# Beyond Action Recognition: Detailed Video Modeling

youInc.com

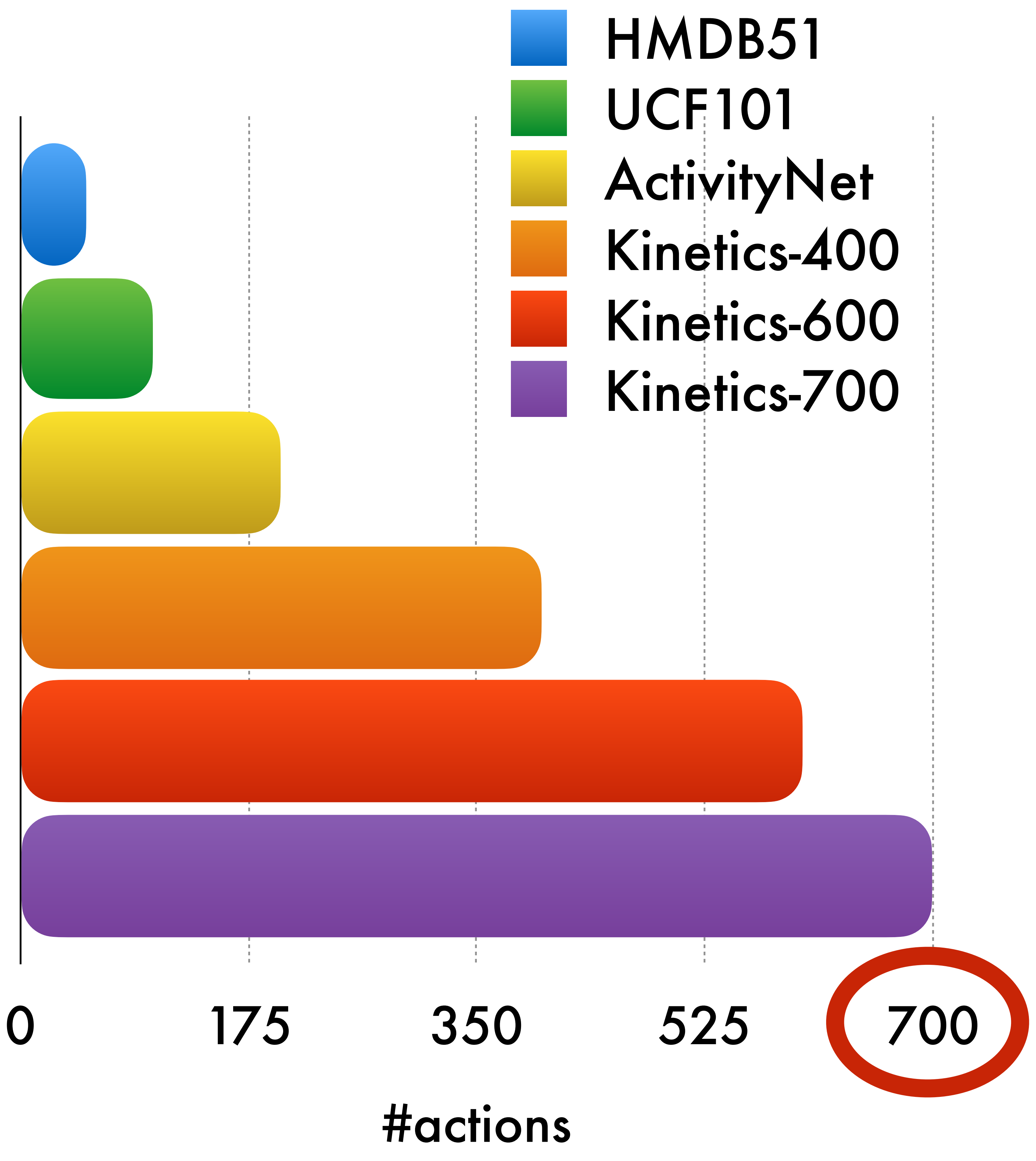

**Gül Varol**

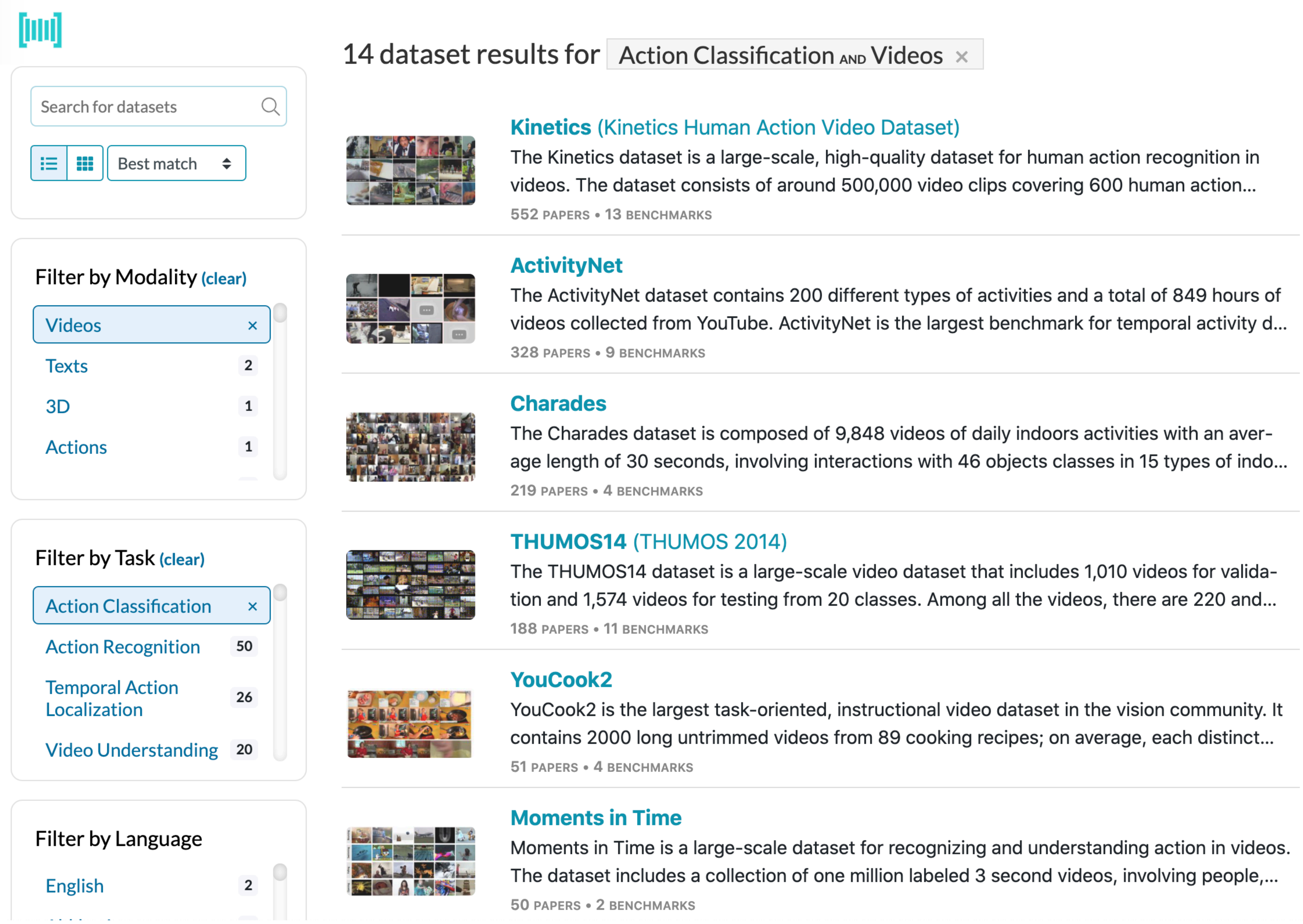
*IMAGINE team, École des Ponts ParisTech*


*@ICCVW, 11.10.2021*

École des Ponts
ParisTech

# What is wrong with action recognition?




paperswithcode.com

# Closed set, loss of information from categorisation

(video from Kinetics)

category: playing ukulele

Description: Two musicians playing ukulele and double bass on the stage.

# Arbitrary level of details

## e.g., Dozen categories for dancing, one category for sign language

Kinetics categories:

Belly dancing
Breakdancing
Country line dancing
Dancing ballet
Dancing charleston
Dancing gangnam style
Dancing macarena
Jumpstyle dancing
Robot dancing
Salsa dancing
Swing dancing
Tango dancing
Tap dancing
Zumba
...

