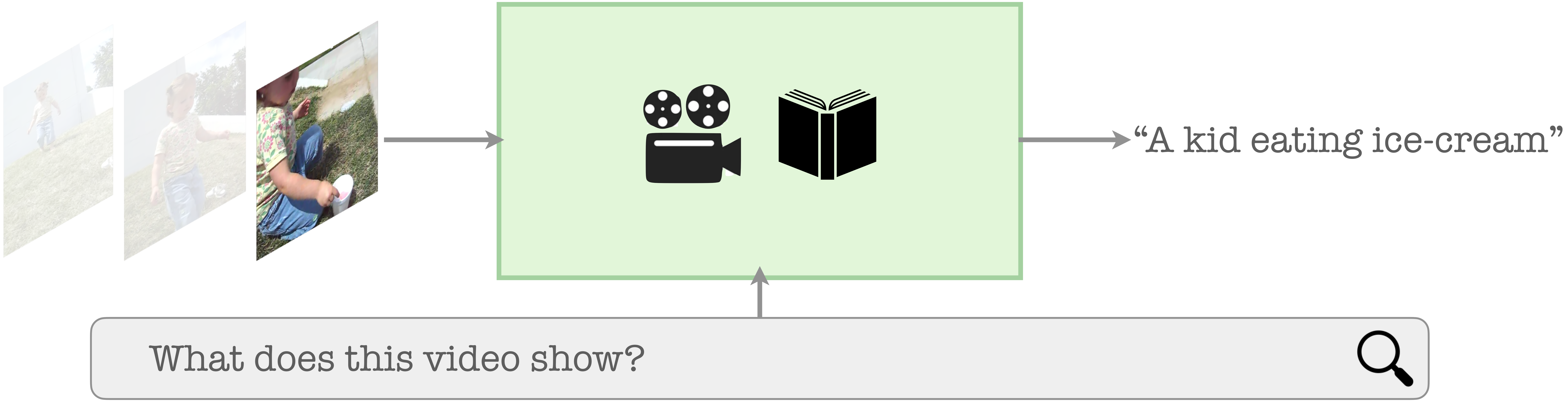


“A kid eating ice-cream”

What does this video show?

# Quick Preview: The problem

- Do video foundation models truly understand time?

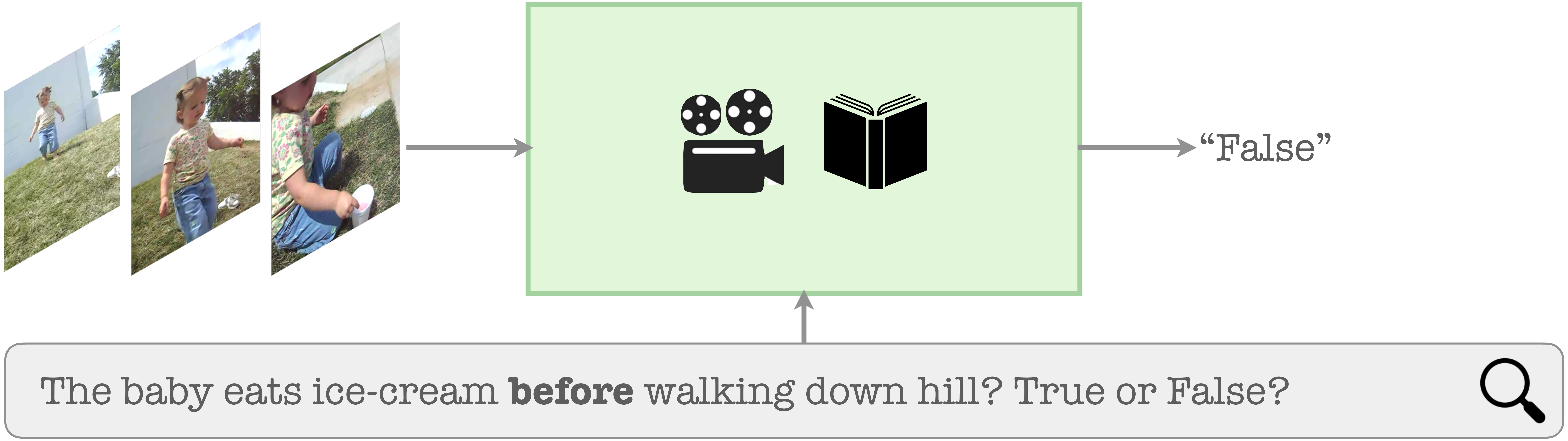


“A kid eating ice-cream”

What does this video show?

# Quick Preview: The problem

- Do video foundation models truly understand time?
- Our idea for a “test of time”: ask questions that have temporal relations

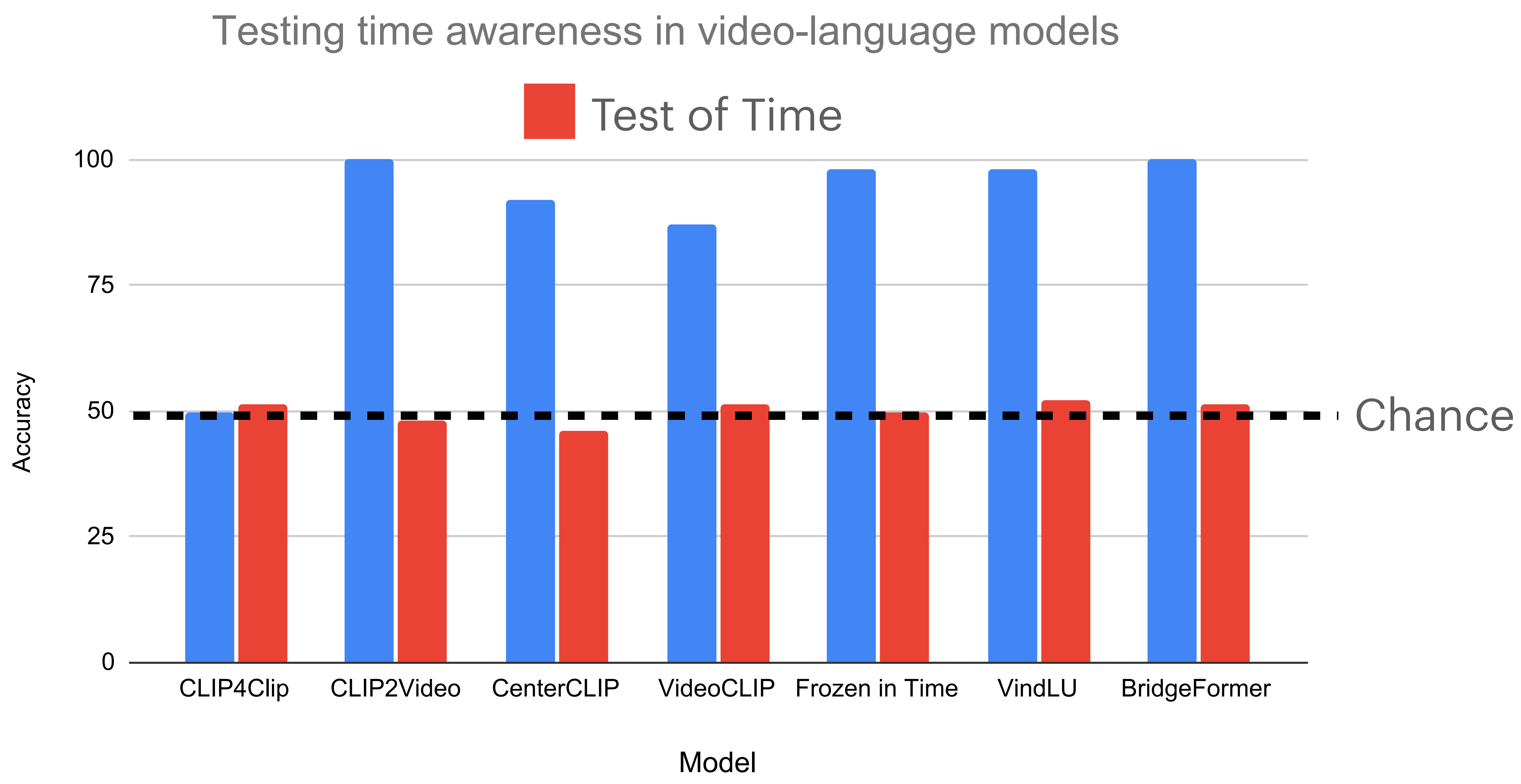


“False”

The baby eats ice-cream **before** walking down hill? True or False?

# Quick Preview: Our contributions

- Key finding: We find that seven existing video-language models fail this test of time



# Quick Preview: Our contributions

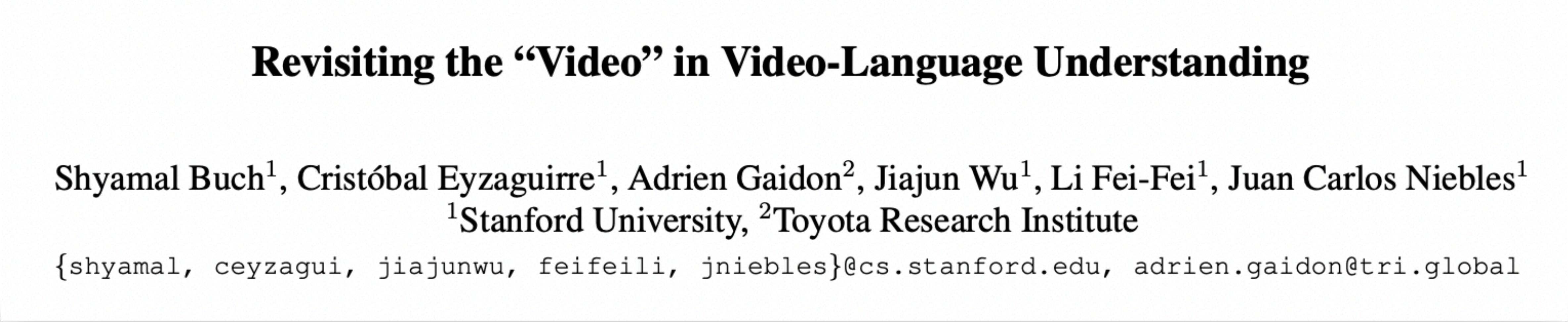
- Key finding: We find that seven existing video-language models fail this test of time
- Our solution: adapt video-language models with contrastive learning on carefully designed negatives

The baby eats ice-cream **before** walking down hill. ✓

The baby walks down hill **before** eating ice-cream. ✗

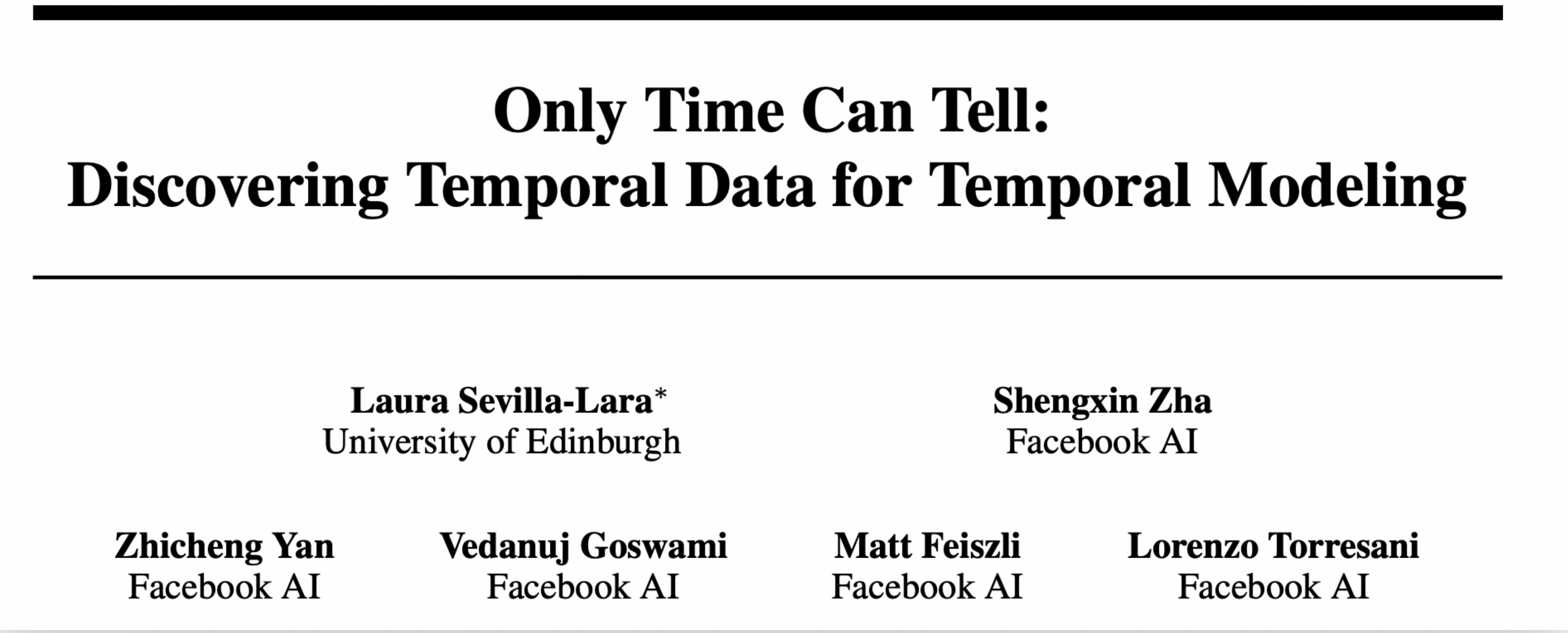
# The test of time

- The static image bias in current video benchmarks



### Revisiting the "Video" in Video-Language Understanding

Shyamal Buch[1], Cristóbal Eyzaguirre[1], Adrien Gaidon[2], Jiajun Wu[1], Li Fei-Fei[1], Juan Carlos Niebles[1]
[1]Stanford University, [2]Toyota Research Institute
{shyamal, ceyzagui, jiajunwu, feifeili, jniebles}@cs.stanford.edu, adrien.gaidon@tri.global