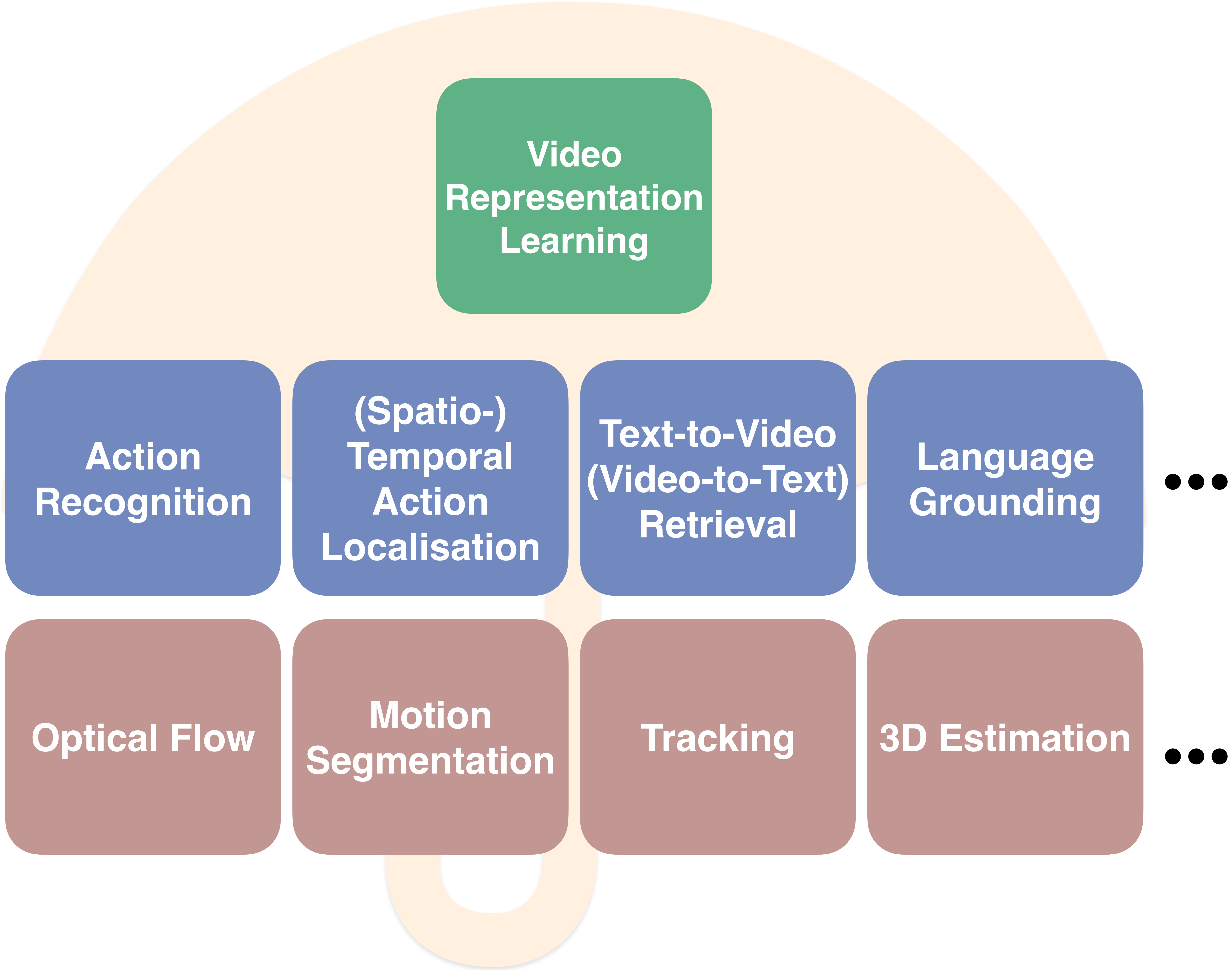
category: sign language interpreting

# Beyond semantics

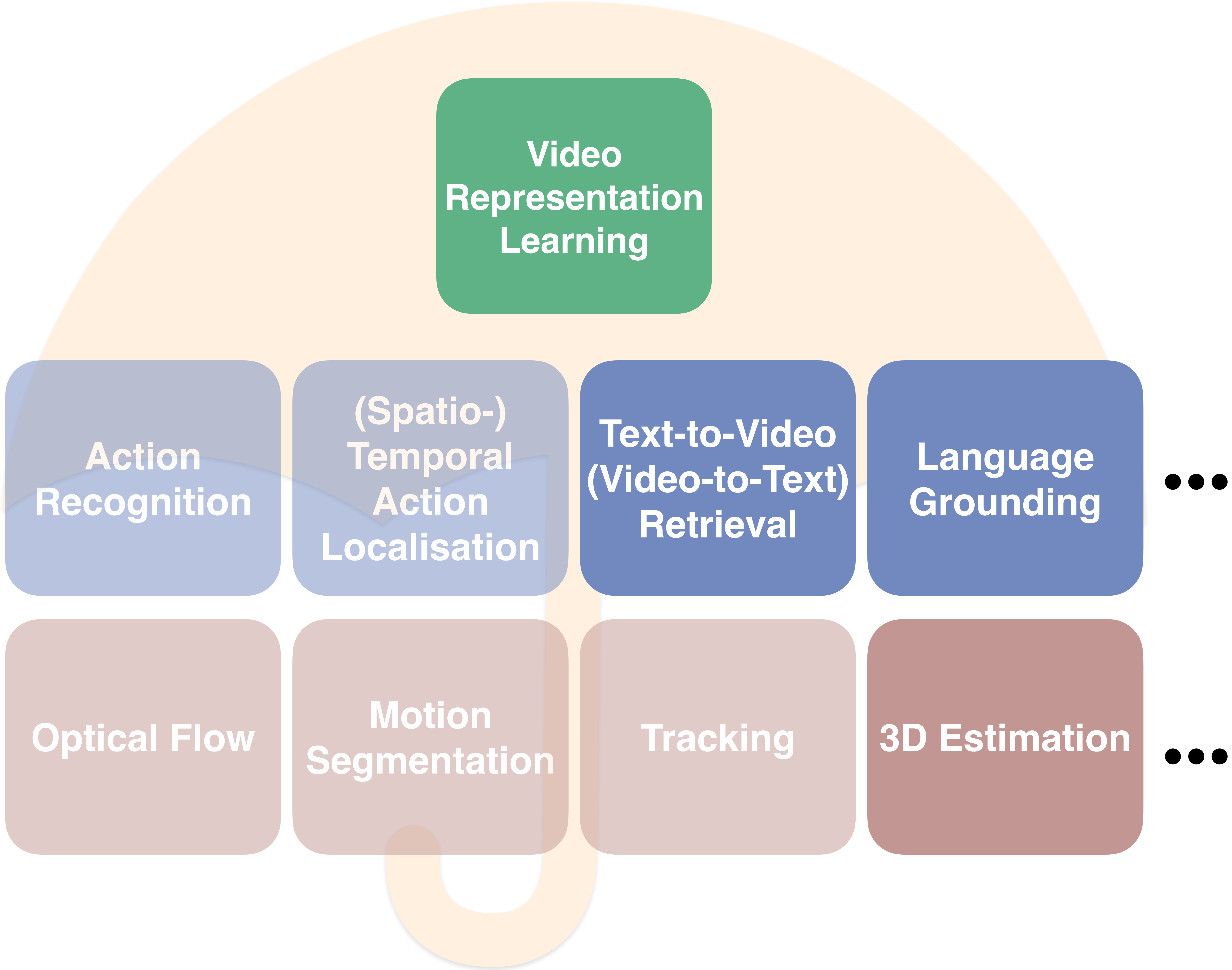
*Robots learning from human demonstrations*

category: washing dishes

# Video Modeling?



# Video Modeling?



# Talk Outline

**Text-to-Video Retrieval**

✦ 1) Text-to-video retrieval

[Bain et al. ICCV 2021]

**Sign Language Localisation**

✦ 2) Temporal localisation in sign language videos

[Varol et al. CVPR 2021] [Bull et al. ICCV 2021]

**3D Estimation**

✦ 3) Hand-object reconstruction from RGB videos

[Hasson et al. 3DV 2021]

# Talk Outline

Text-to-Video Retrieval

✦ 1) Text-to-video retrieval

[Bain et al. ICCV 2021]

Sign Language Localisation

✦ 2) Temporal localisation in sign language videos

[Varol et al. CVPR 2021] [Bull et al. ICCV 2021]



3D Estimation

✦ 3) Hand-object reconstruction from RGB videos

[Hasson et al. 3DV 2021]

2021 ICCV OCTOBER 11-17 VIRTUAL

# Frozen in Time:

## A Joint Video and Image Encoder For End-to-End Retrieval

ICCV 2021

https://www.robots.ox.ac.uk/~vgg/research/frozen-in-time/


Max Bain

Arsha Nagrani

Gül Varol

Andrew Zisserman

École des Ponts ParisTech

VGG University of Oxford

# Text-to-Video Retrieval

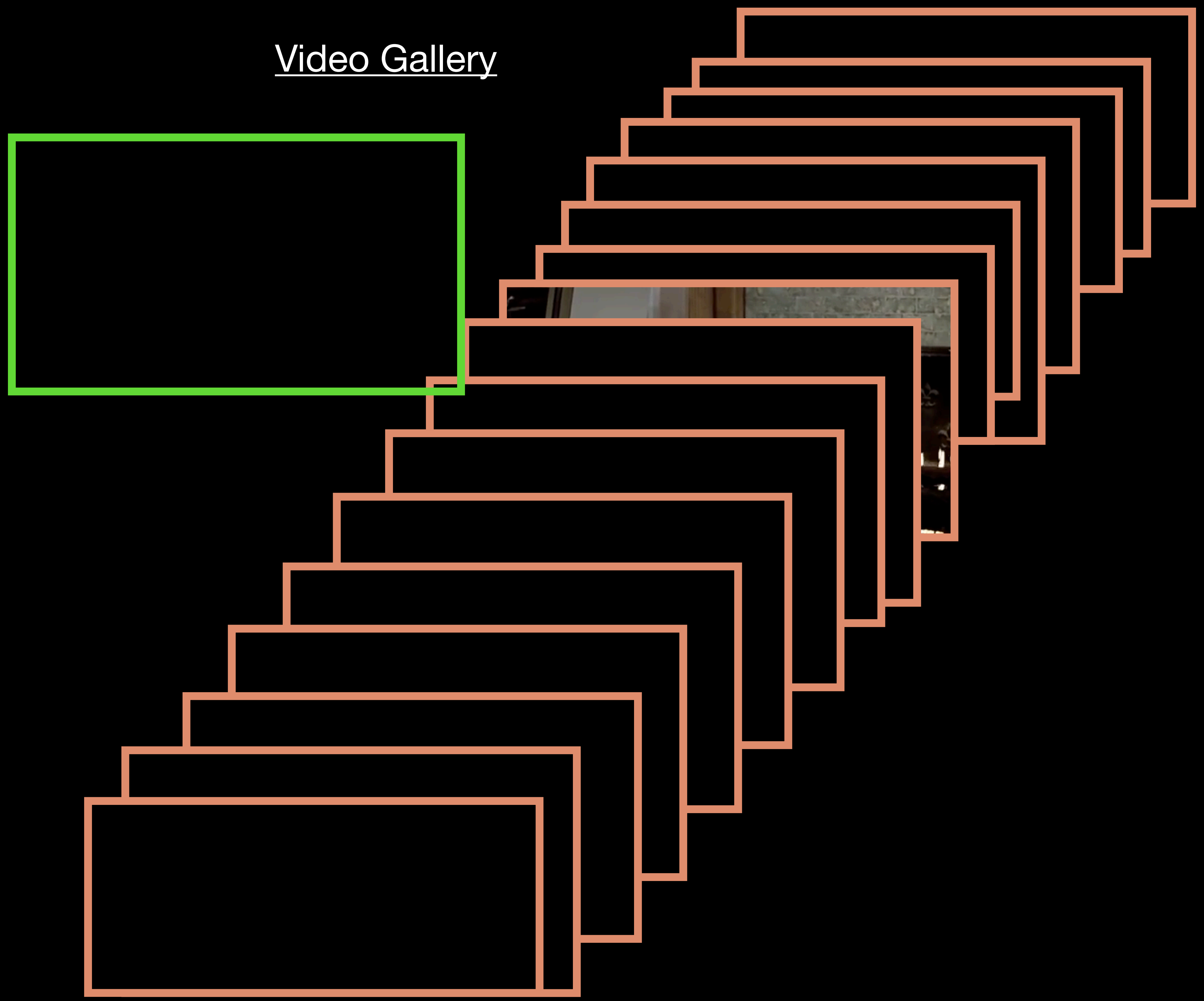
Text Query

Billy reveals the truth to Louis about the Duke's bet which changed both their lives

Text-video retrieval model

Similarity: **0.89**

Video Gallery



# Demo:

# ❄️ Frozen in Time ⏳

# 🔍 Video Search Demo 🎬

e.g. empty street in nepa Search

display: 8

Visual search of ~2.6M videos are based on research described in
Frozen in time: A joint video and image encoder for end-to-end retrieval.

# State of the art in video retrieval:

✘ Image and video retrieval progress largely disjoint

✘ Not end-to-end

- Pre-extracted “expert” features, typically trained on other tasks (ImageNet, Kinetics…)
- Performance limited and strongly linked to quality of features
- Experts typically not trained for vision & language space (MoEE[1], CE[2], MMT[3])

✘ Lack of large scale text-video data


[1] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. arXiv, 2018

[2] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. BMVC, 2019.

[3] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. ECCV, 2020.

# State of the art in video retrieval:

**This work:**

✗ Image and video retrieval progress largely disjoint

✓ Joint image-video training

✗ Not end-to-end

- Pre-extracted “expert” features, typically trained on other tasks (ImageNet, Kinetics…)
- Performance limited and strongly linked to quality of features
- Experts typically not trained for vision & language space (MoEE[1], CE[2], MMT[3])

✓ End-to-end video representation

✗ Lack of large scale text-video data

✓ Introduces WebVid-2M data


[1] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. arXiv, 2018

[2] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. BMVC, 2019.

[3] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. ECCV, 2020.

# State of the art in video representations:

✘ Image and video representations progress largely disjoint

✘ 3D Spatio-temporal Convolutions (I3D, ResNet3D, S3D, X3D…)

Either:

✘ Trained for action classification on Kinetics dataset (e.g., X3D [1]) - closed set of categories

✘ Trained for retrieval on HowTo100M dataset (e.g., MIL-NCE [2]) - noisy speech data, long training

✘ Trained with “self-supervision” (e.g., [3]) - requires audio or other modality


[1] Christoph Feichtenhofer. X3D: Expanding architectures for efficient video recognition. CVPR, 2020.

[2] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. CVPR, 2020.

[3] Yuki M. Asano*, Mandela Patrick*, Christian Rupprecht, Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. NeurIPS , 2020.

# State of the art in video representations:

**This work:**

- ✗ Image and video representations progress largely disjoint
- ✗ 3D Spatio-temporal Convolutions (I3D, ResNet3D, S3D, X3D…)
  - ✓✓ Joint image-video training via Transformer encoder

Either:

- ✗ Trained for action classification on Kinetics dataset (e.g., X3D [1]) - closed set of categories
- ✗ Trained for retrieval on HowTo100M dataset (e.g., MIL-NCE [2]) - noisy speech data, long training
- ✗ Trained with “self-supervision” (e.g., [3]) - requires audio or other modality
  - ✓✓✓ Trained for retrieval efficiently on WebVid-2M dataset


[1] Christoph Feichtenhofer. X3D: Expanding architectures for efficient video recognition. CVPR, 2020.

[2] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. CVPR, 2020.

[3] Yuki M. Asano*, Mandela Patrick*, Christian Rupprecht, Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. NeurIPS , 2020.

# Frozen in Time

## A Joint Video and Image Encoder

In this work aim to overcome these with:

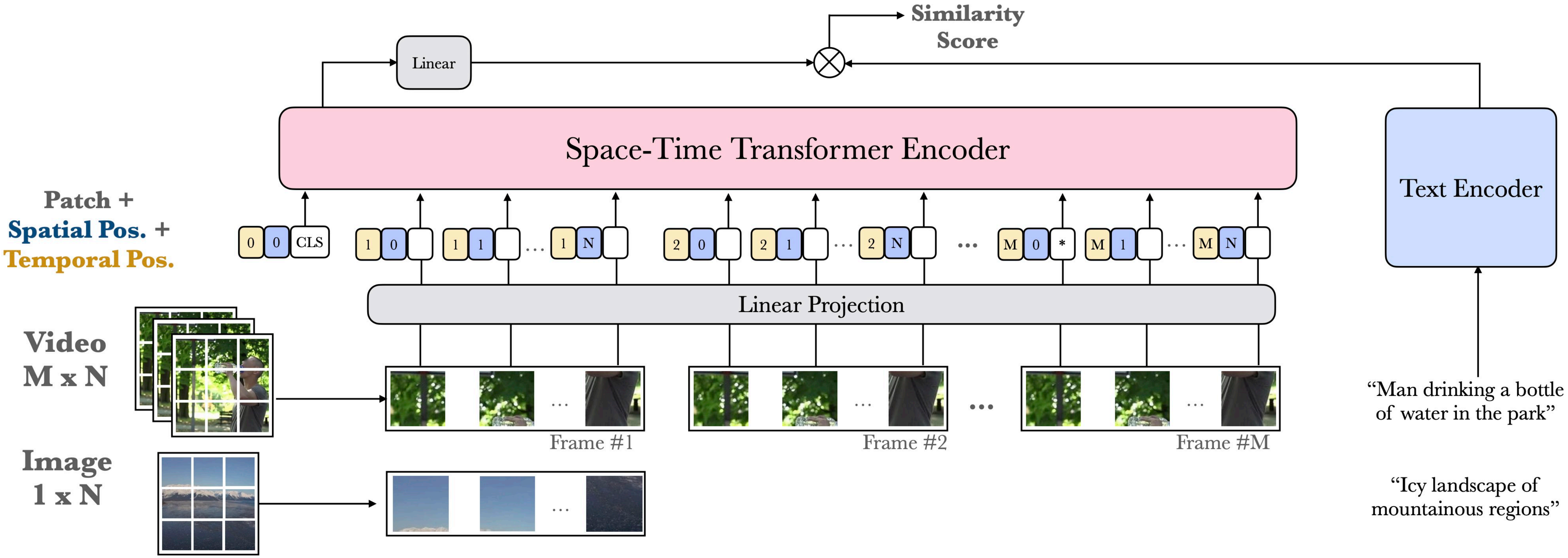
✓ **End-to-end** retrieval model (from pixels)

✓ **Jointly training** on image-text and video-text pairs

- Using a **Transformer** encoder that accepts a variable-length sequence
- Treating images as 1-frame videos, *frozen in time*

✓ Collecting a **large-scale video-text dataset**, WebVid-2M, for pretraining

# End-to-end retrieval

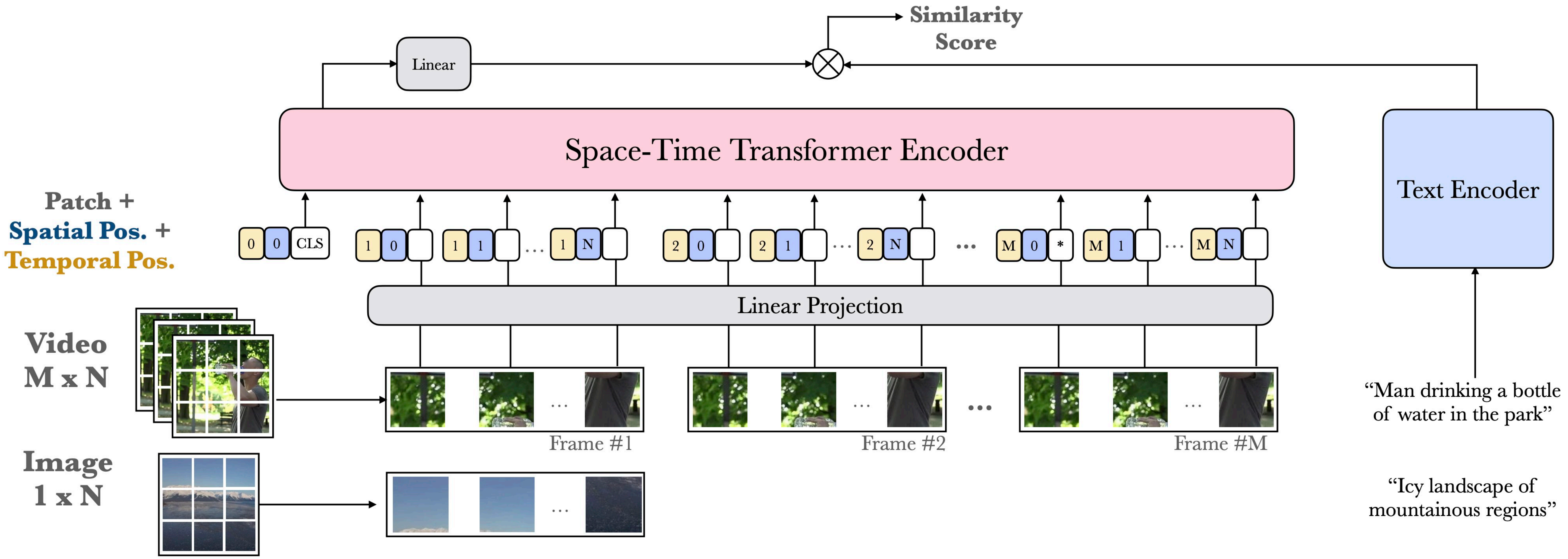



- Dual encoder for efficient retrieval

# End-to-end retrieval

**Video** encoder:

- inspired from Timesformer [1]
- initialized from ViT [2] weights pretrained on ImageNet
  - Temporal embeddings zero-initialized



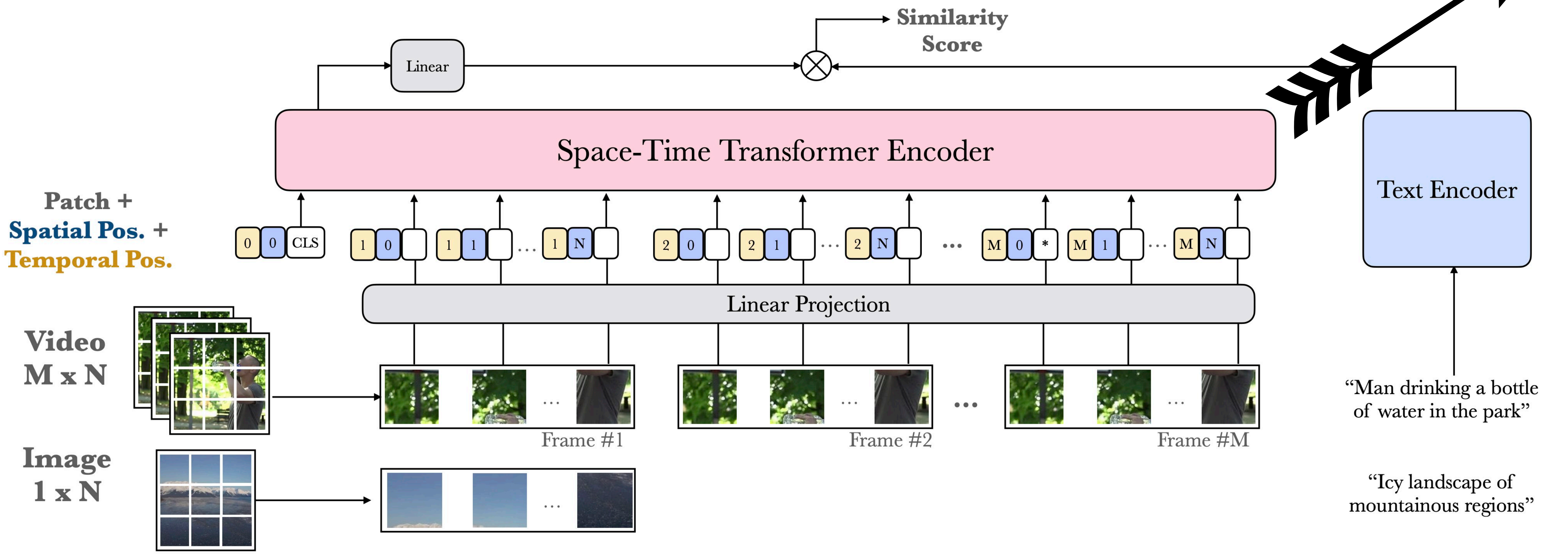


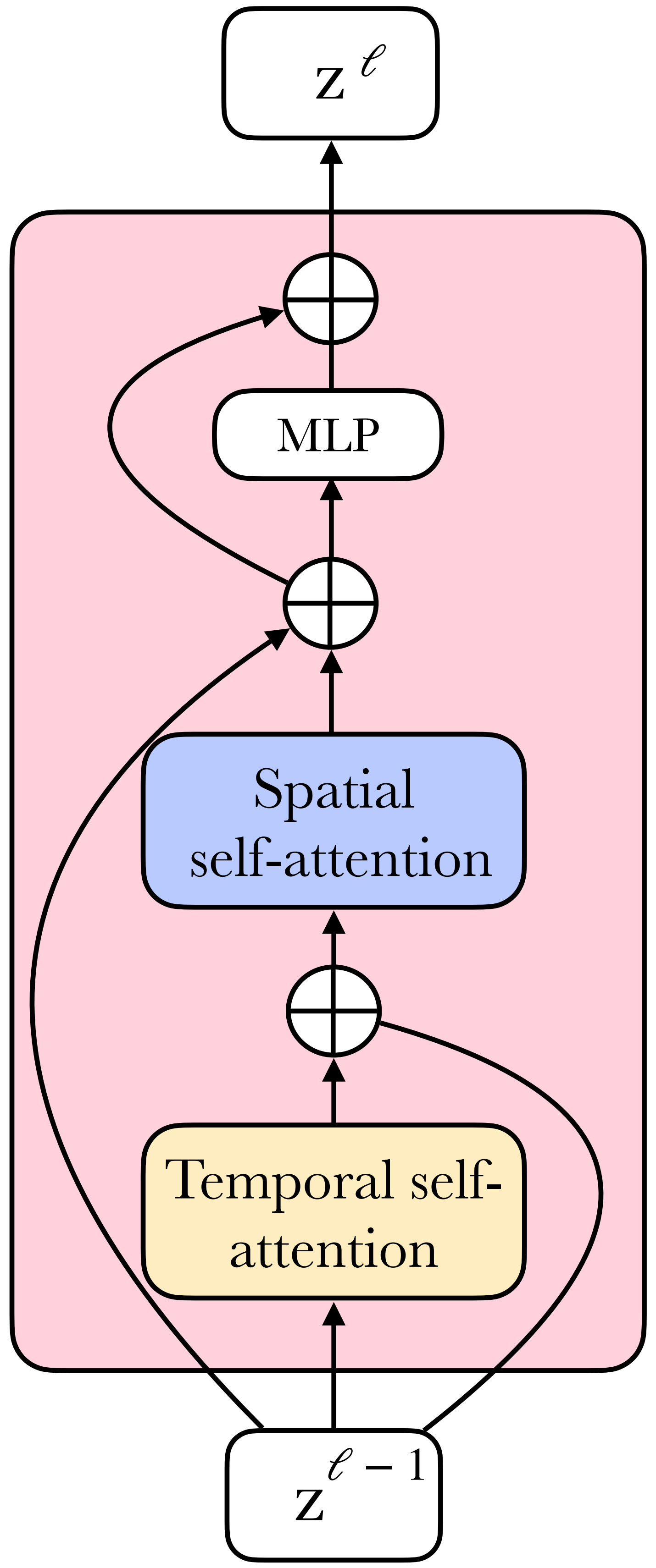
[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? arXiv, 2021.
[2] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR, 2021.

# End-to-end retrieval

**Video** encoder:

- inspired from Timesformer [1]
- initialized from ViT [2] weights pretrained on ImageNet
  - Temporal embeddings zero-initialized



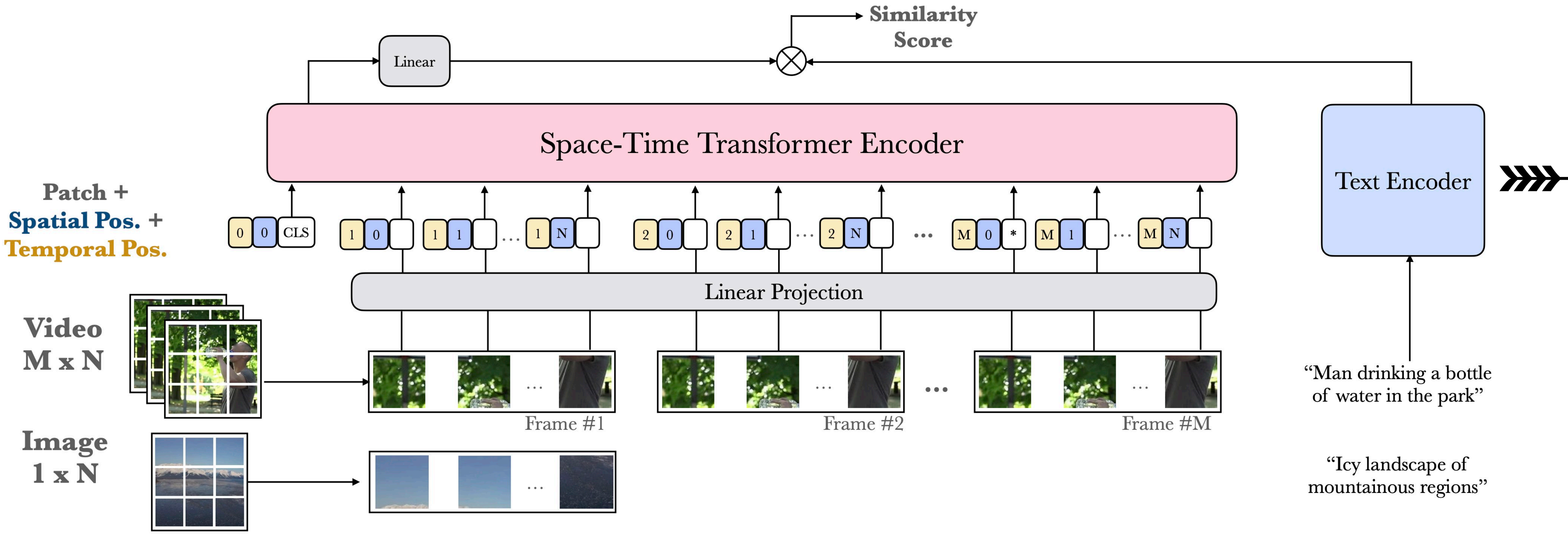




[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? arXiv, 2021.
[2] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR, 2021.

# End-to-end retrieval

**Text** encoder:

- initialized from DistilBERT [1]





DistilBERT

[1] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv, 2019.

# WebVid-2M Dataset

## 2.5M video-text pairs scraped from the web

“Runners feet in a sneakers close up. realistic three dimensional animation.”