### Only Time Can Tell: Discovering Temporal Data for Temporal Modeling

**Laura Sevilla-Lara***
University of Edinburgh

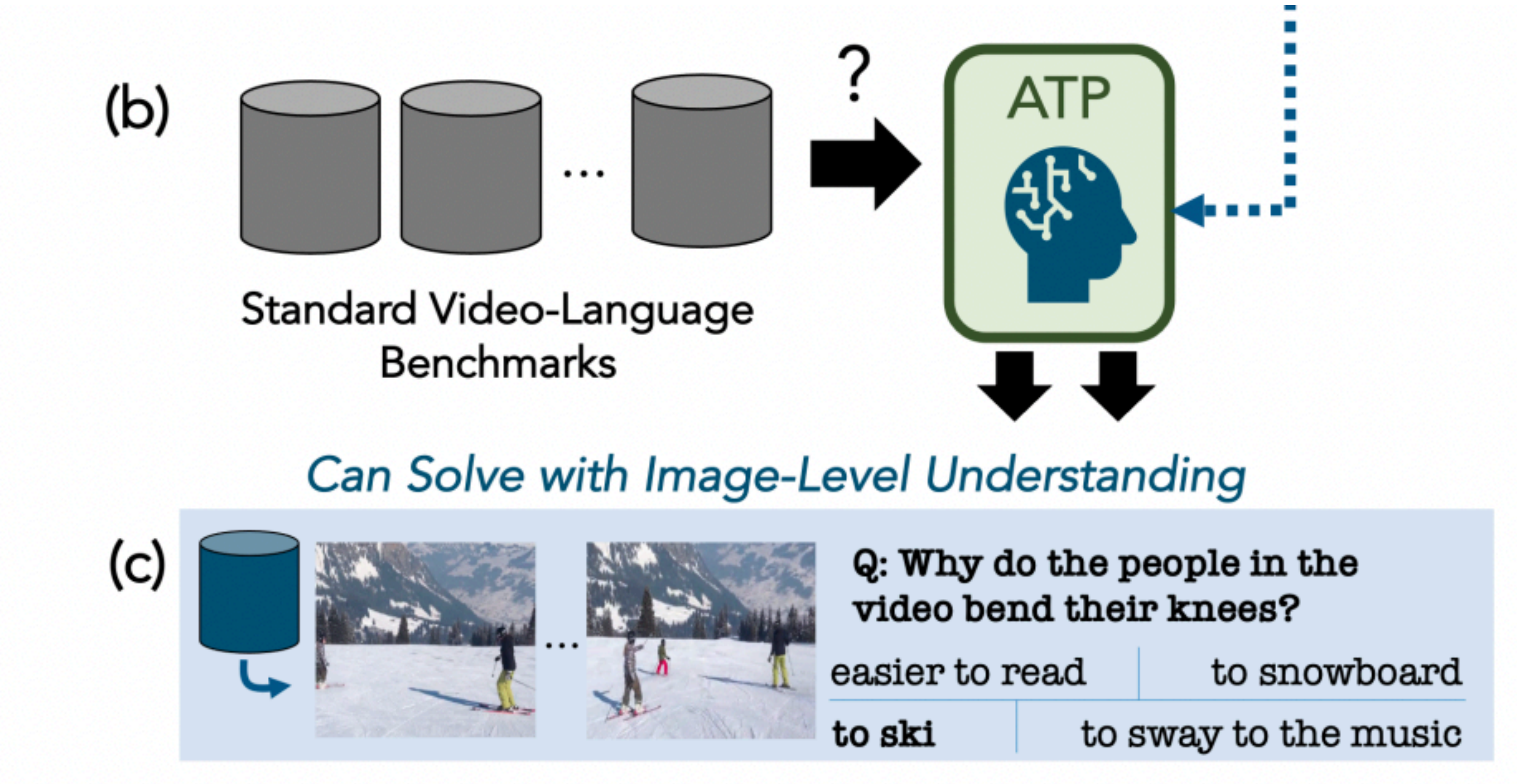
**Shengxin Zha**
Facebook AI

**Zhicheng Yan**
Facebook AI

**Vedanuj Goswami**
Facebook AI

**Matt Feiszli**
Facebook AI

**Lorenzo Torresani**
Facebook AI


(b)
...
?
ATP
Standard Video-Language Benchmarks
Can Solve with Image-Level Understanding
(c)
...
Q: Why do the people in the video bend their knees?
easier to read
to snowboard
to ski
to sway to the music

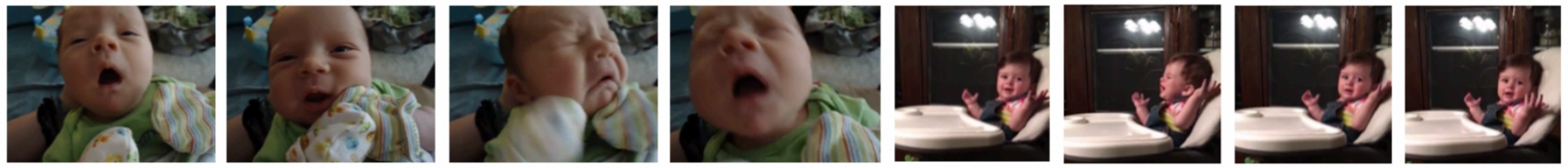
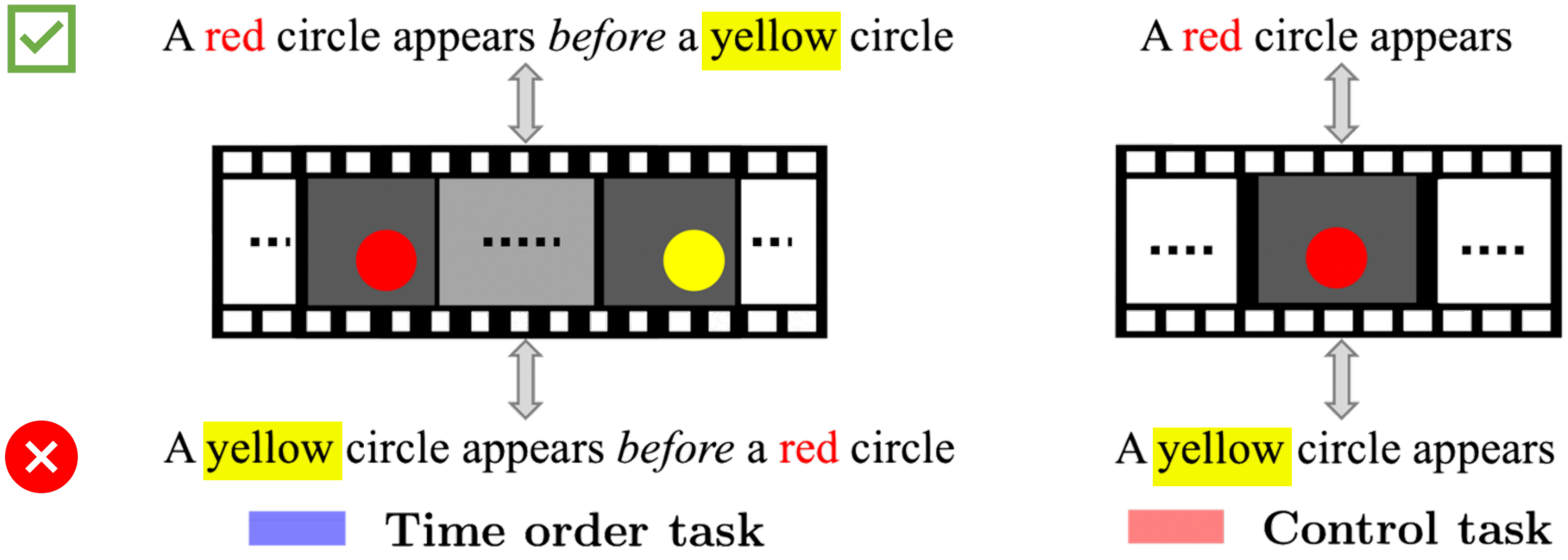


Figure 1: **Can you guess these actions? "yawning", "sneezing" or "crying"?** Temporal information is essential to discriminate some actions, while for others it is redundant. Shuffling frames in time removes temporal information, revealing the actions where it actually matters. (Solution at the end of the paper.)

# The test of time

- The static image bias in current video benchmarks
- Synthetic benchmark



✓ A red circle appears *before* a yellow circle

✗ A yellow circle appears *before* a red circle

🟦 Time order task

A red circle appears

A yellow circle appears

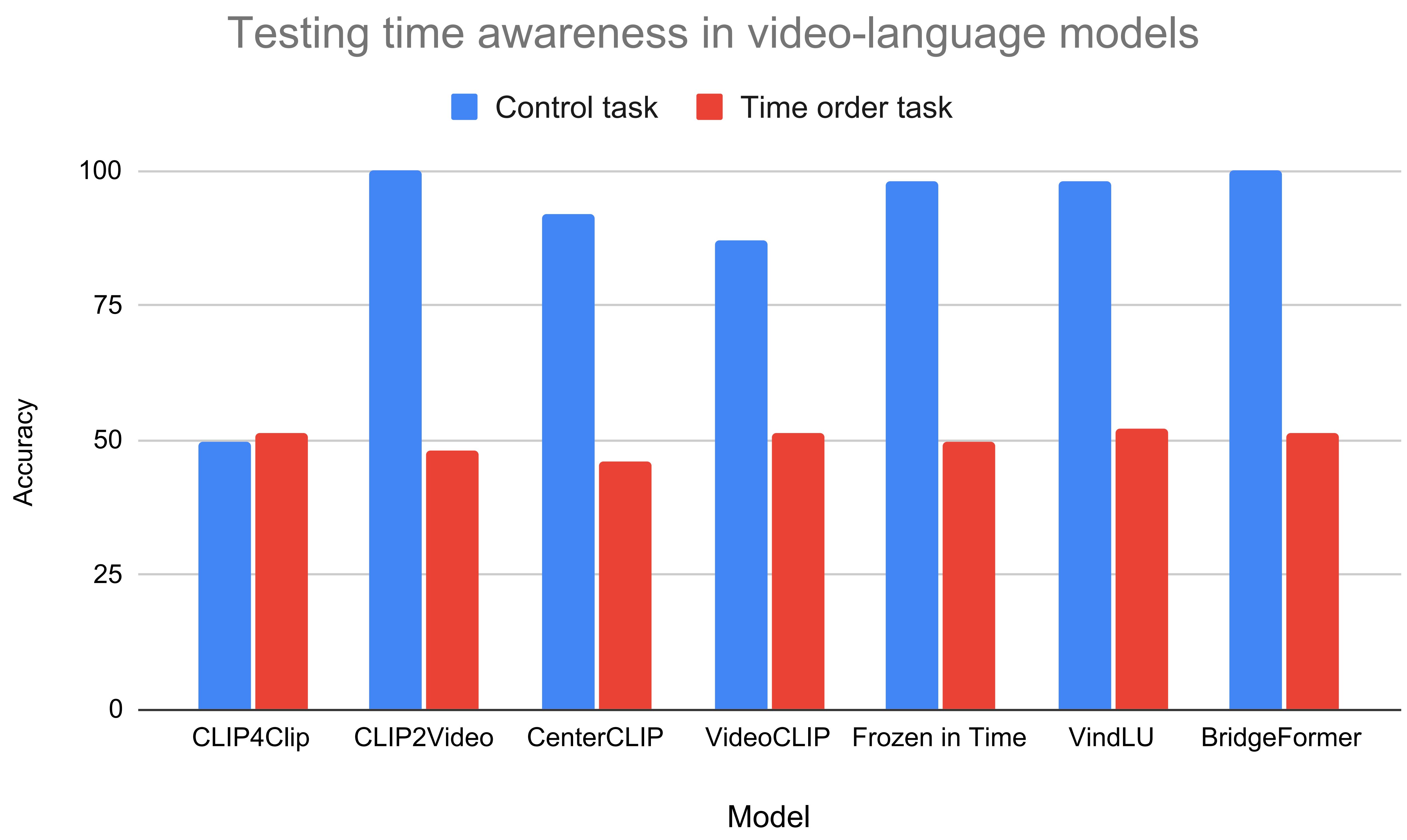
🟥 Control task

# Existing model fail this test of time

- We pick a suite of seven openly available video-language models

# Existing model fail this test of time

- We pick a suite of seven openly available video-language models
- While excelling at the control task, they all fail at the time-order task