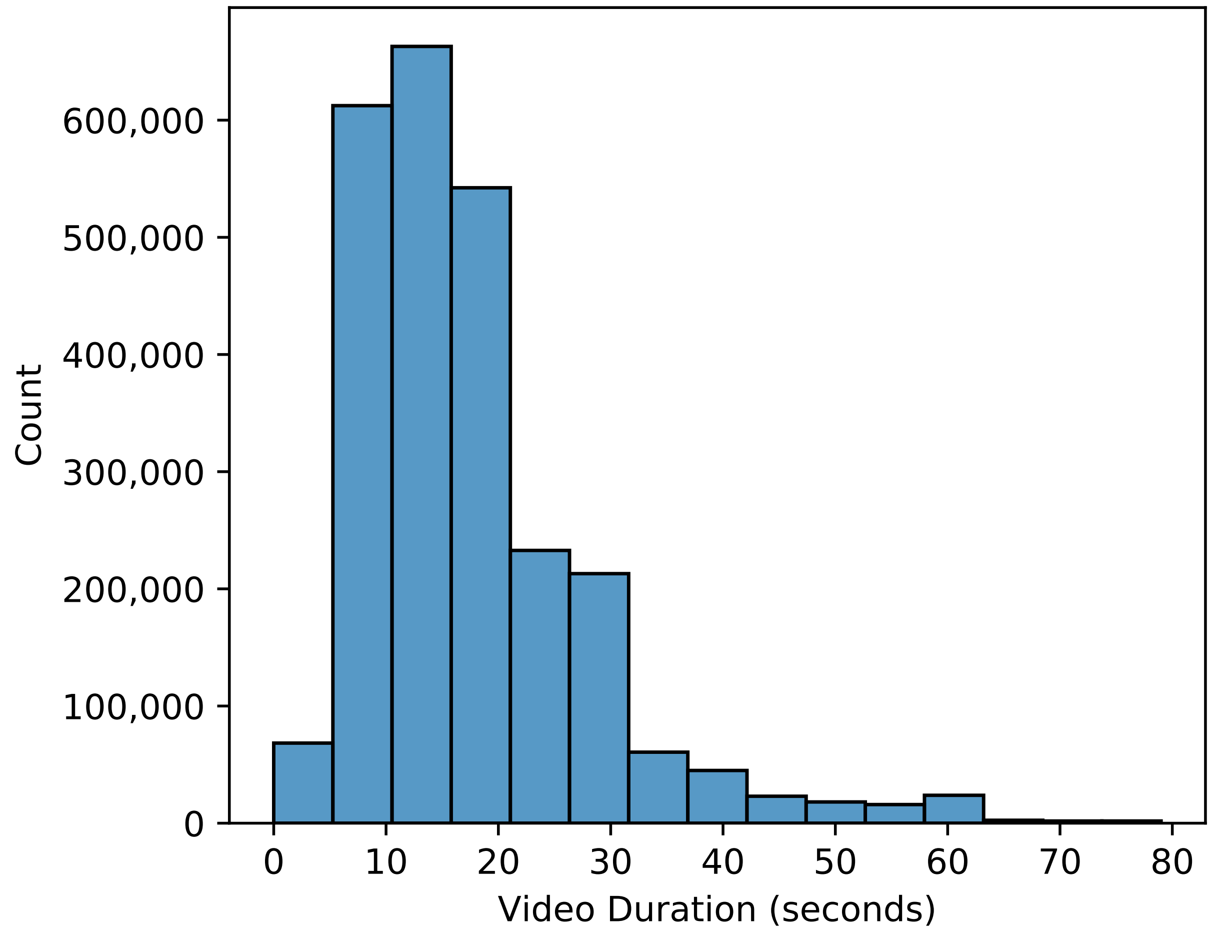
“Female cop talking on walkietalkie, responding emergency call, crime prevention”

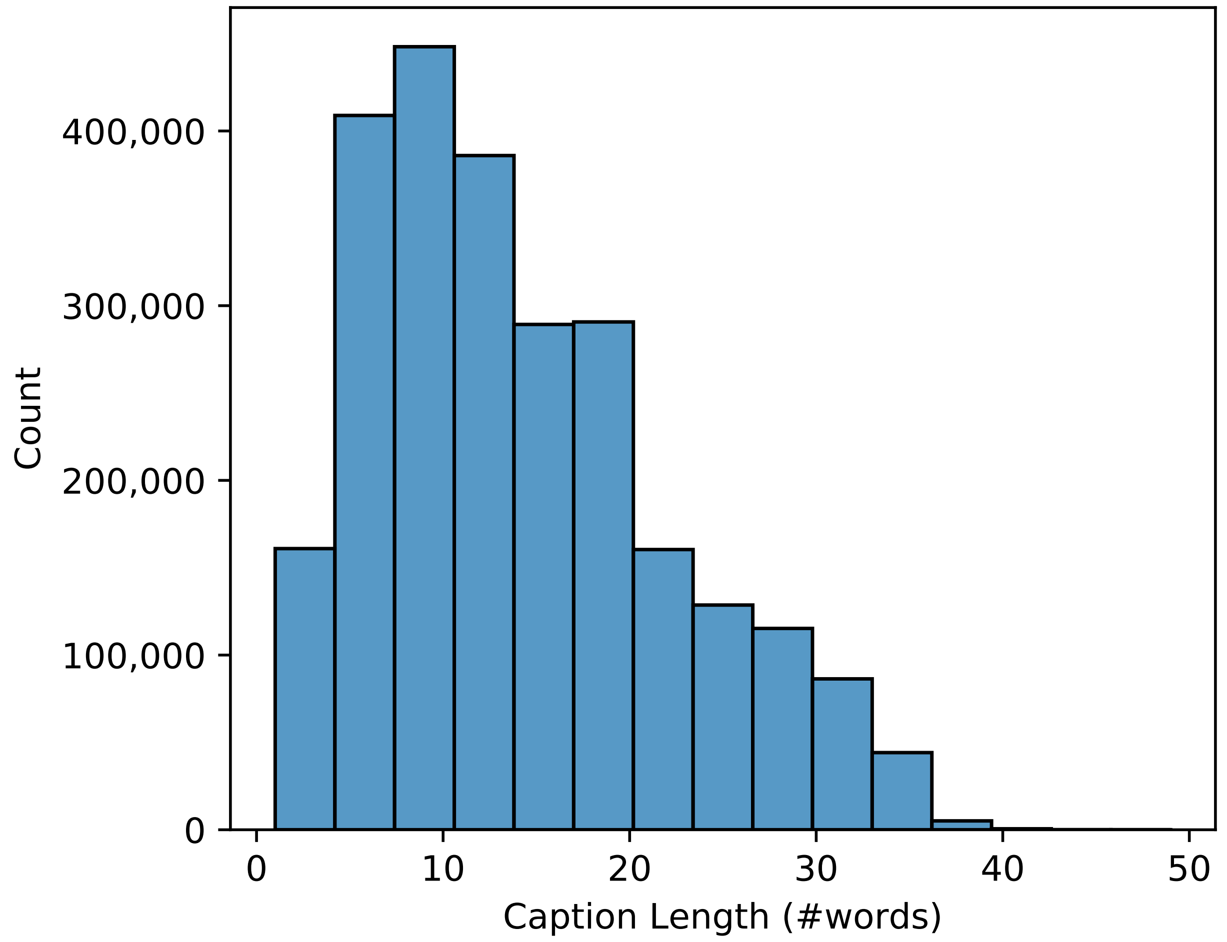
“Billiards, concentrated young woman playing in club”

“Lonely beautiful woman sitting on the tent looking outside. wind on the hair and camping on the beach near the colors of water and shore. freedom and alternative tiny house for traveler lady drinking”

“Kherson, ukraine - 20 may 2016: open, free, rock music festival crowd partying at a rock concert. hands up, people, fans cheering clapping applauding in kherson, ukraine - 20 may 2016. band performing”

“Cabeza de toro, punta cana/ dominican republic - feb 20, 2020: 4k drone flight over coral reef with manta"



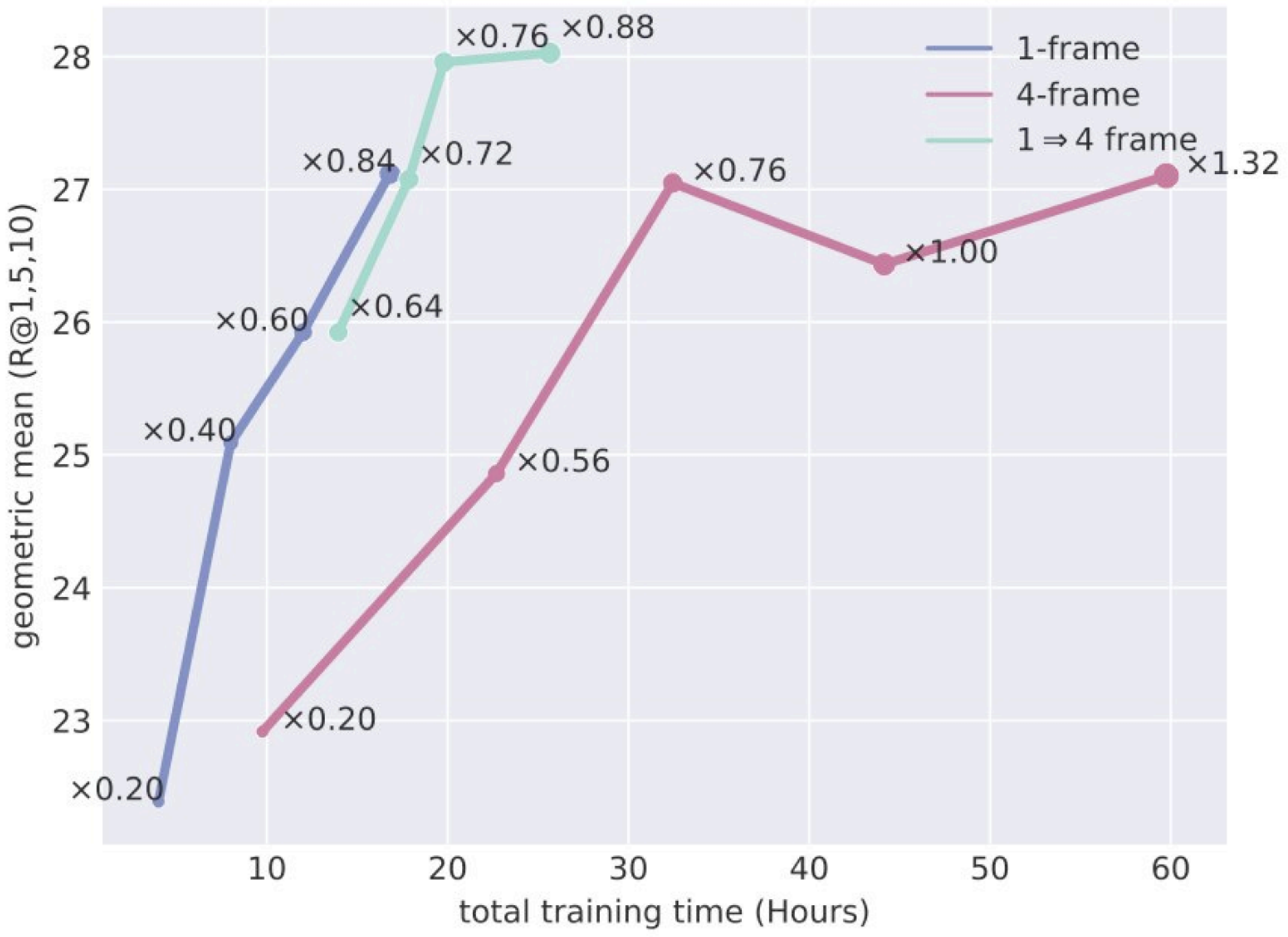
# Effect of pretraining

| Pre-training (for 1 epoch) | #pairs | ↑R@1 | ↑R@10 | ↓MedR |
|---|---|---|---|---|
| - | - | 5.6 | 22.3 | 55 |
| ImageNet | | 15.2 | 54.4 | 9.0 |
| HowTo-17M subset | 17.1M | 24.1 | 63.9 | 5.0 |
| CC3M | 3.0M | 24.5 | 62.7 | 5.0 |
| WebVid2M | 2.5M | 26.0 | 64.9 | 5.0 |
| **CC3M + WebVid2M** | 5.5M | **27.3** | **68.1** | **4.0** |

MSRVTT Benchmark

# Efficient training with curriculum learning

- Transformers are costly to train, and scale with the length of the sequence.
- We investigate curriculum learning:
    - Similar / better performance in much less training time
    - Initially train with only one frame
    - Gradually increase



# Comparison to the state of the art

## MSRVTT benchmark

| Method | E2E† | Vis Enc. Init. | Visual-Text PT | #pairs PT | R@1 | R@5 | R@10 | MedR |
|---|---|---|---|---|---|---|---|---|
| JSFusion [62] | ✓ | - | - | - | 10.2 | 31.2 | 43.2 | 13.0 |
| HT MIL-NCE [35] | ✓ | - | HowTo100M | 100M | 14.9 | 40.2 | 52.8 | 9.0 |
| ActBERT [67] | ✓ | VisGenome | HowTo100M | 100M | 16.3 | 42.8 | 56.9 | 10.0 |
| HERO [27] | ✓ | ImageNet, Kinetics | HowTo100M | 100M | 16.8 | 43.4 | 57.7 | - |
| VidTranslate [22] | ✓ | IG65M | HowTo100M | 100M | 14.7 | - | 52.8 | |
| NoiseEstimation [2] | ✗ | ImageNet, Kinetics | HowTo100M | 100M | 17.4 | 41.6 | 53.6 | 8.0 |
| CE [29] | ✗ | Numerous experts† | - | | 20.9 | 48.8 | 62.4 | 6.0 |
| UniVL [31] | ✗ | - | HowTo100M | 100M | 21.2 | 49.6 | 63.1 | 6.0 |
| ClipBERT [25] | ✓ | - | COCO, VisGenome | 5.6M | 22.0 | 46.8 | 59.9 | 6.0 |
| AVLnet [44] | ✗ | ImageNet, Kinetics | HowTo100M | 100M | 27.1 | 55.6 | 66.6 | 4.0 |
| MMT [15] | ✗ | Numerous experts† | HowTo100M | 100M | 26.6 | 57.1 | 69.6 | 4.0 |
| Support Set [39] | ✗ | IG65M, ImageNet | - | - | 27.4 | 56.3 | 67.7 | 3.0 |
| Support Set [39] | ✗ | IG65M, ImageNet | HowTo100M | 100M | 30.1 | 58.5 | 69.3 | **3.0** |
| **Ours** | ✓ | ImageNet | CC3M | 3M | 25.5 | 54.5 | 66.1 | 4.0 |
| **Ours** | ✓ | ImageNet | CC3M, WebVid-2M | 5.5M | **31.0** | **59.5** | **70.5** | **3.0** |
| **Zero-shot** | | | | | | | | |
| HT MIL-NCE [35] | ✓ | - | HowTo100M | 100M | 7.5 | 21.2 | 29.6 | 38.0 |
| **Ours** | ✓ | ImageNet | CC3M, WebVid-2M | 5.5M | **18.7** | **39.5** | **51.6** | **10.0** |

# Take-home messages

- An **end-to-end** trained video retrieval model can outperform "expert" feature models, even without multiple modalities such as audio.
- Video Transformers, with their flexible input sequence sizes, can benefit from **joint image-video** training, exploiting cheaper image-text datasets.
- **Curriculum** in sequence length (time) achieves competitive performance with far less compute.