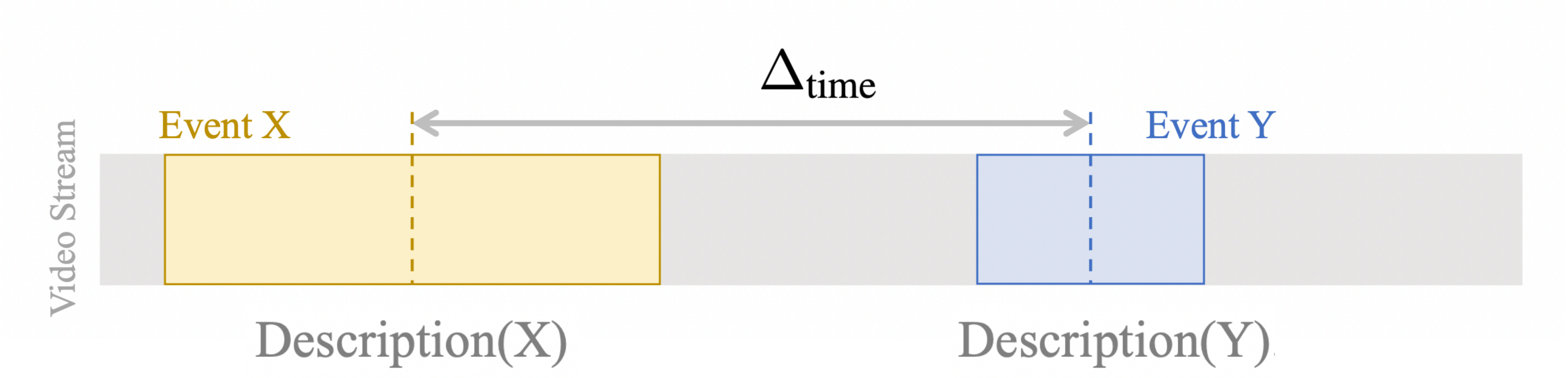


# How to instil this sense of time?

- Post-pretraining: instead of training for scratch, we run another round of pre-training

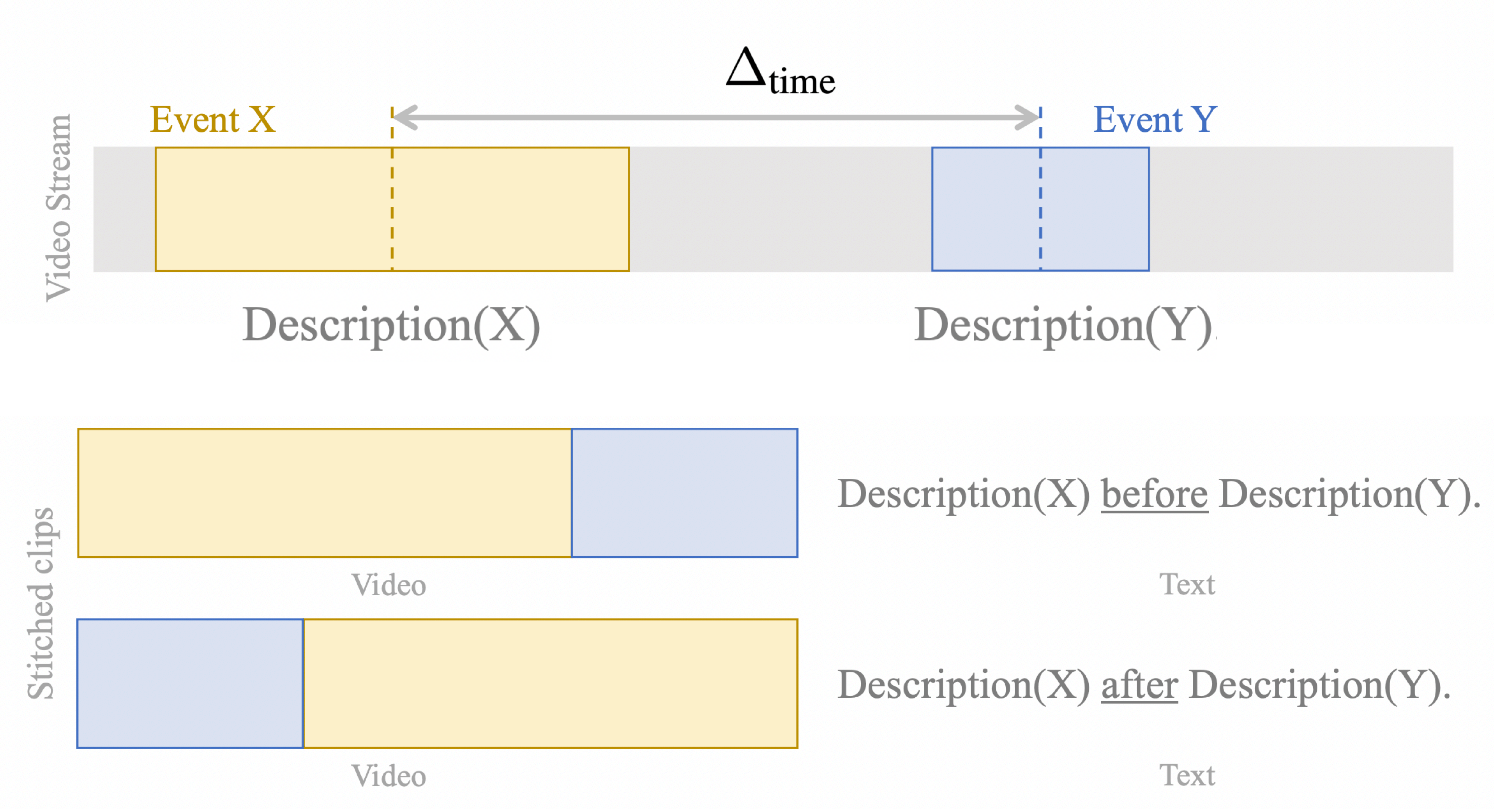
# How to instil this sense of time?

- Data: any dense video-captioning dataset!



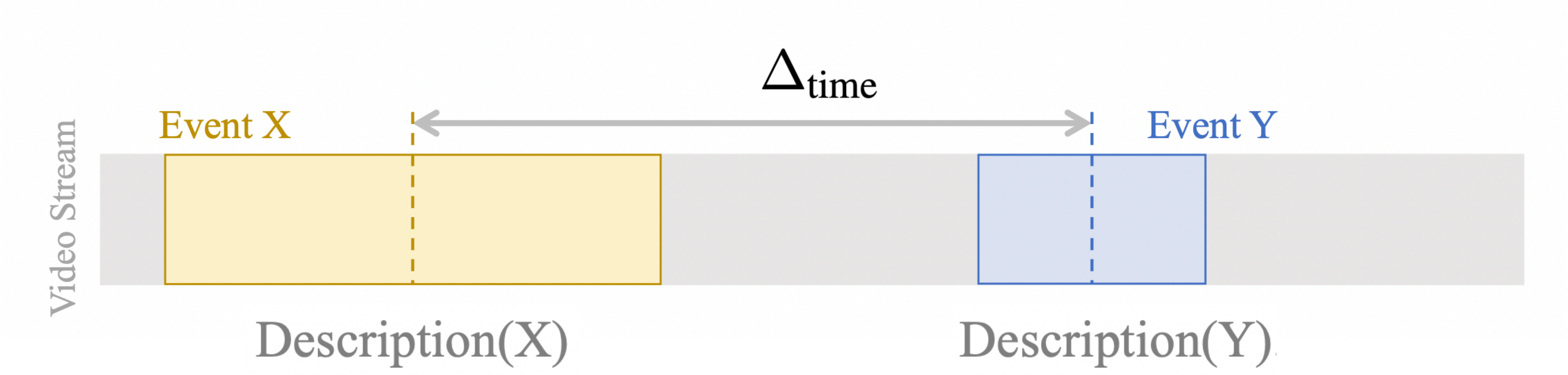
# How to instil this sense of time?

- Data: any dense video-captioning dataset!



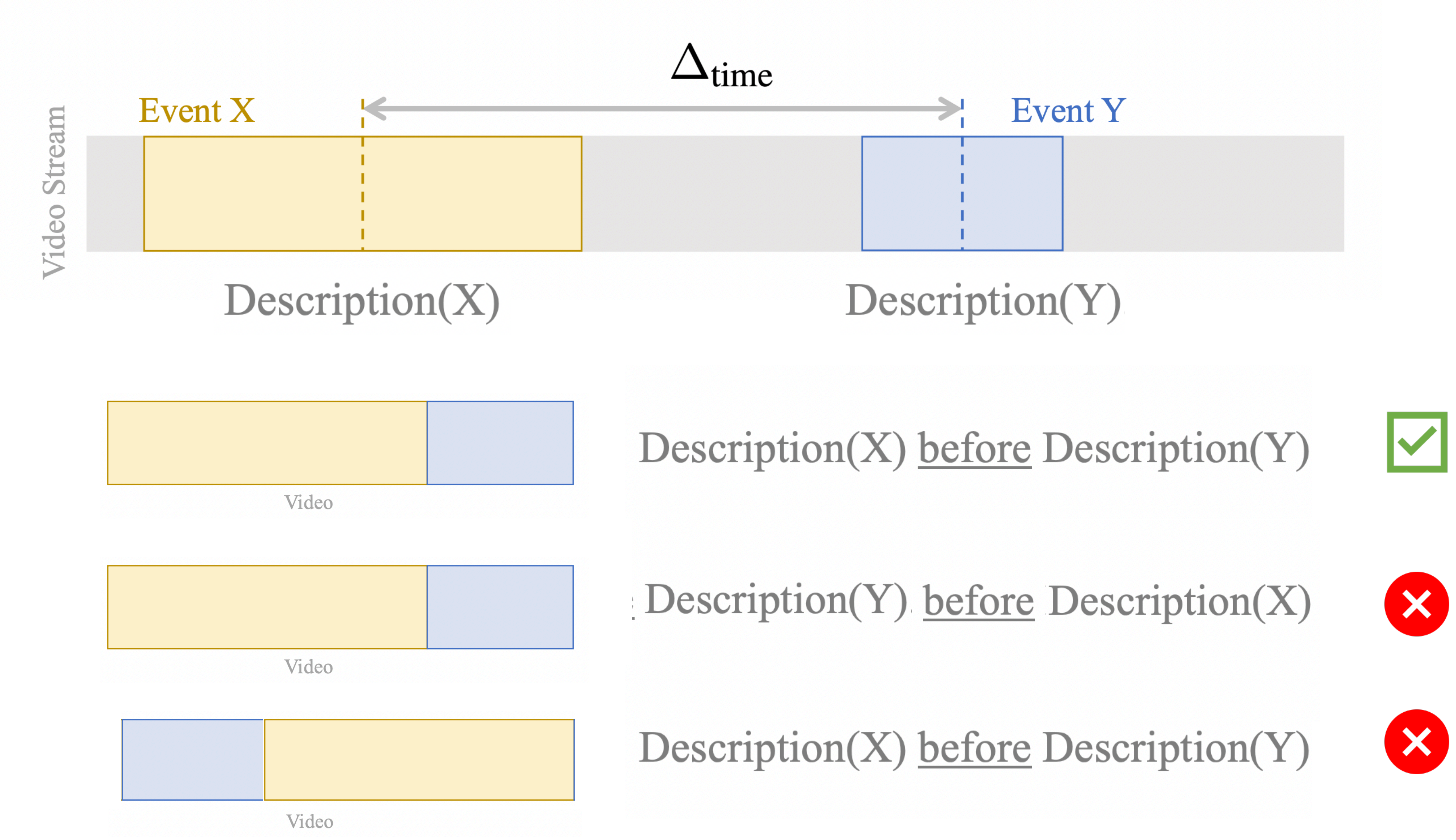
# How to instil this sense of time?

- Data: any dense video-captioning dataset!



# How to instil this sense of time?

- Data: any dense video-captioning dataset!



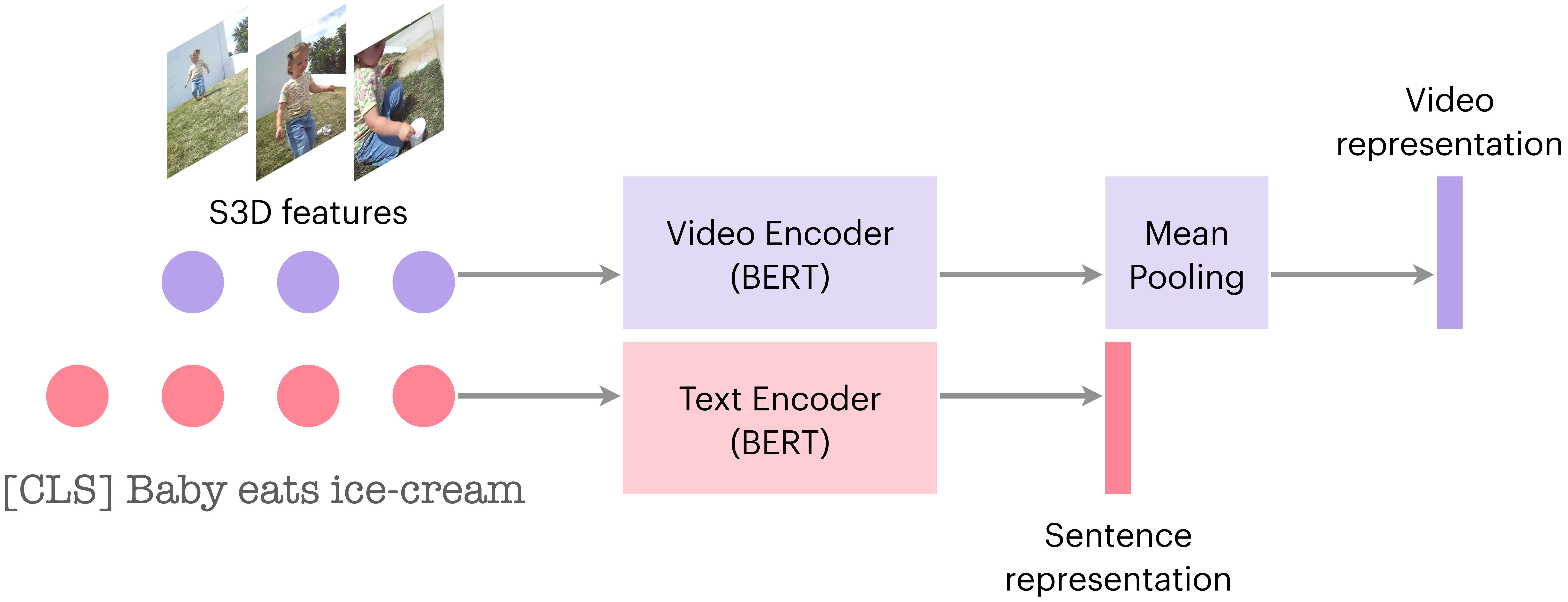
Description(X) before Description(Y) ☑

Description(Y) before Description(X) ✗

Description(X) before Description(Y) ✗

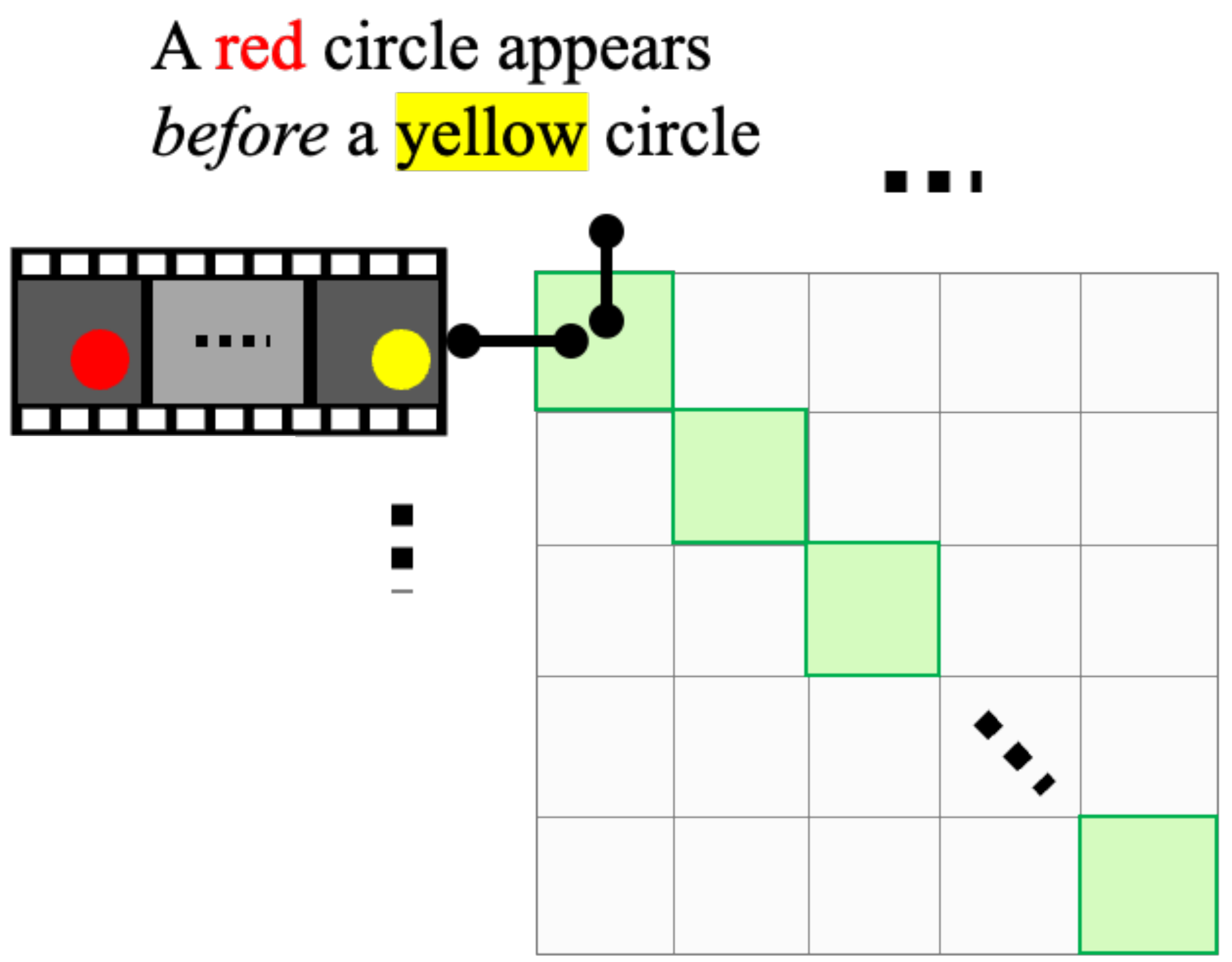
# How to instil this sense of time?

- Base model: We start with a pre-trained model: VideoCLIP [1]



[1] Xu et al, VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding, EMNLP 2021.

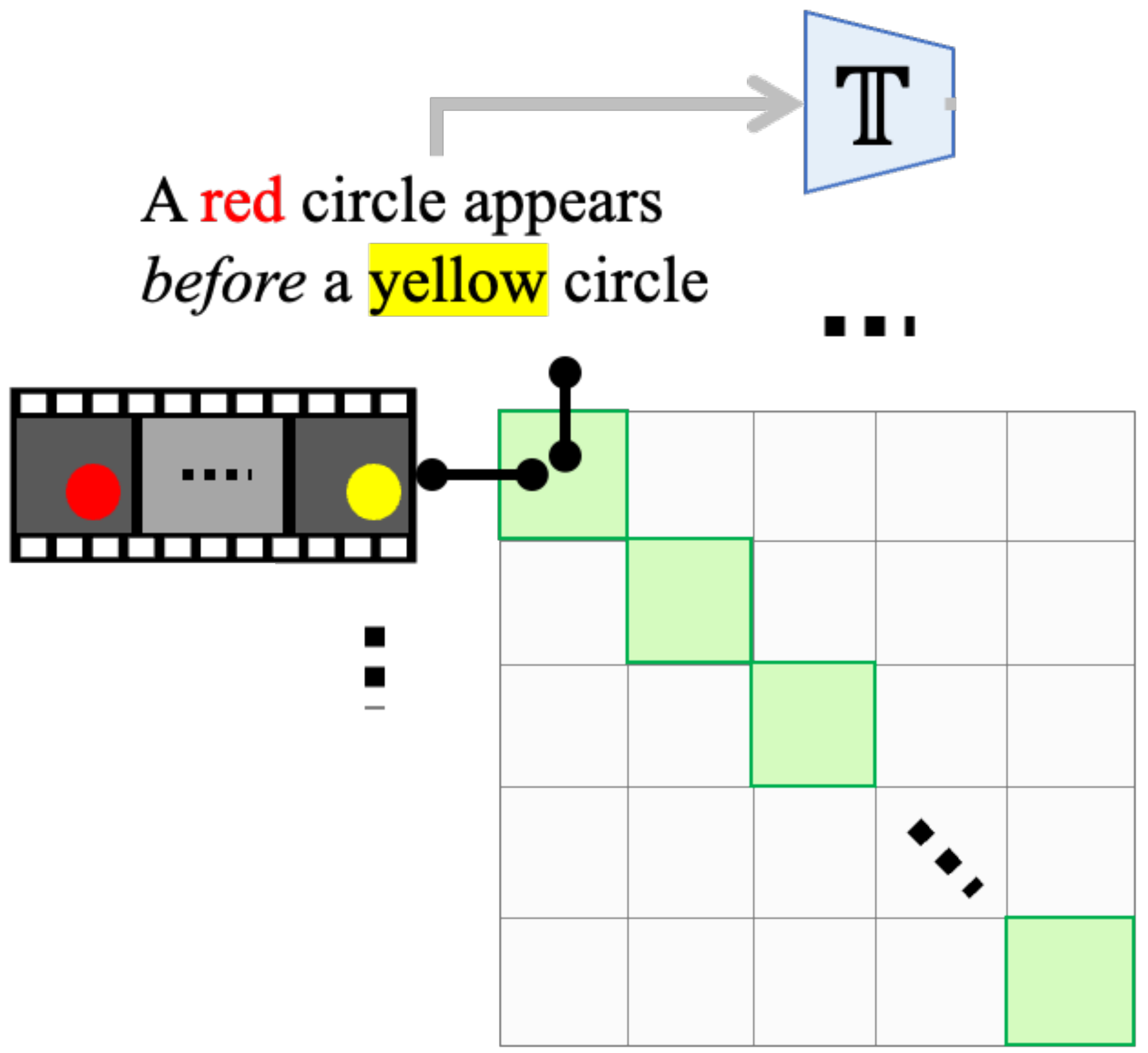
# How to instil this sense of time?