# Talk Outline

Text-to-Video Retrieval

✦ 1) Text-to-video retrieval

[Bain et al. ICCV 2021]

Sign Language Localisation

✦ 2) Temporal localisation in sign language videos

[Varol et al. CVPR 2021] [Bull et al. ICCV 2021]

3D Estimation

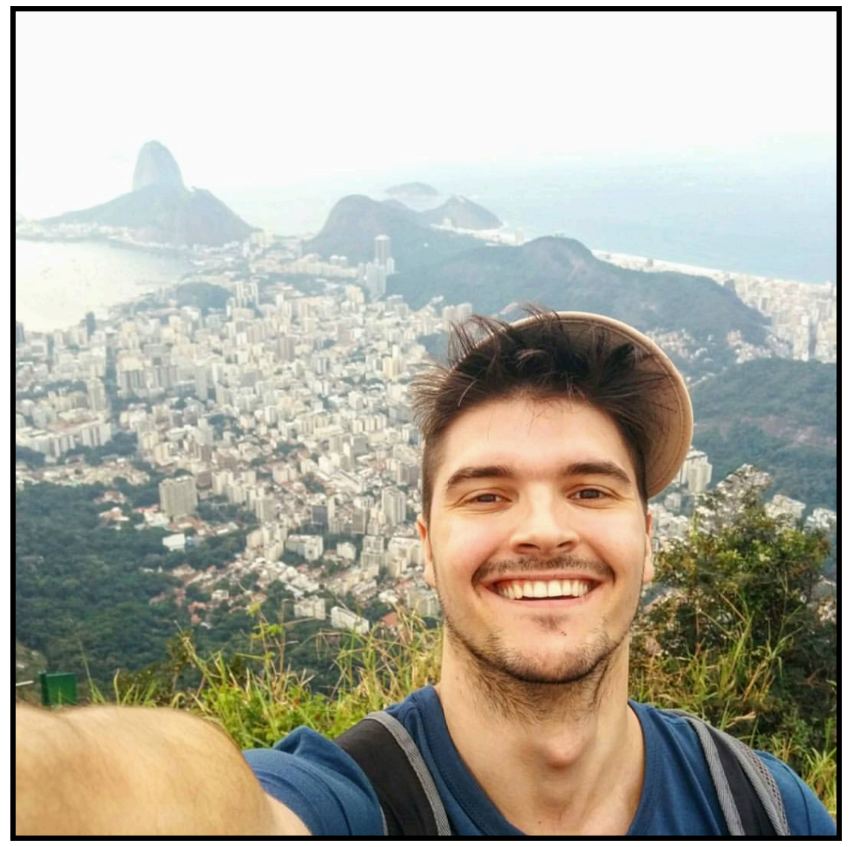
✦ 3) Hand-object reconstruction from RGB videos

[Hasson et al. 3DV 2021]

# Read and Attend:
## Temporal Localisation in Sign Language Videos

CVPR 2021

https://www.robots.ox.ac.uk/~vgg/research/bslattend/

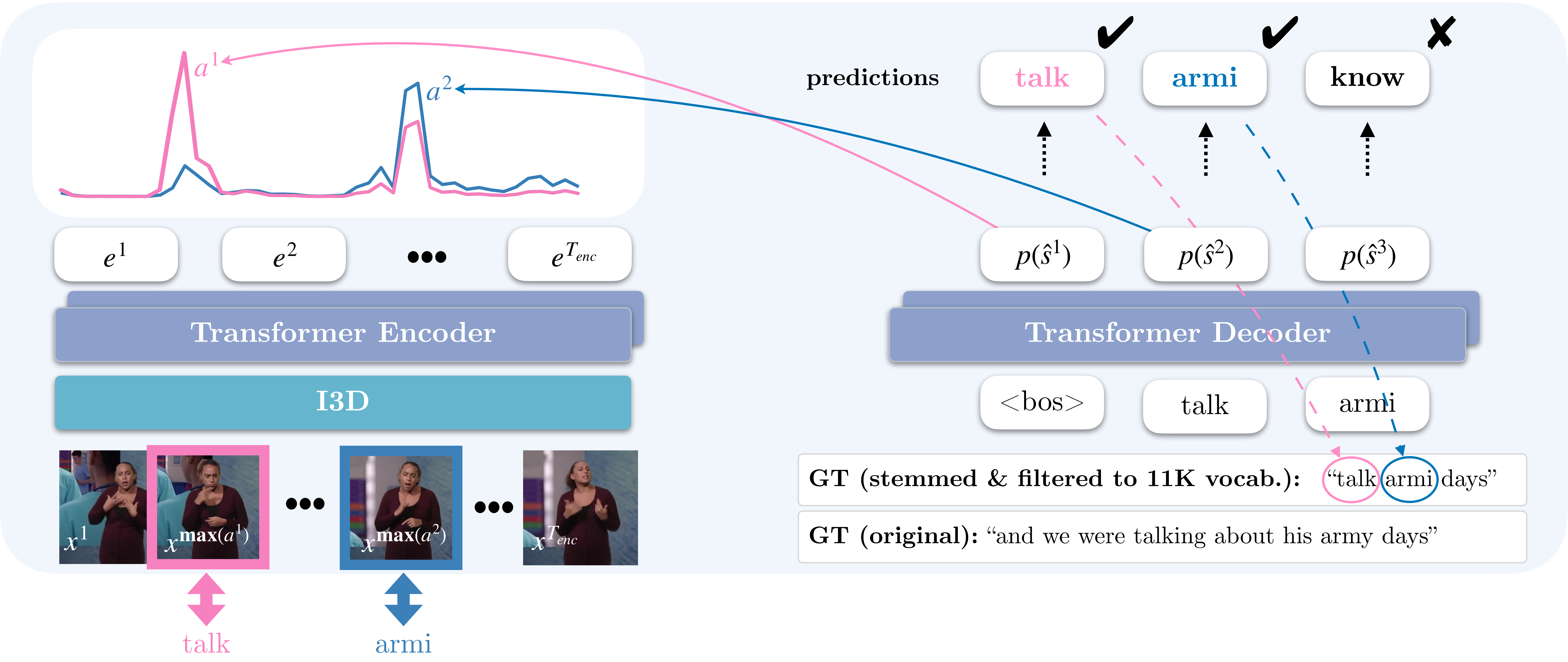


Gül Varol* Liliane Momeni* Samuel Albanie* Triantafyllos Afouras* Andrew Zisserman

*equal contribution

# Localising Words

- ➡ We train a **large-vocabulary** video-to-text Transformer model.
- ➡ **Translation** performance of this sequence prediction task is **low** (~20% recall).
- ➡ However, a **localisation** ability emerges from the **attention** mechanism.

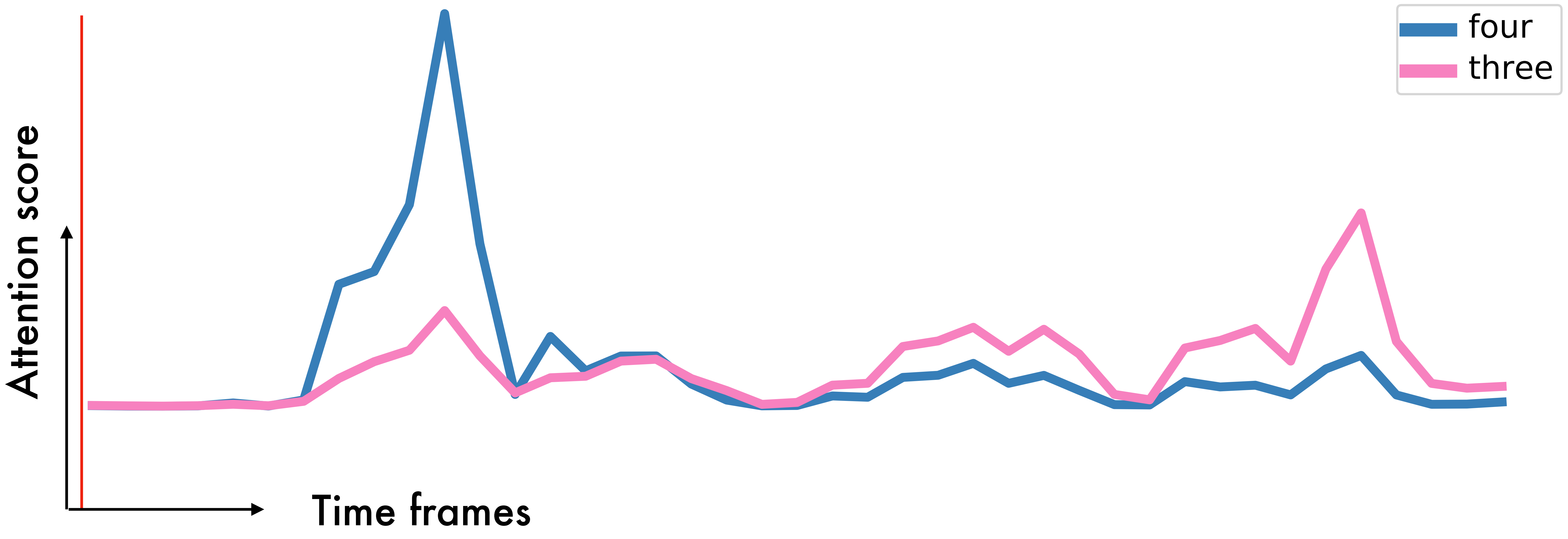


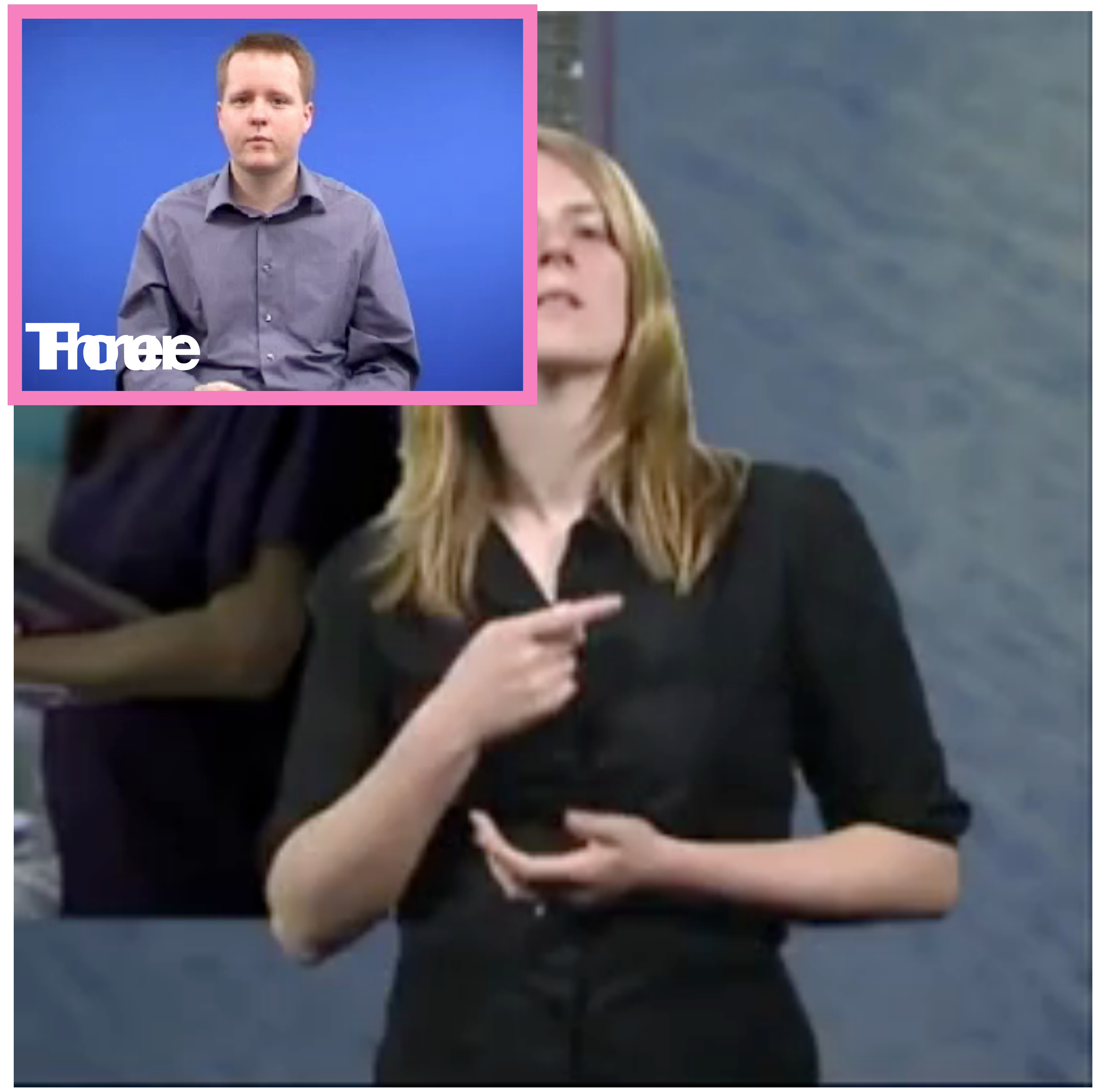
# Attention visualisations

**GT (original)** : "Mr Hanssen said three you gave him four"

**GT (processed)** : "mr said three gave four"

**Prediction** : "four week ago three"



**We show an example of the ground-truth sign from a dictionary to aid visual assessment.*

# Mining automatic annotations for sign recognition

**Sign: “World”**

# Localising Sentences

- ➡ We localised **individual** words that correspond to signs in the videos.
- ➡ Can we localise **multiple** words, i.e., phrases/sentences?

2021 ICCV OCTOBER 11-17 VIRTUAL

cnrs École des Ponts ParisTech VGG University of Oxford

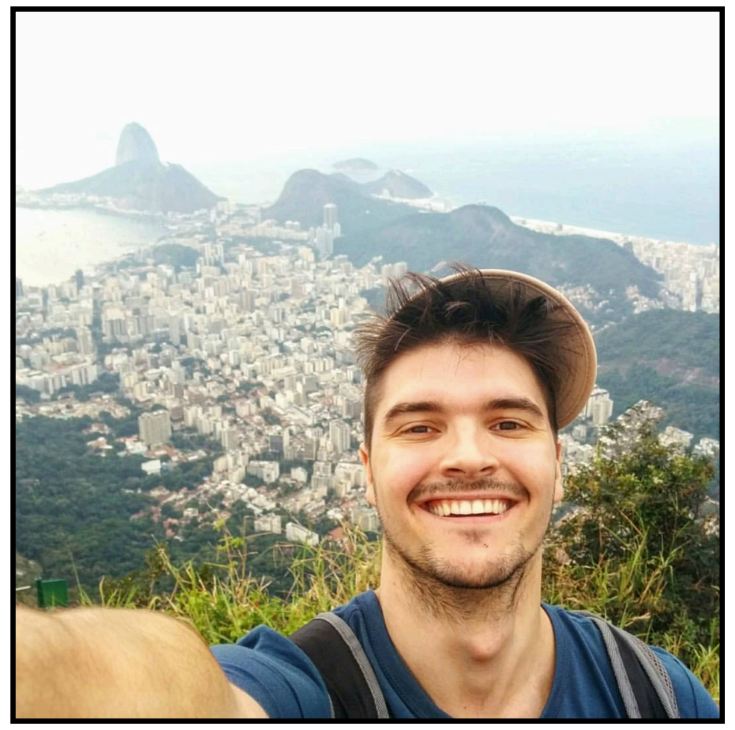
# Aligning Subtitles in Sign Language Videos

ICCV 2021

https://www.robots.ox.ac.uk/~vgg/research/bslalign/



Hannah Bull*


Triantafyllos Afouras*


Gül Varol


Samuel Albanie

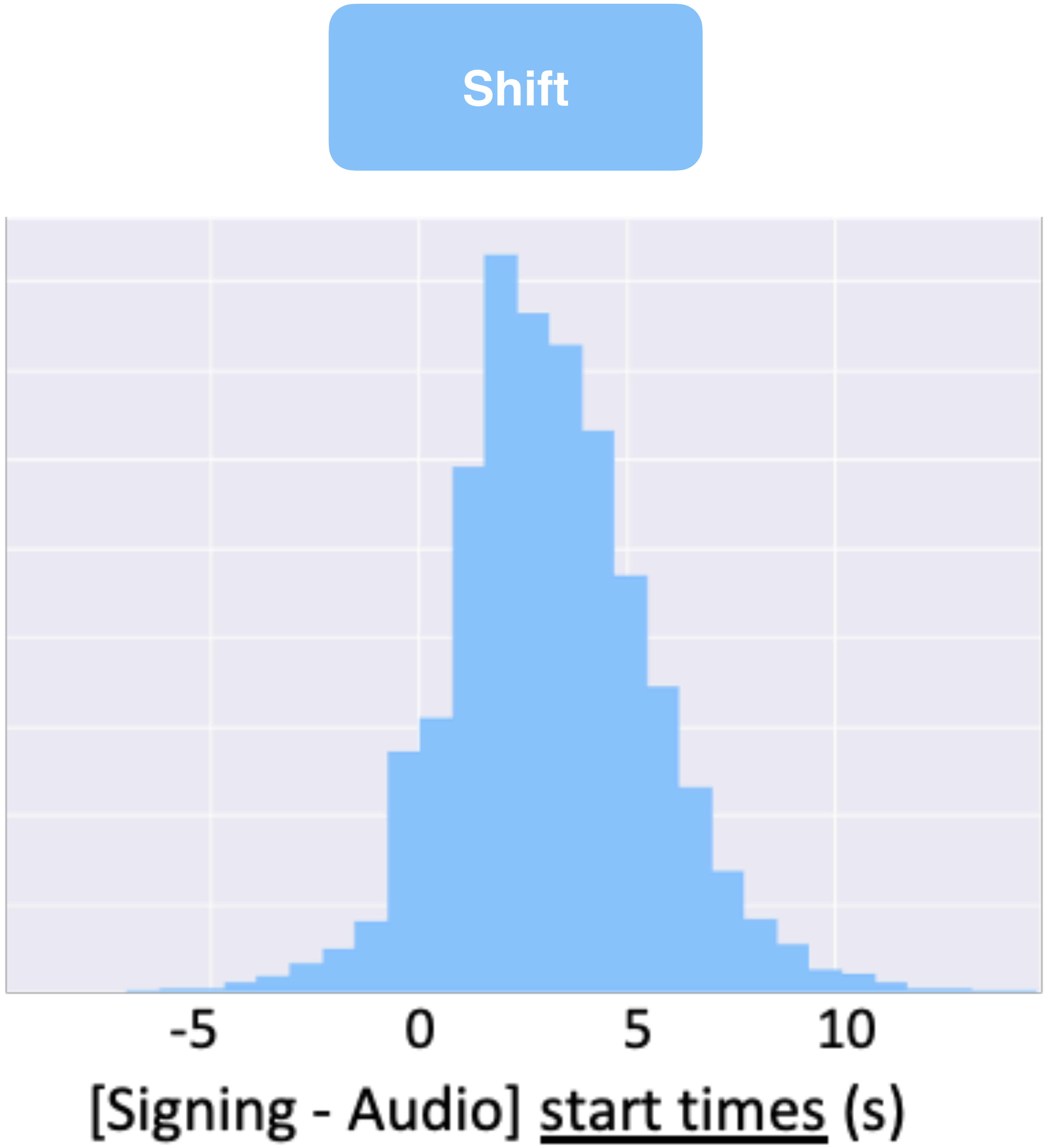

Liliane Momeni

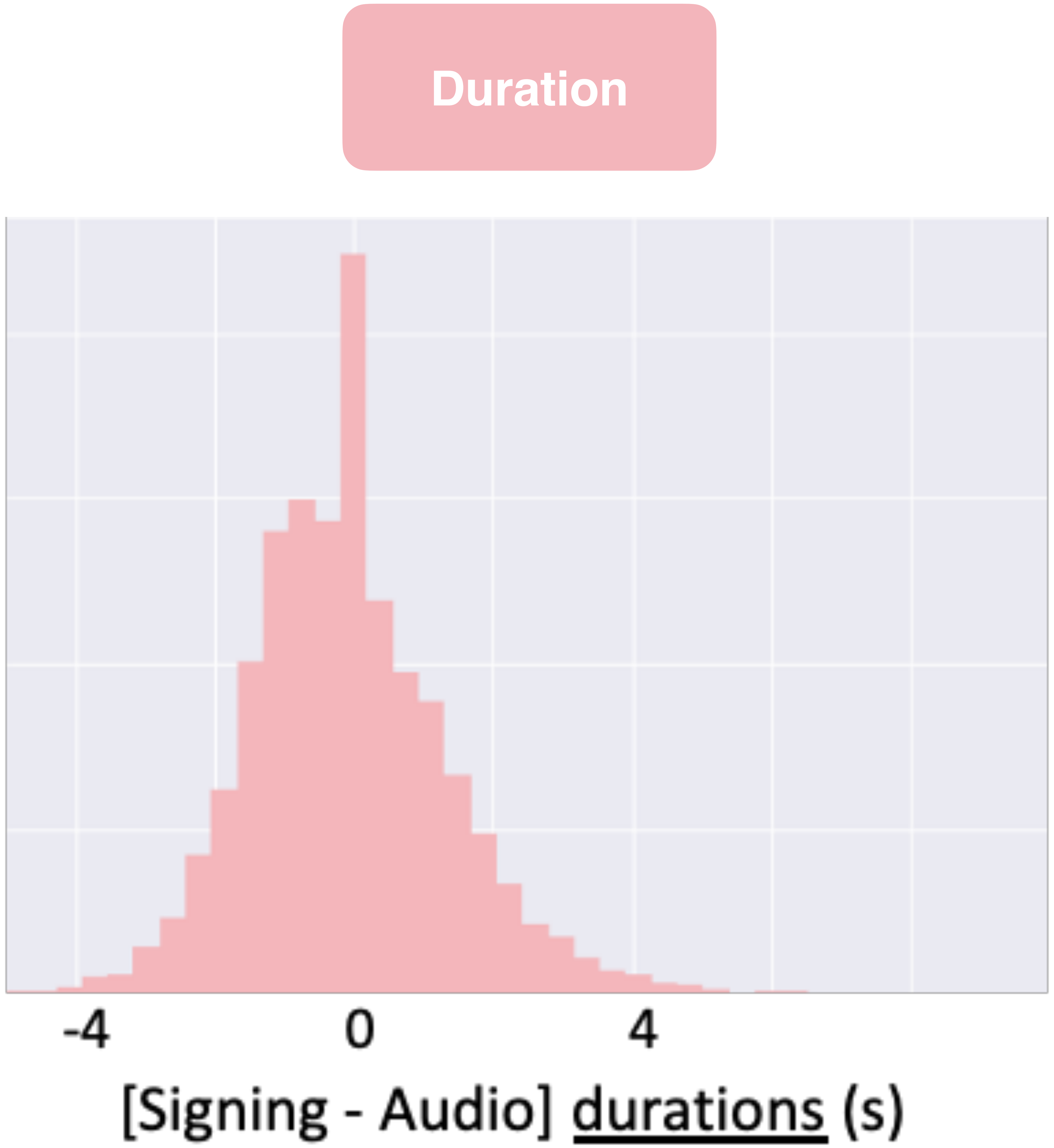

Andrew Zisserman

*equal contribution

# Problem formulation: Subtitle alignment

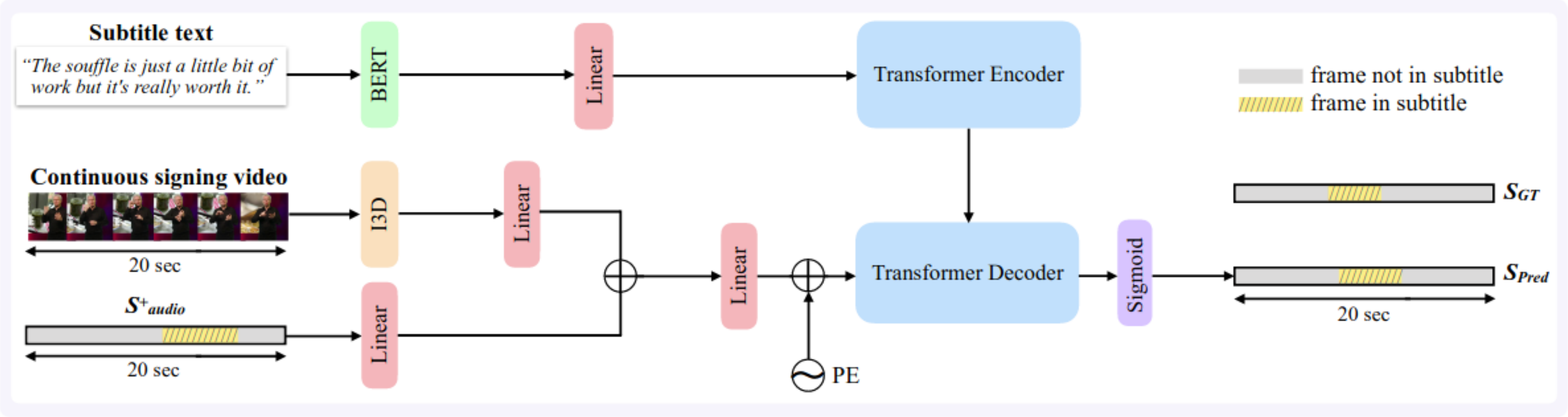
# Difference between speech- and sign-aligned subtitles