🟩 Usual Positives

□ Usual Negatives

# How to instil this sense of time?

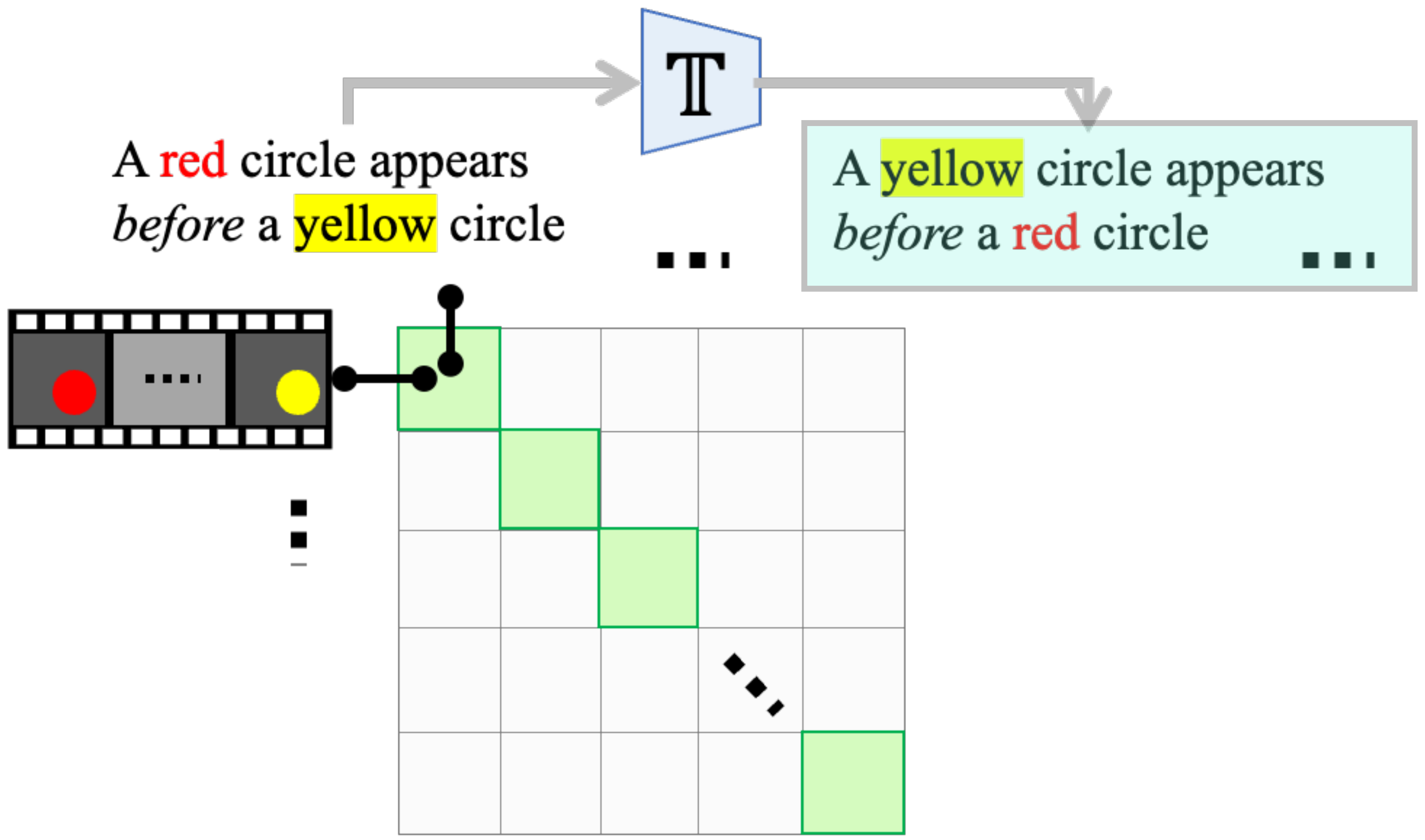


🟩 Usual Positives

□ Usual Negatives

𝕋 Time-order reversal function

# How to instil this sense of time?



A red circle appears *before* a yellow circle

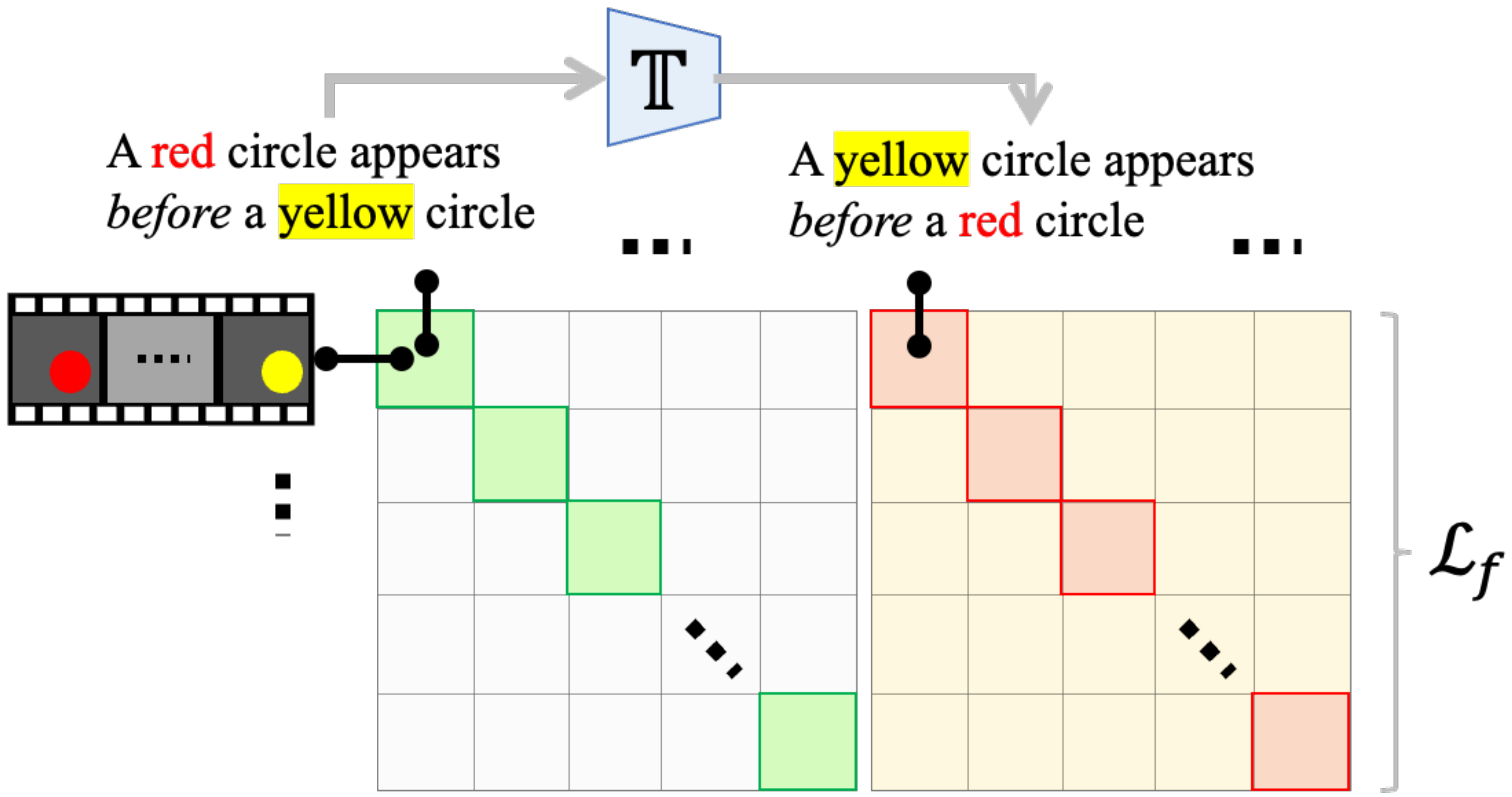
A yellow circle appears *before* a red circle

🟩 Usual Positives

□ Usual Negatives

𝕋 Time-order reversal function

# How to instil this sense of time?



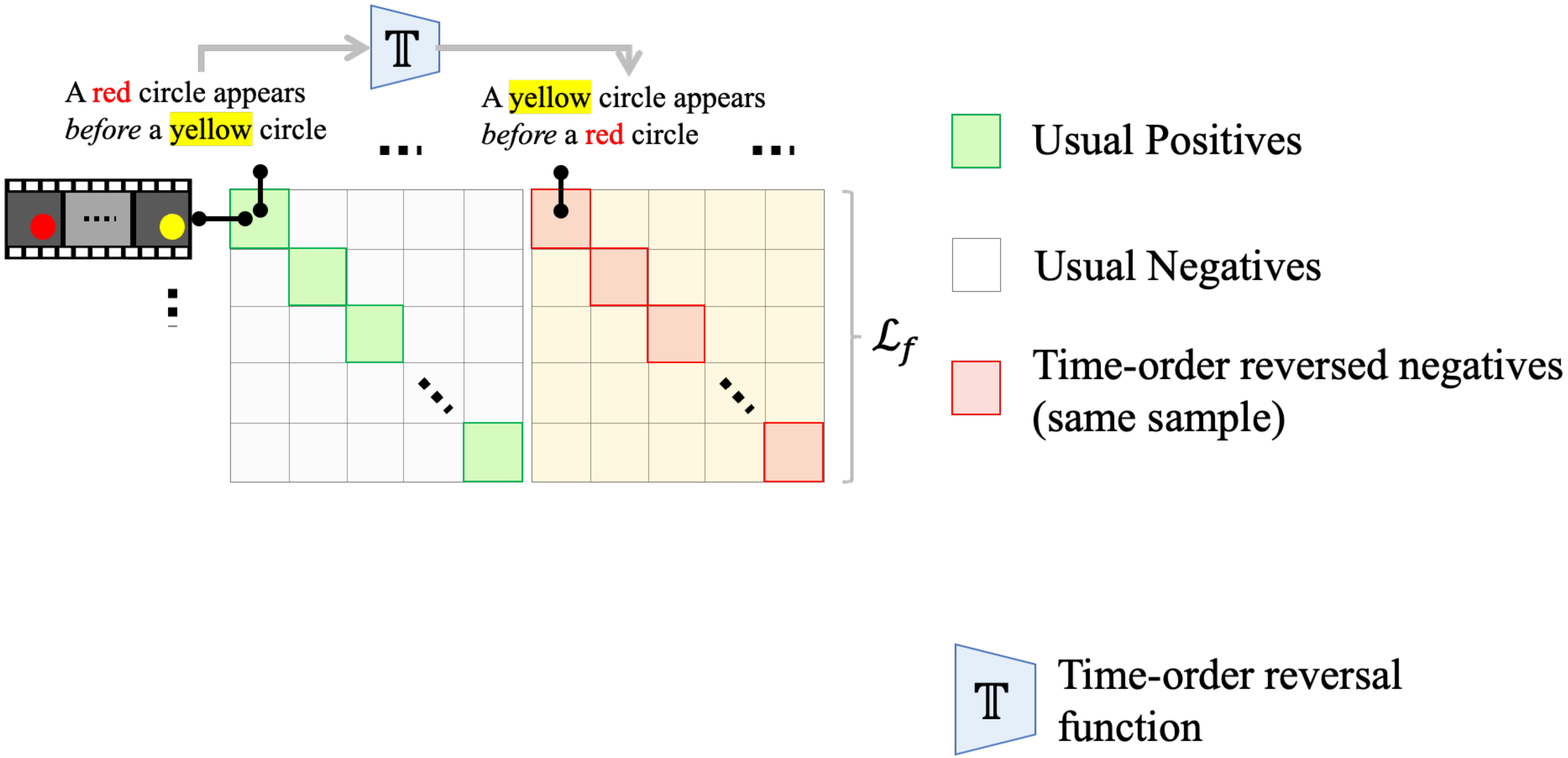
$\mathbb{T}$
A red circle appears *before* a yellow circle
A yellow circle appears *before* a red circle
$\mathcal{L}_f$

🟩 Usual Positives

□ Usual Negatives

$\mathbb{T}$ Time-order reversal function

# How to instil this sense of time?



A red circle appears *before* a yellow circle

A yellow circle appears *before* a red circle

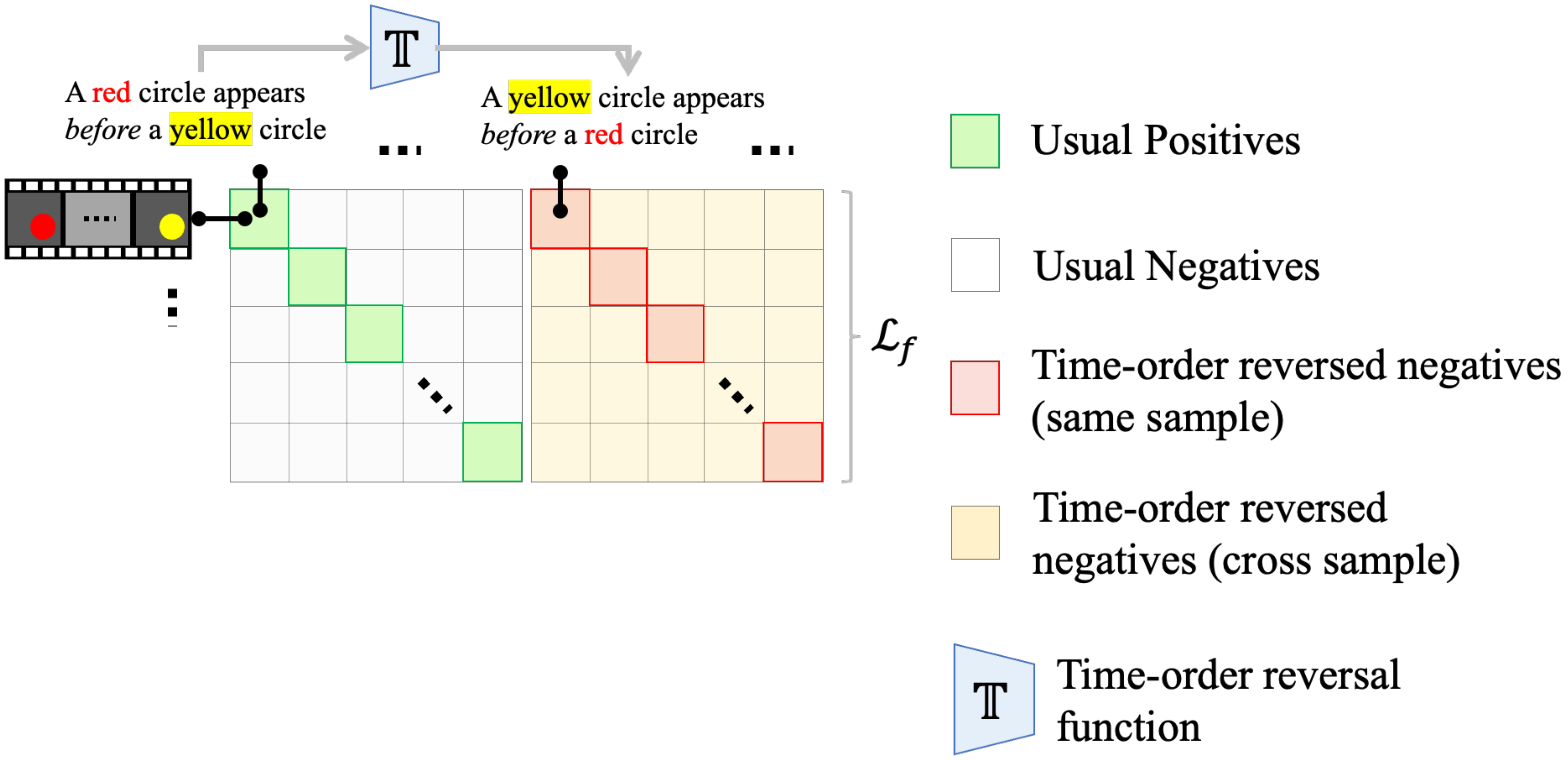
🟩 Usual Positives

□ Usual Negatives

🟥 Time-order reversed negatives (same sample)

Time-order reversal function

# How to instil this sense of time?



𝕋
A red circle appears *before* a yellow circle
A yellow circle appears *before* a red circle
$\mathcal{L}_f$