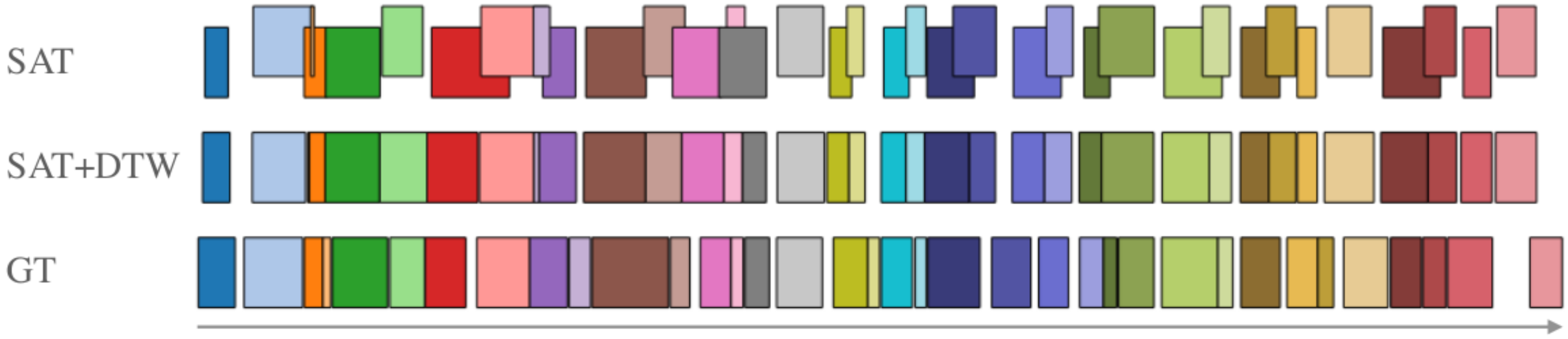
# Subtitle Aligner Transformer (SAT)

- Single subtitle
- "Inverted" Transformer

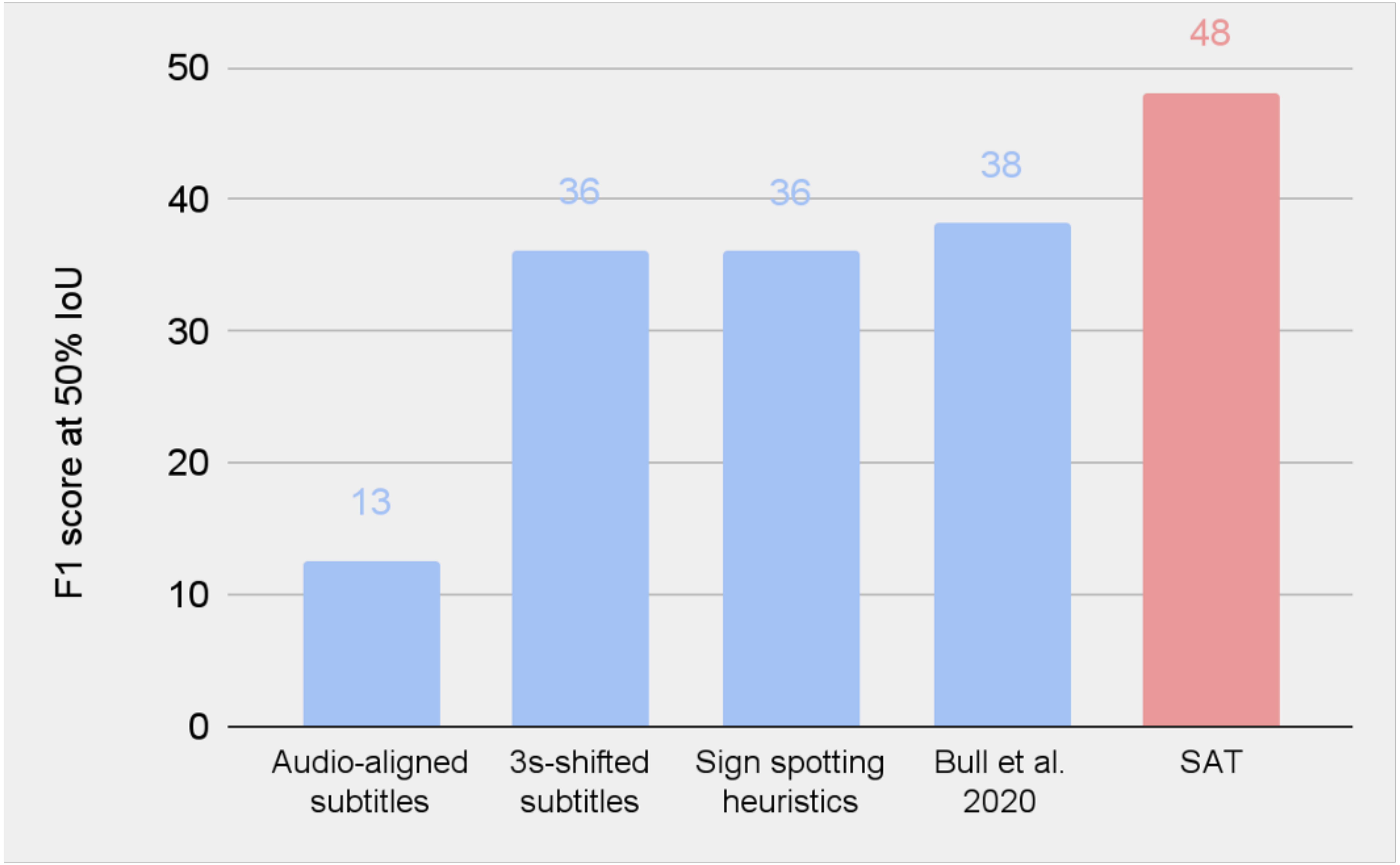


# Global alignment with DTW

- Multiple subtitles

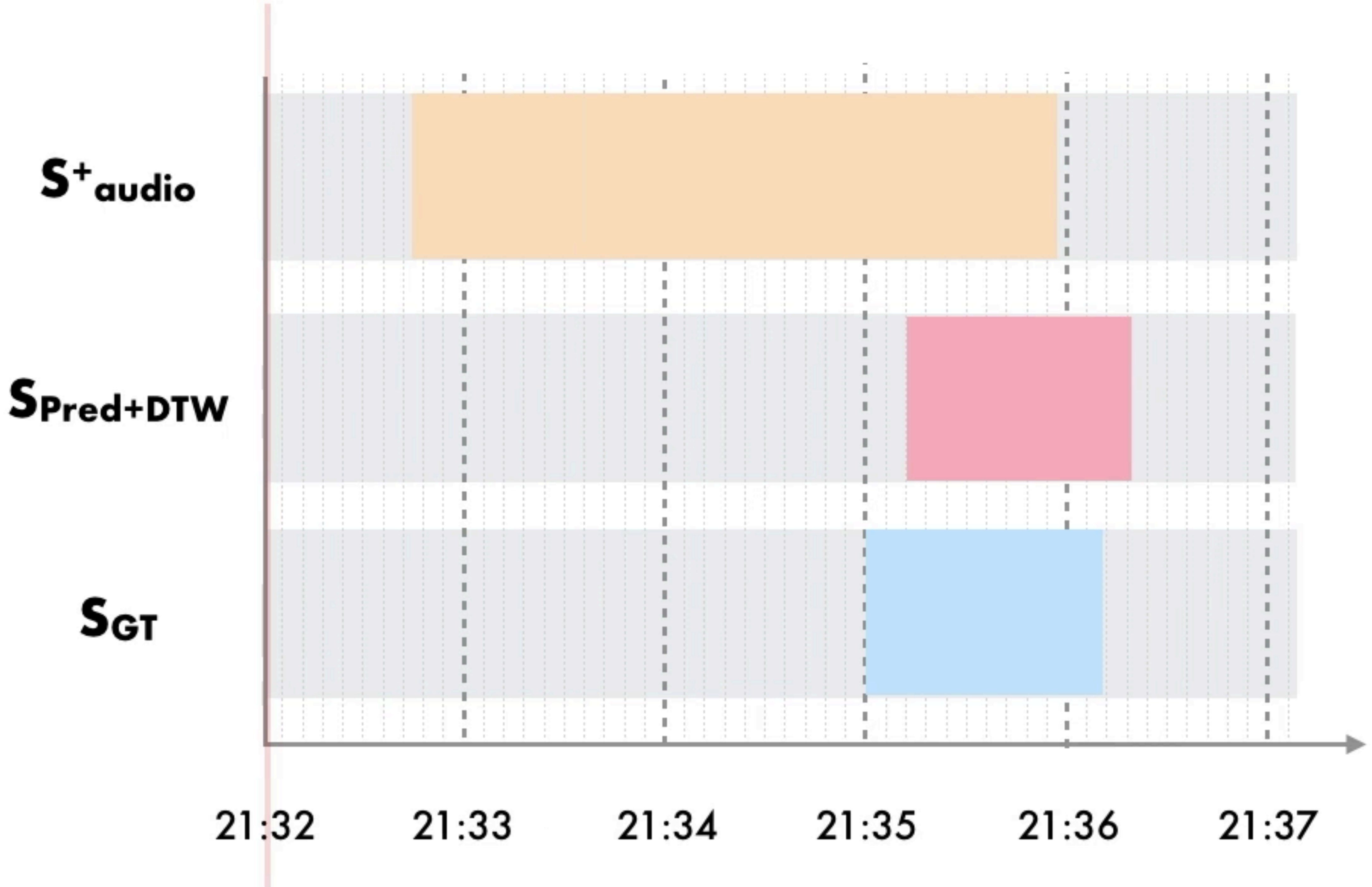


# Main results



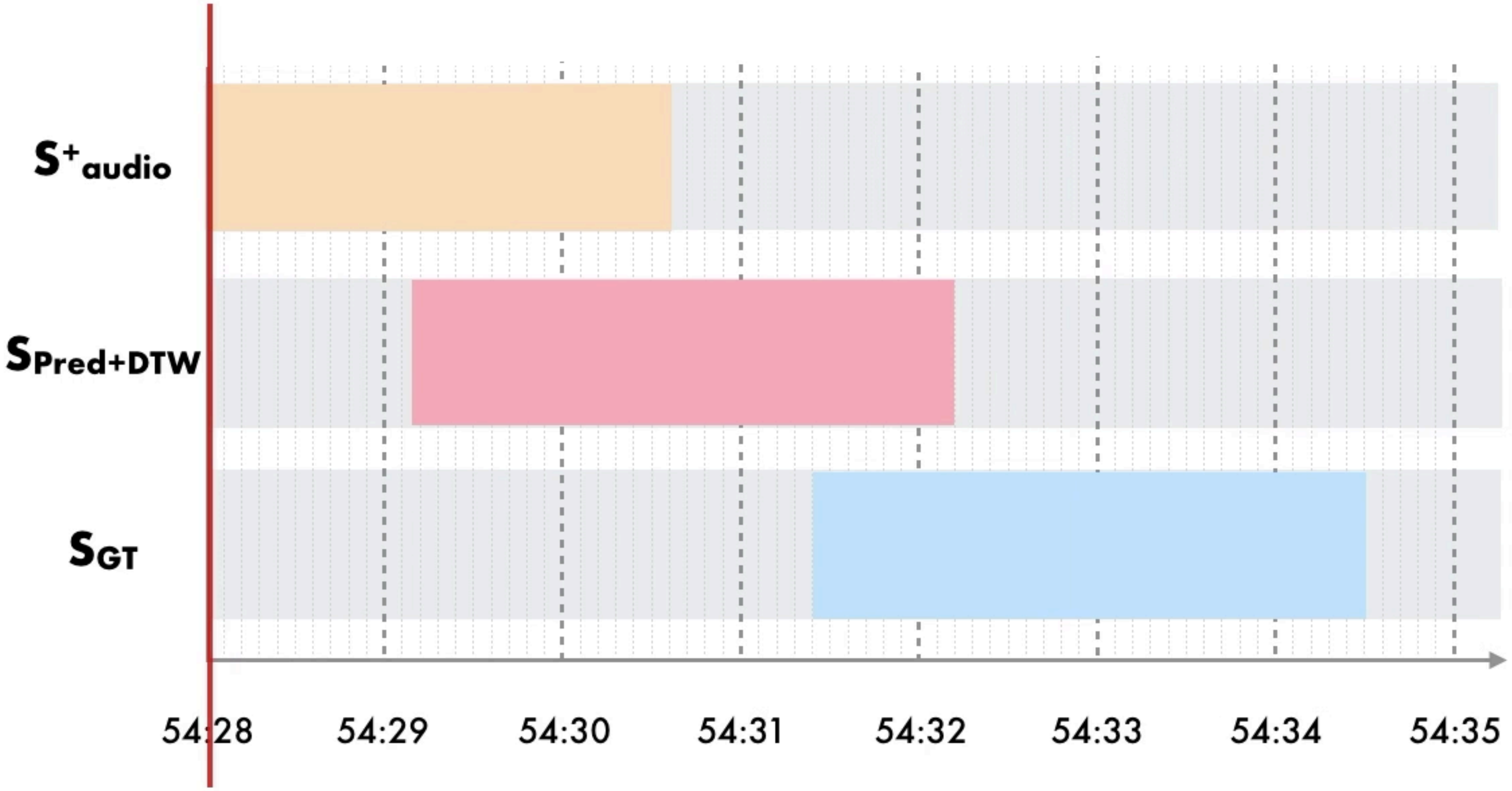
# Qualitative results

**Subtitle text:** "But when it tastes this good, who cares?"




# Qualitative results

**Subtitle text:** "It's the fate of the weakening daughter."




# Talk Outline

Text-to-Video Retrieval

✦ 1) Text-to-video retrieval

[Bain et al. ICCV 2021]