🟩 Usual Positives

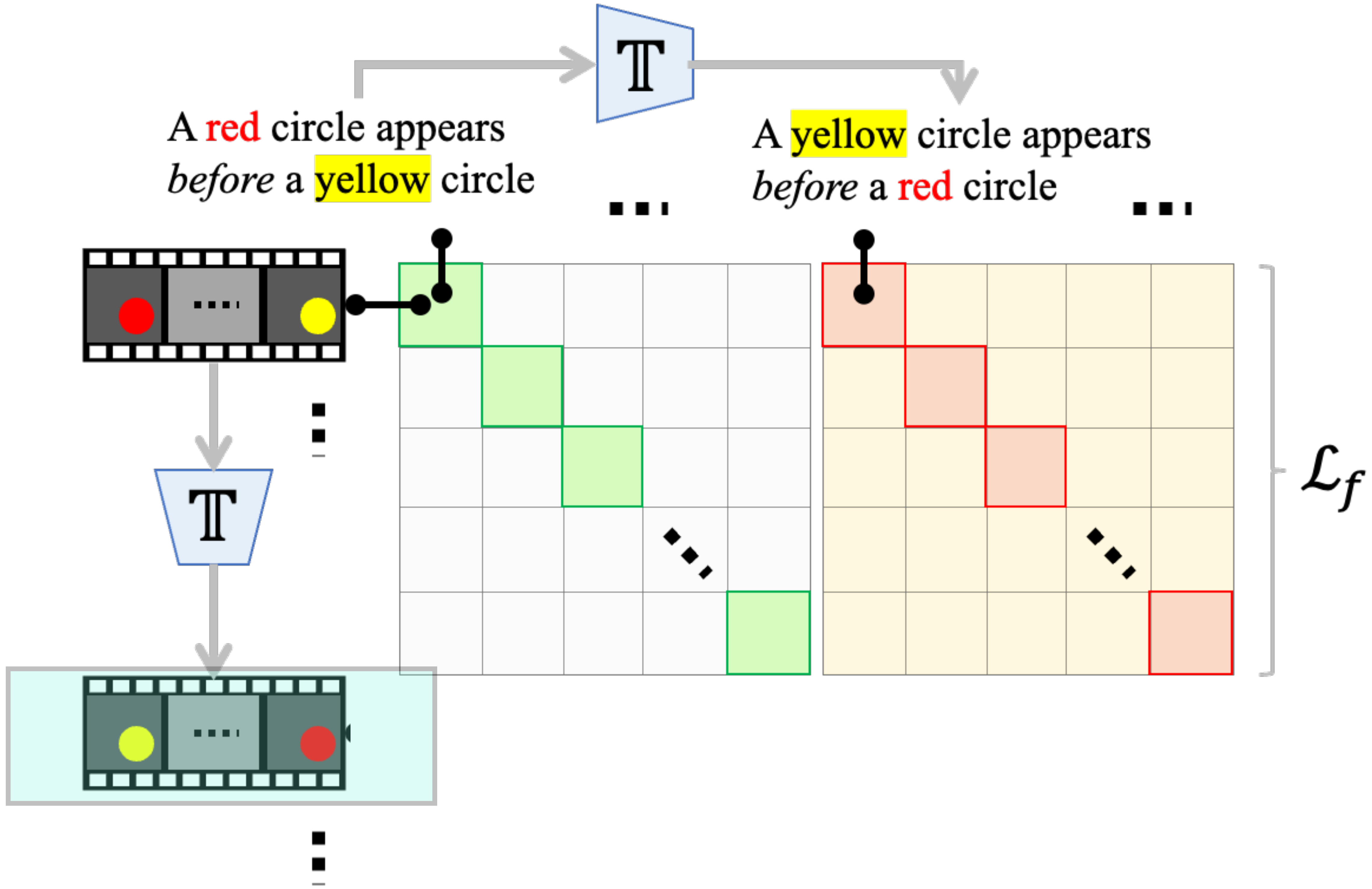
□ Usual Negatives

🟥 Time-order reversed negatives (same sample)

🟨 Time-order reversed negatives (cross sample)

𝕋 Time-order reversal function

# How to instil this sense of time?



A red circle appears *before* a yellow circle

A yellow circle appears *before* a red circle

🟩 Usual Positives

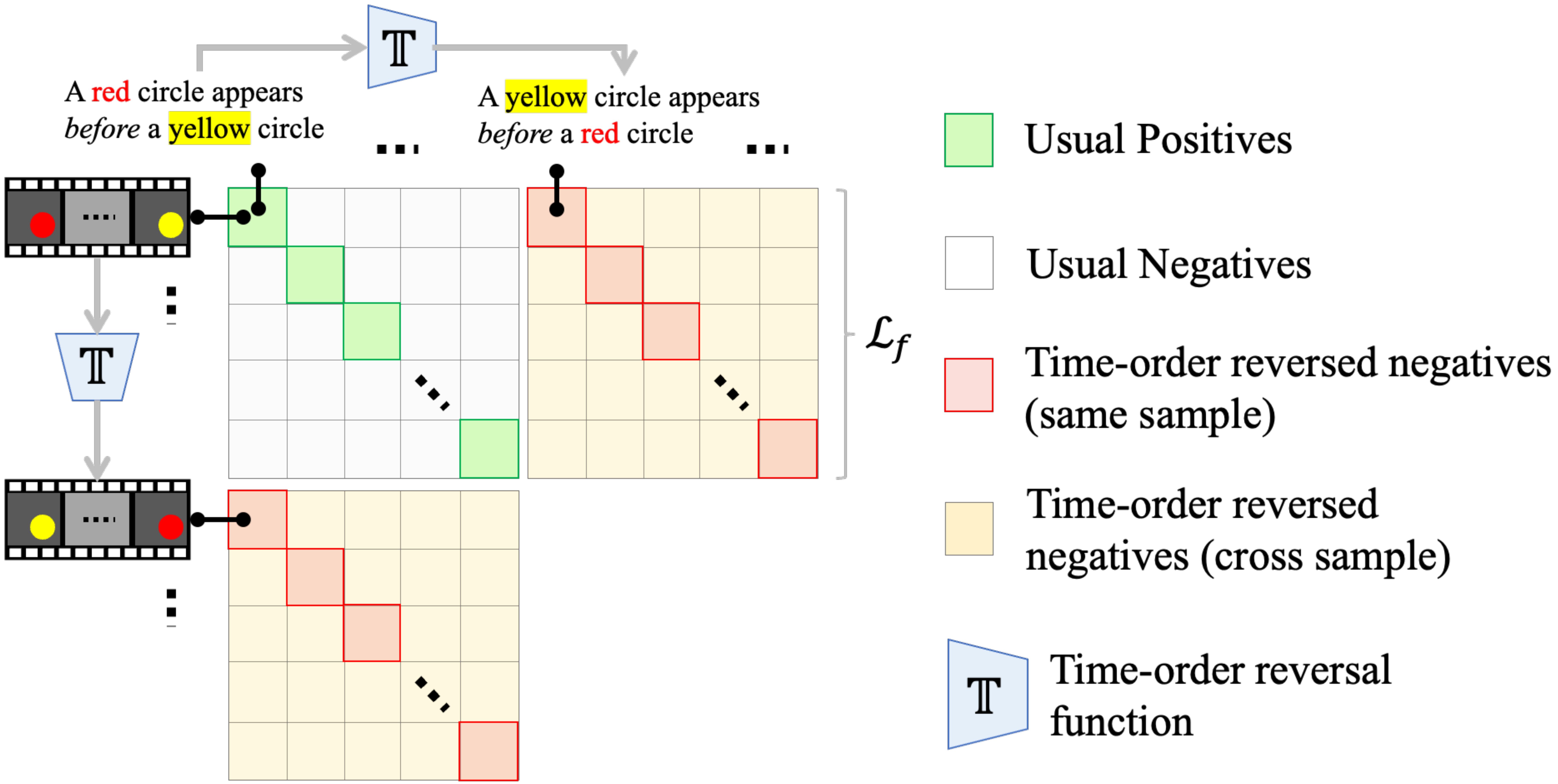
□ Usual Negatives

🟥 Time-order reversed negatives (same sample)

🟨 Time-order reversed negatives (cross sample)

𝕋 Time-order reversal function

# How to instil this sense of time?



A red circle appears *before* a yellow circle

A yellow circle appears *before* a red circle

$\mathbb{T}$

$\mathcal{L}_f$

🟩 Usual Positives

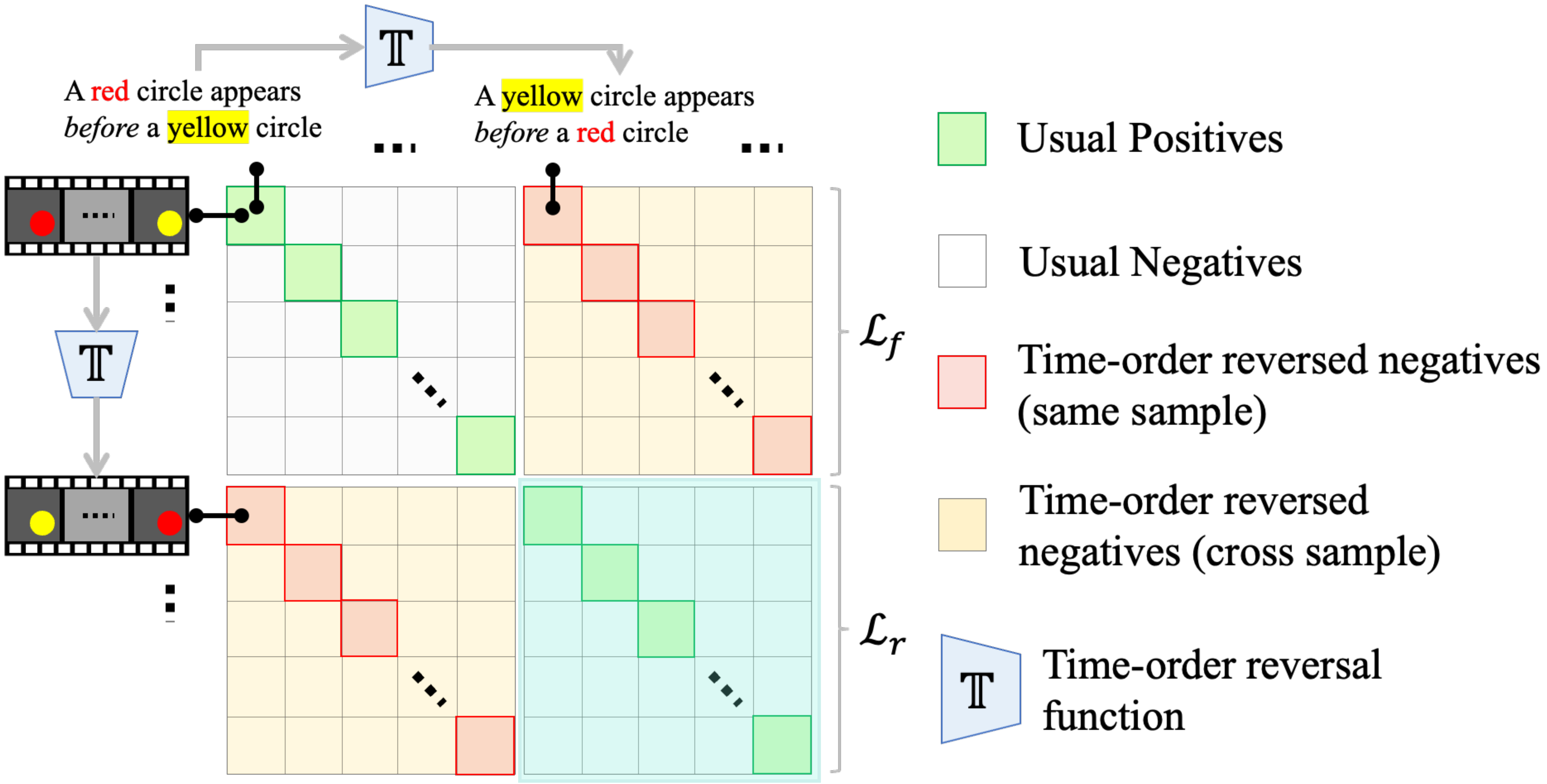
□ Usual Negatives

🟥 Time-order reversed negatives (same sample)

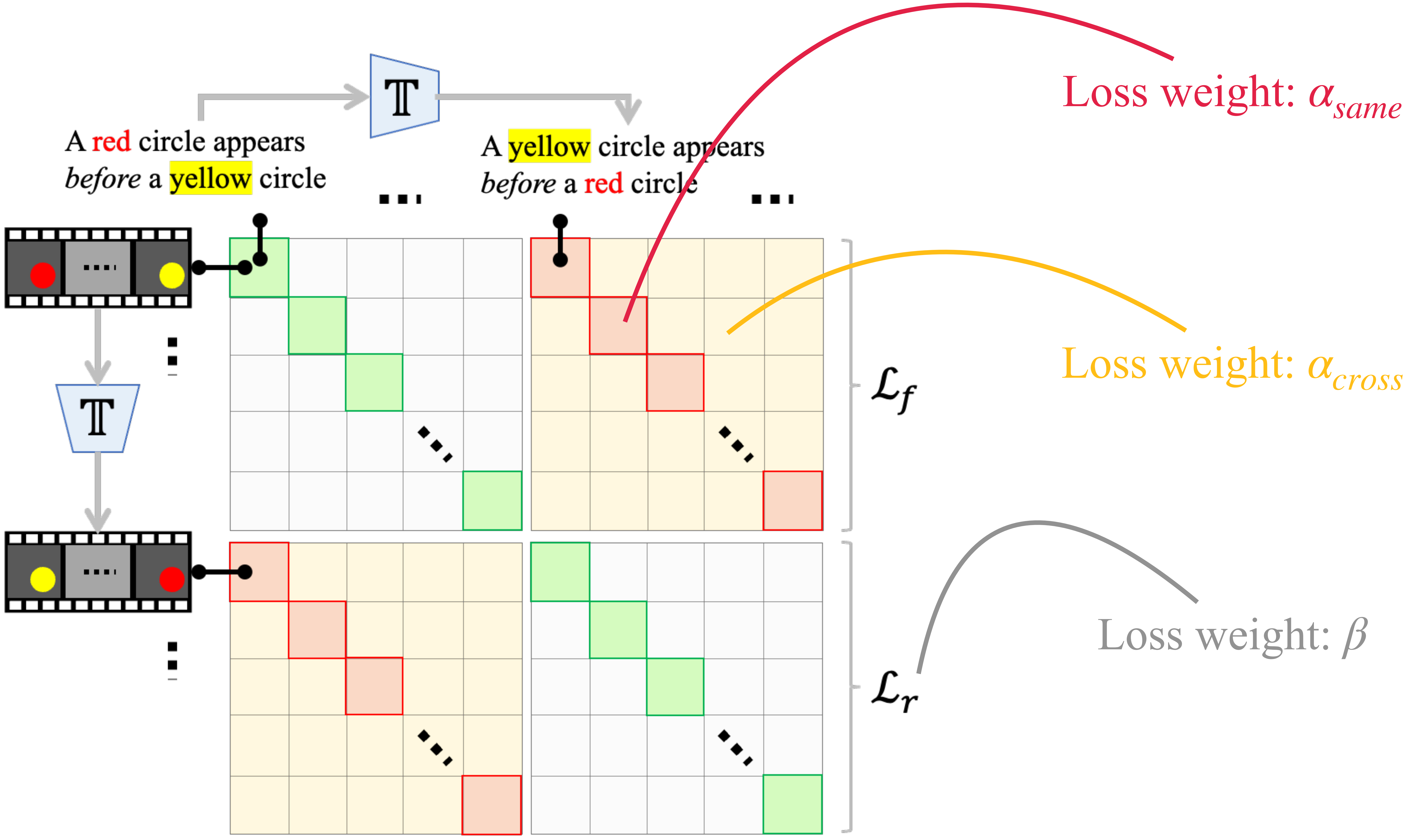
🟨 Time-order reversed negatives (cross sample)

$\mathbb{T}$ Time-order reversal function

# How to instil this sense of time?

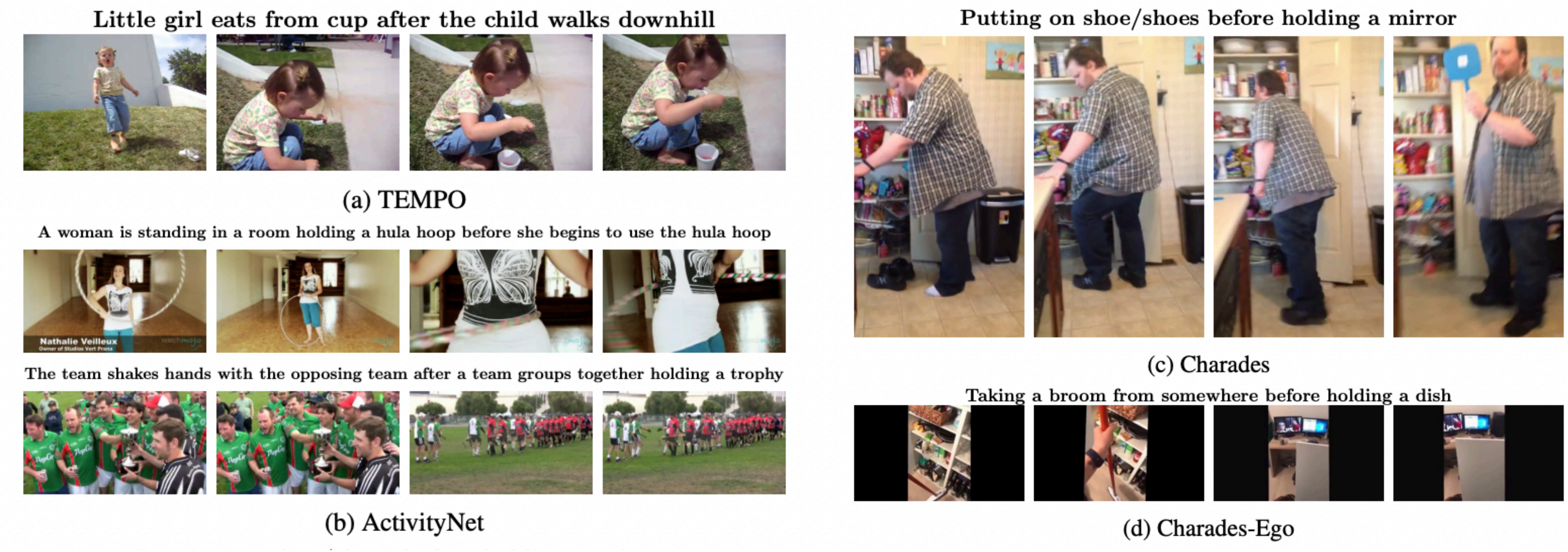

A red circle appears *before* a yellow circle
A yellow circle appears *before* a red circle
$\mathbb{T}$
$\mathcal{L}_f$
$\mathcal{L}_r$
🟩 Usual Positives
□ Usual Negatives
🟥 Time-order reversed negatives (same sample)
🟨 Time-order reversed negatives (cross sample)
$\mathbb{T}$ Time-order reversal function

# How to instil this sense of time?


A red circle appears *before* a yellow circle
A yellow circle appears *before* a red circle
$\mathcal{L}_f$
$\mathcal{L}_r$
Loss weight: $\alpha_{same}$
Loss weight: $\alpha_{cross}$
Loss weight: $\beta$

TACT: Temporal Adaptation by Consistent Time-ordering

# Experiments


Little girl eats from cup after the child walks downhill
A woman is standing in a room holding a hula hoop before she begins to use the hula hoop
Nathalie Veilleux
The team shakes hands with the opposing team after a team groups together holding a trophy
Putting on shoe/shoes before holding a mirror
Taking a broom from somewhere before holding a dish

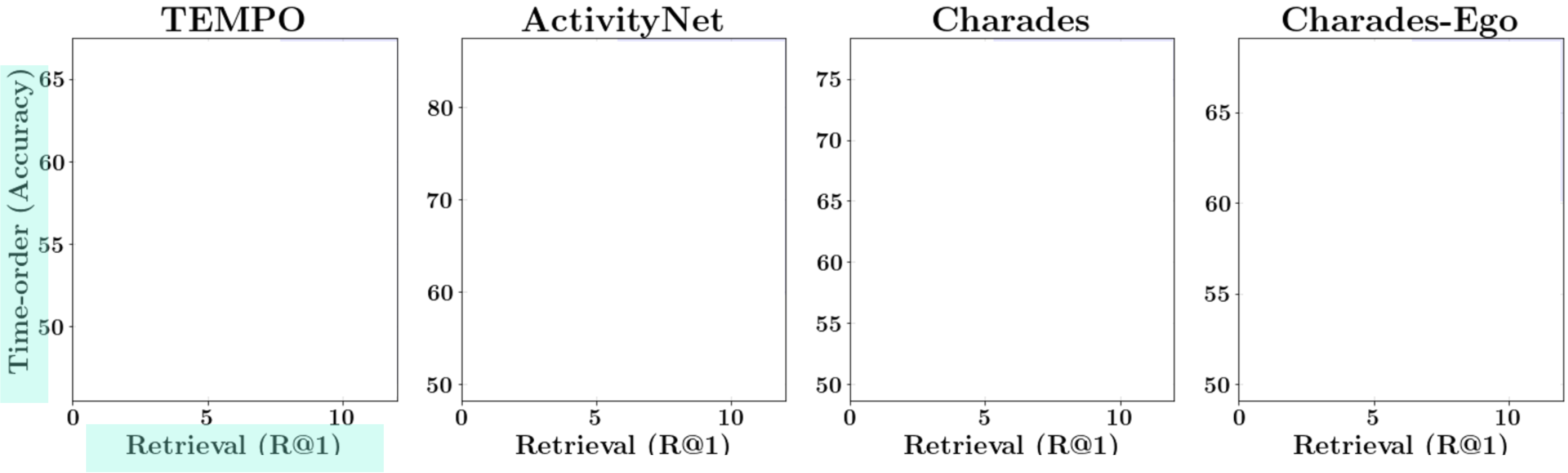
(a) TEMPO

(b) ActivityNet

(c) Charades

(d) Charades-Ego

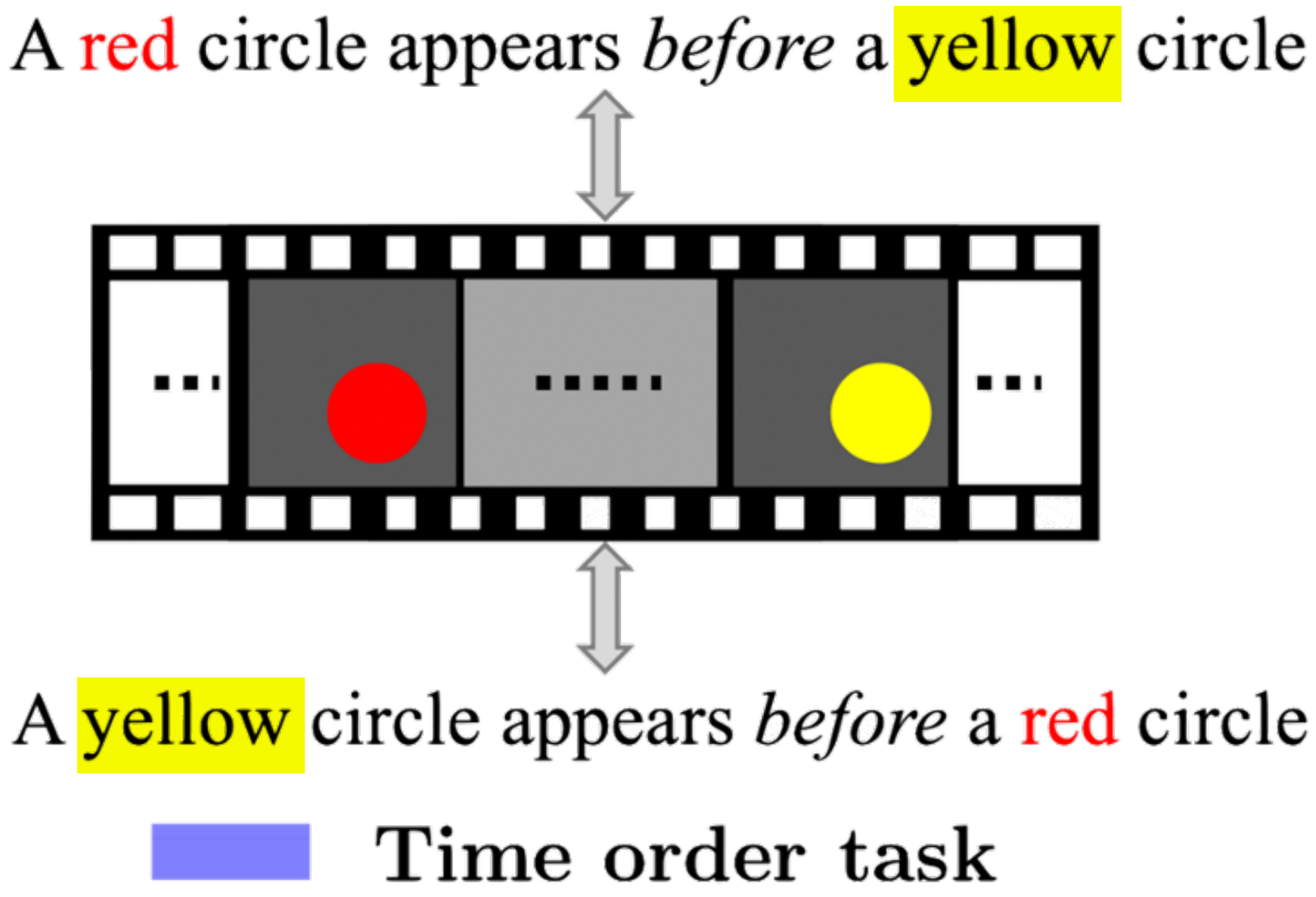
# Experiments

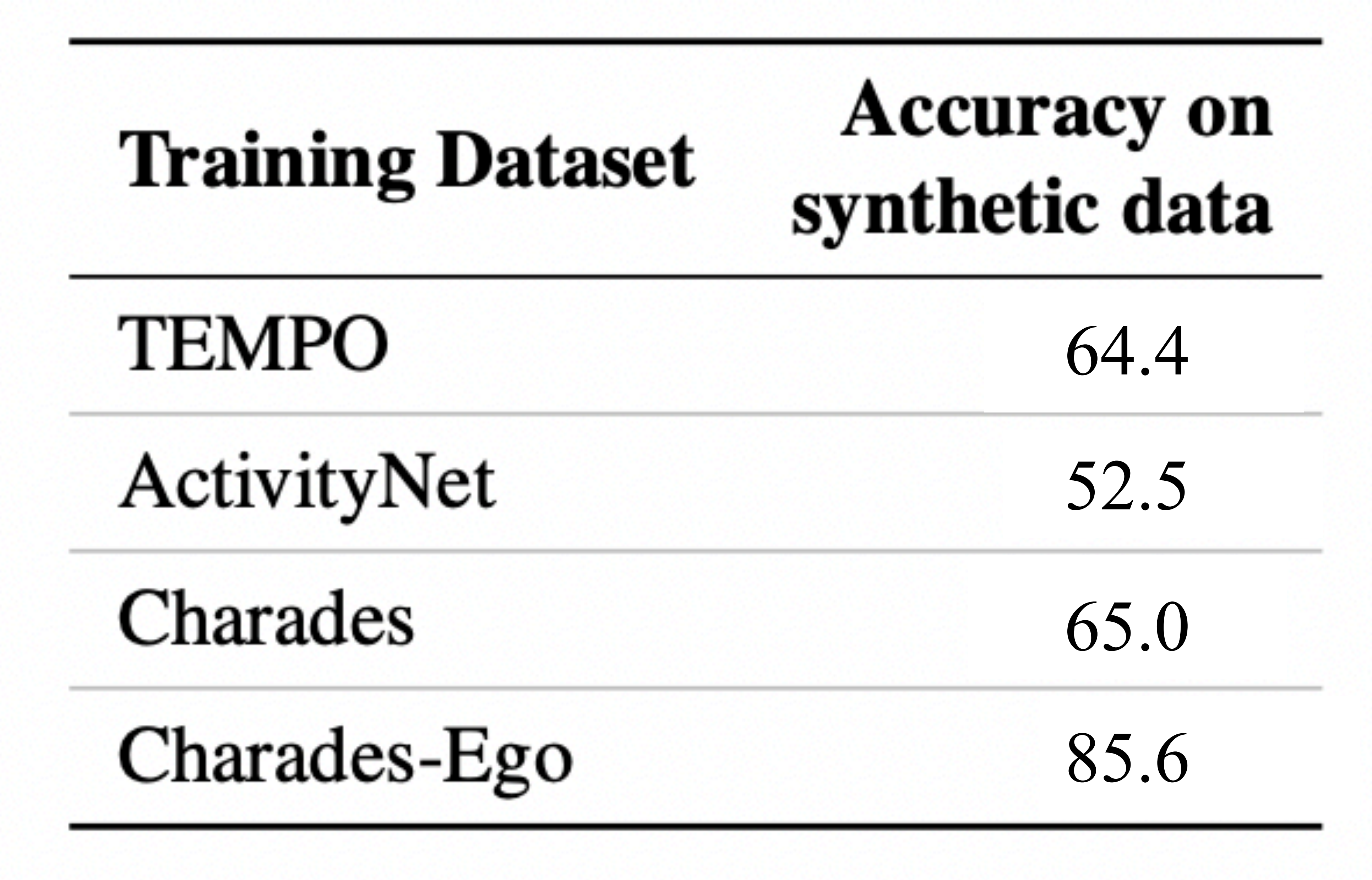


# Experiments

TEMPO
ActivityNet
Charades
Charades-Ego
Time-order (Accuracy)
Retrieval (R@1)
65
60
55
50
80
70
60
50
75
70
65
60
55
50
65
60
55
50
0
5
10
Random
Baseline: VideoCLIP without temporal ordering
Ours
Desirable area

# Experiments: Synthetic benchmark


A red circle appears *before* a yellow circle
.....
A yellow circle appears *before* a red circle
Time order task



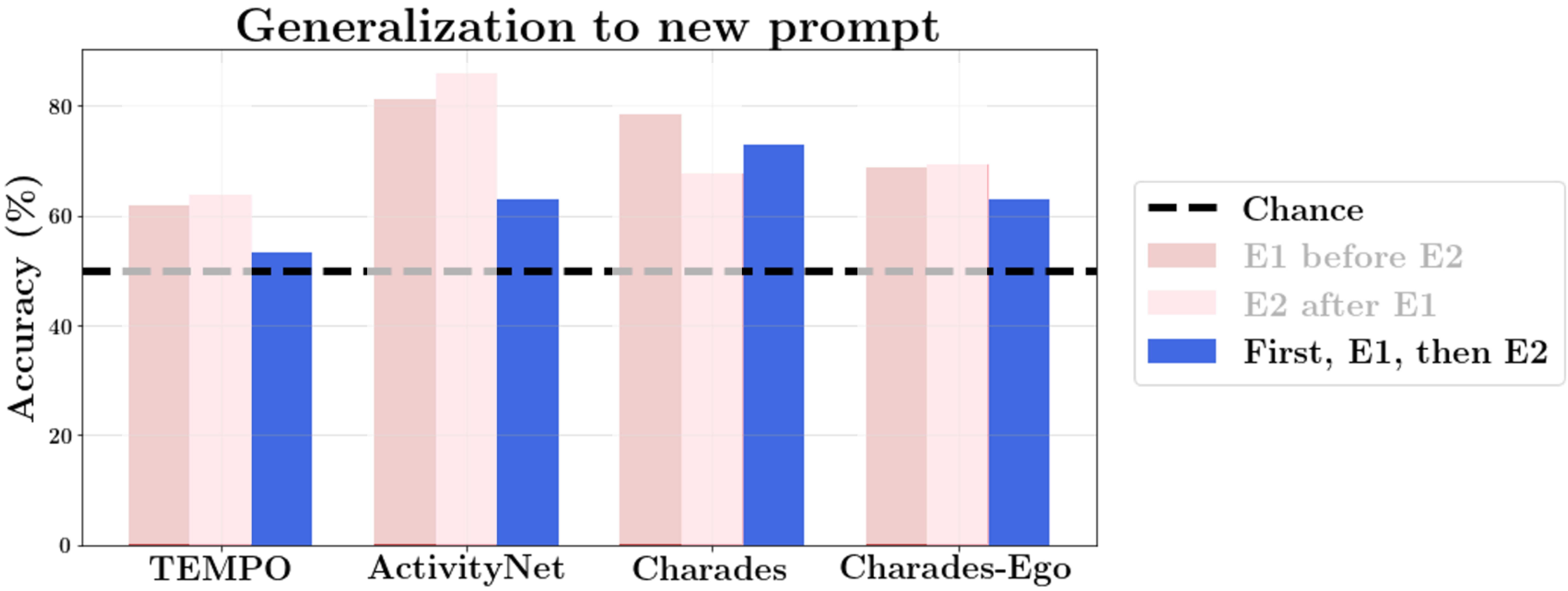
| Training Dataset | Accuracy on synthetic data |
|---|---|
| TEMPO | 64.4 |
| ActivityNet | 52.5 |
| Charades | 65.0 |
| Charades-Ego | 85.6 |

# Does it work beyond before-after relations?

- We evaluate with sentences of the form: “First, [event 1], then, [event 2].”

# Does it work beyond before-after relations?

- We evaluate with sentences of the form: “First, [event 1], then, [event 2].”

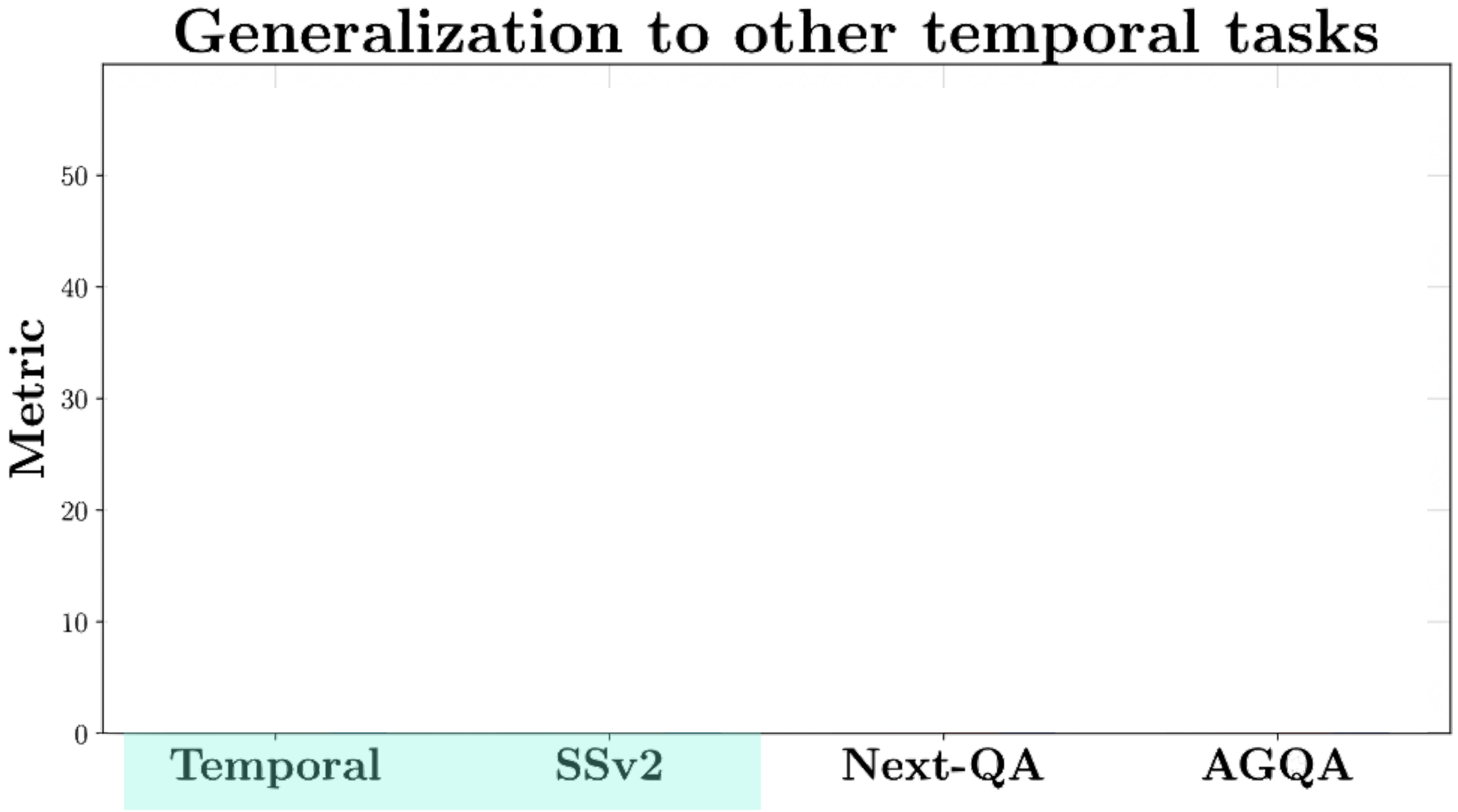


# Does it work beyond this narrow sense of time?

- Does acquiring this narrow sense of time help other general temporal tasks? We find benefits on several temporal reasoning tasks.

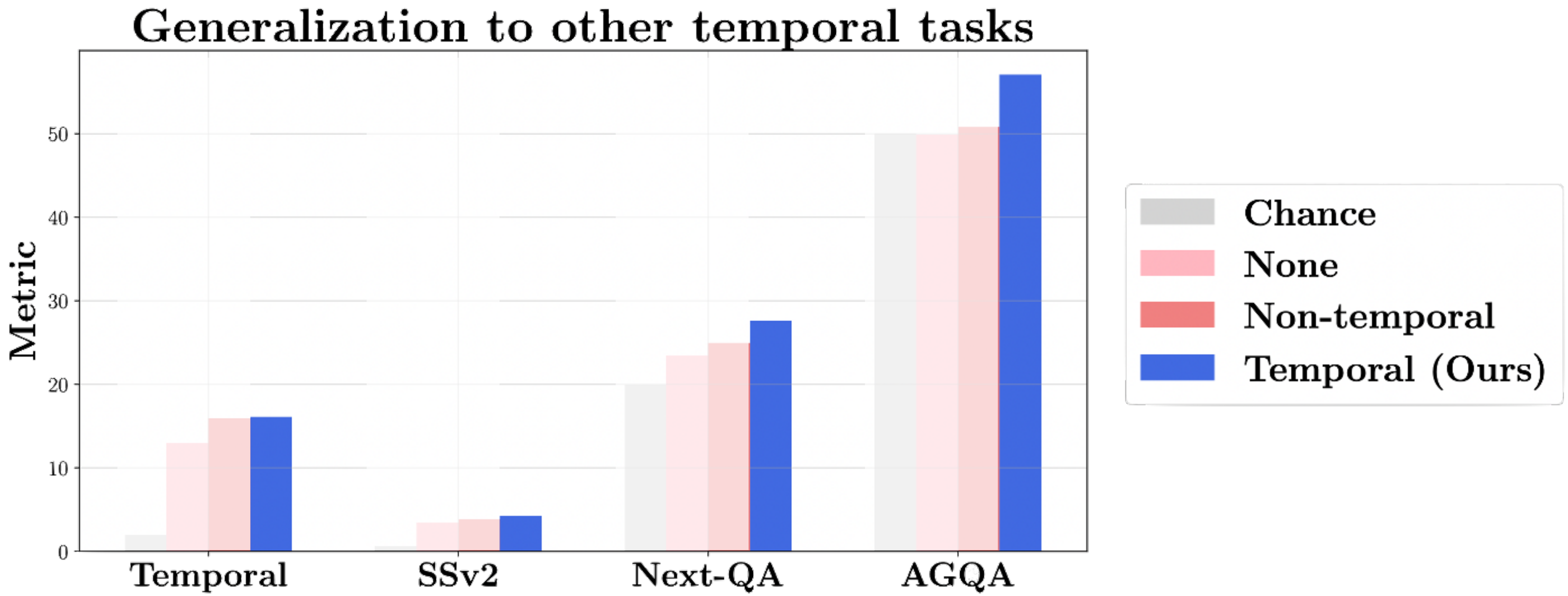
# Does it work beyond this narrow sense of time?

- Does acquiring this narrow sense of time help other general temporal tasks? We find benefits on several temporal reasoning tasks.



# Does it work beyond this narrow sense of time?

- Does acquiring this narrow sense of time help other general temporal tasks? We find benefits on several temporal reasoning tasks.



# Summary

- We propose a “test of time” for video-language models. We show existing models fail on this test.
- We propose a simple recipe, TACT, to instil this sense of time without re-training from scratch.
- We show that adapted models show promise beyond the temporal relations considered and to more general temporal reasoning tasks

# Thank you!

bpiyush.github.io/testoftime-website/

piyush.bagad@student.uva.nl