Sign Language Localisation

✦ 2) Temporal localisation in sign language videos

[Varol et al. CVPR 2021] [Bull et al. ICCV 2021]



3D Estimation

✦ 3) Hand-object reconstruction from RGB videos

[Hasson et al. 3DV 2021]

3DV 2021

# Towards unconstrained hand-object reconstruction from RGB videos


Yana Hasson

Gül Varol

Cordelia Schmid

Ivan Laptev

Inria

École des Ponts ParisTech


https://hassony2.github.io/homan.html

# Towards unconstrained reconstruction from RGB videos

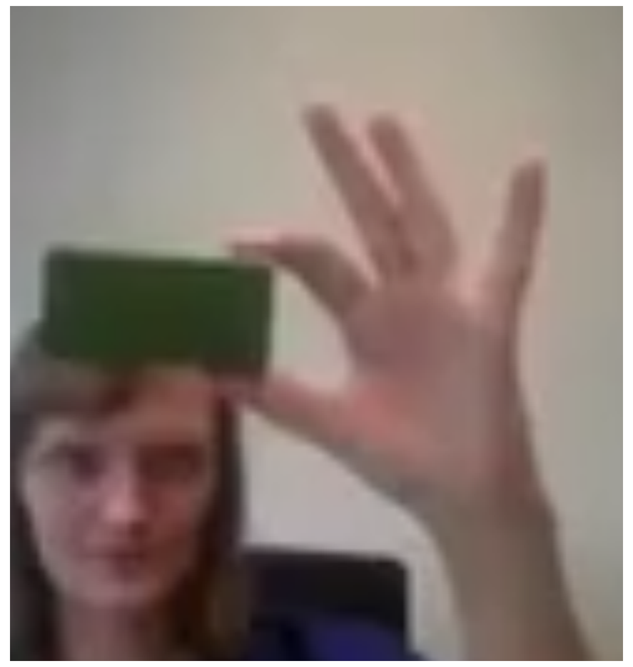
# *In-the-wild* hand-object interactions


[Epic Kitchens-100](), TPAMI 2020

# Limited datasets with 3D annotations

**Hands In Action**
ICCV 2015

**Dexter+O**
ECCV 2016



**HO-3D**
CVPR 2020

**ContactPose**
ECCV 2020

**Contact Force**
TPAMI 2018

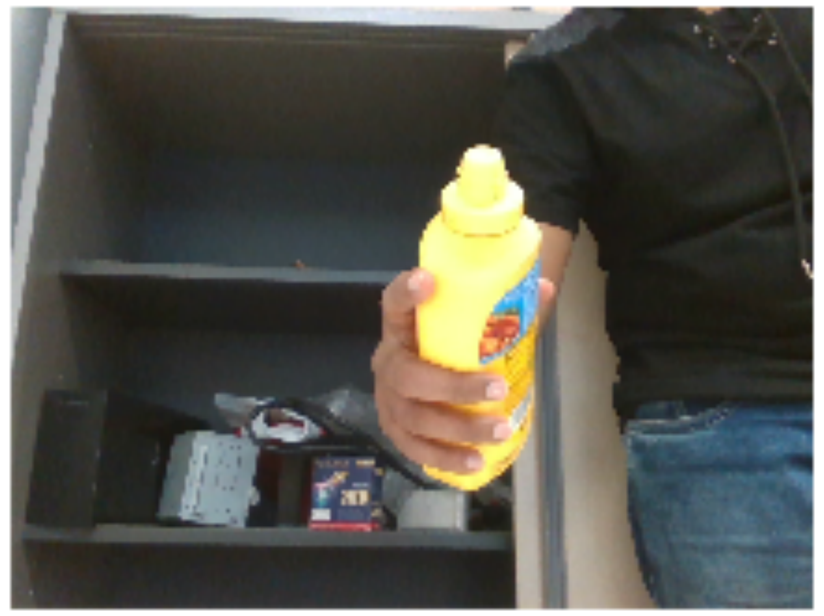
**FPHAB**
CVPR 2018

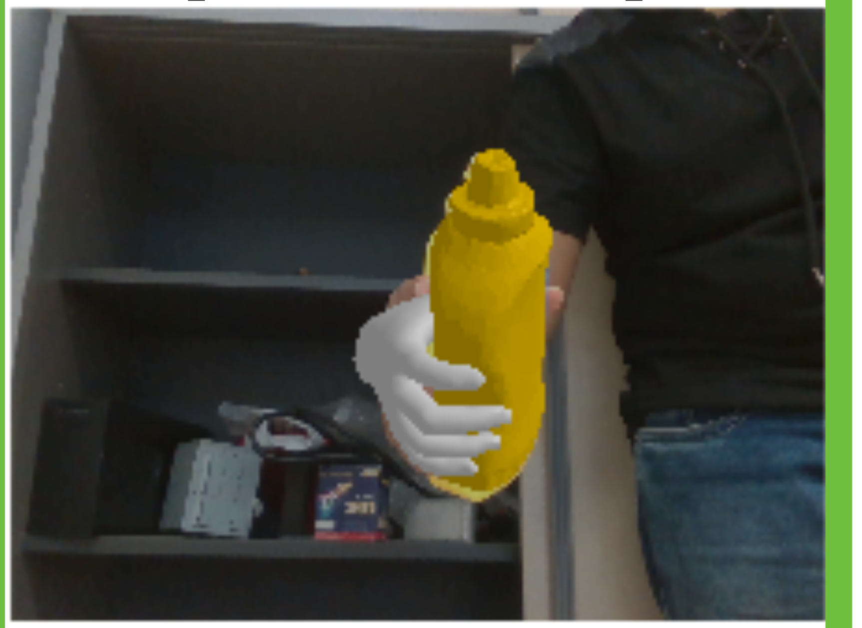
# Learning-based methods fail on different domains

**Learning**
[Hasson 2019]

**Fitting**
[Hasson 2021]

Seen object,
Same domain
[HO-3D]

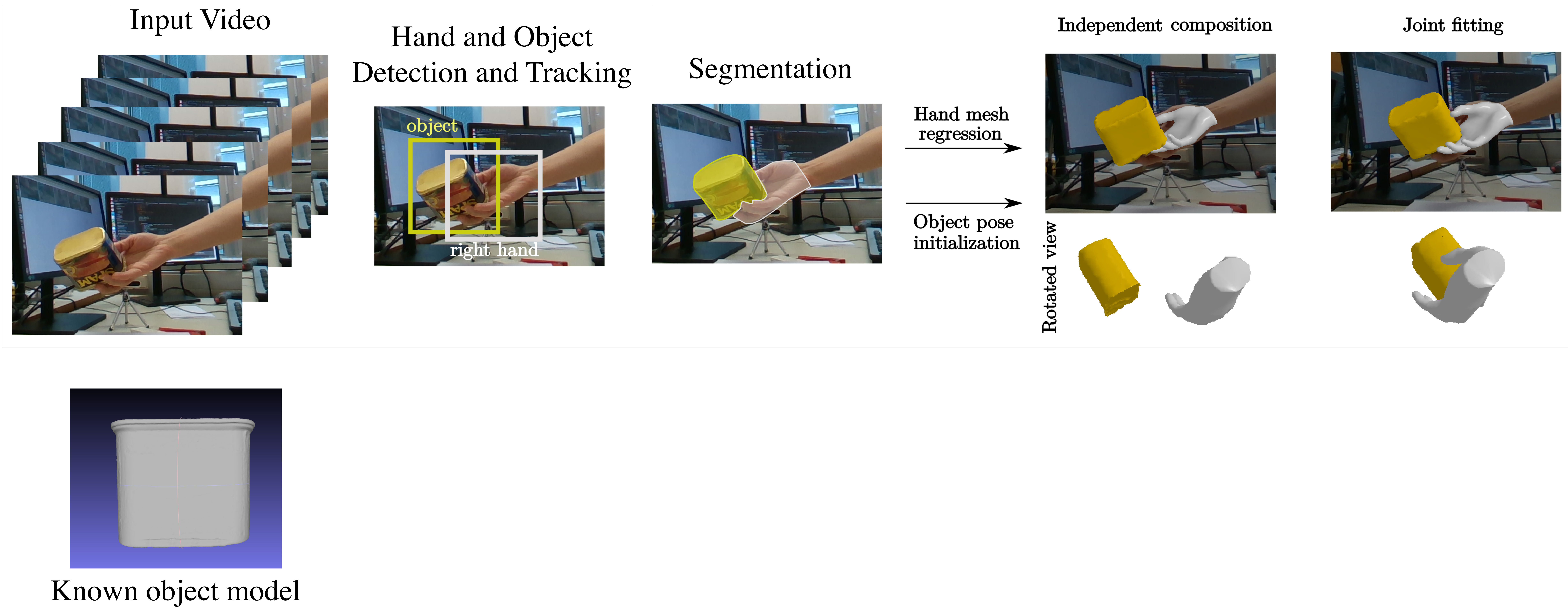




Seen object,
Different domain
[Dex-YCB]

Unseen object,
Different domain
[Core50]

# Towards unconstrained joint hand-object reconstruction from RGB videos



Input clip

Object model

Joint fitting

Camera view

Rotated views

Input clip

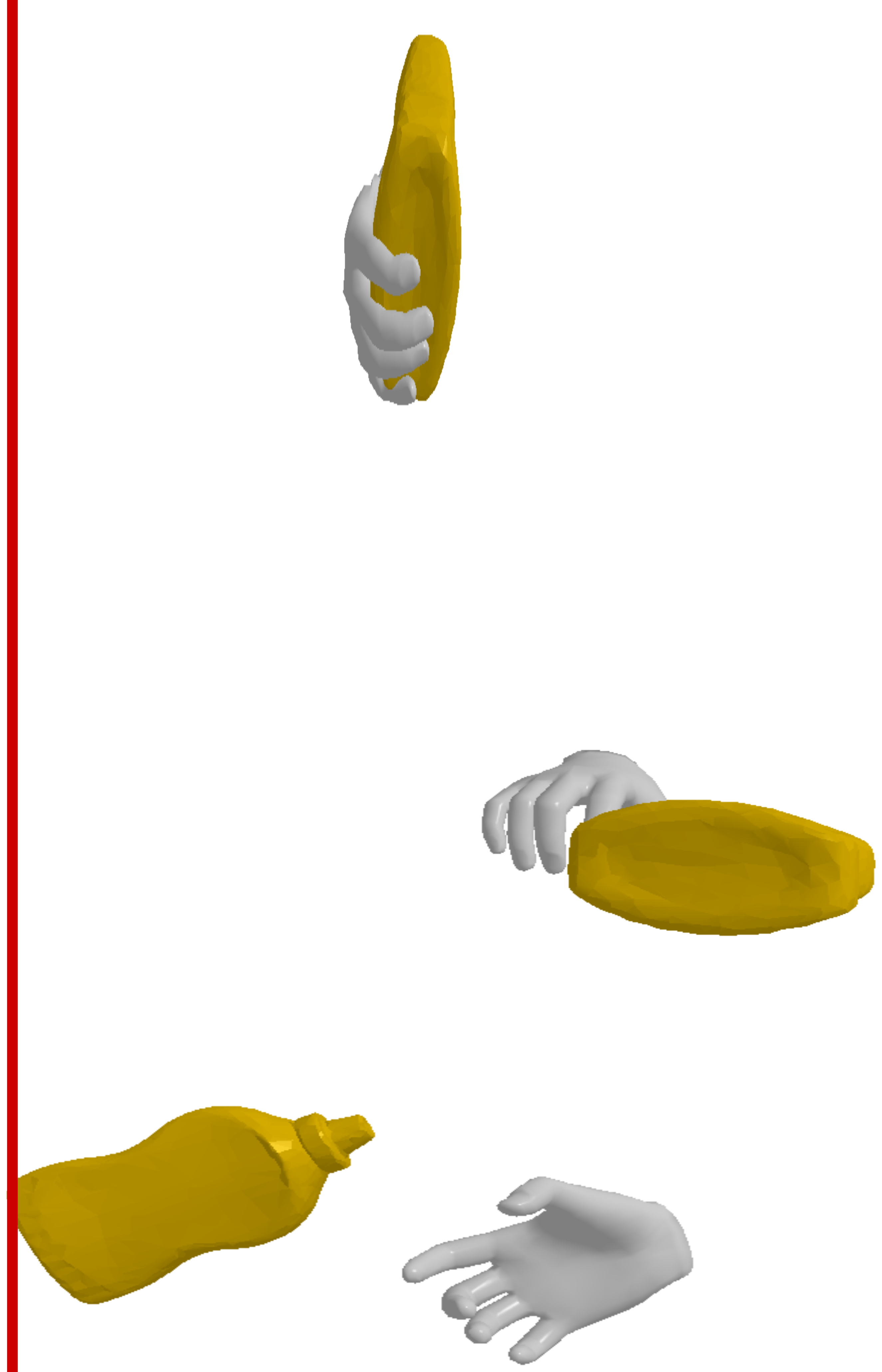
Object model

Joint fitting

Camera view

Rotated views

Independent composition



Input clip

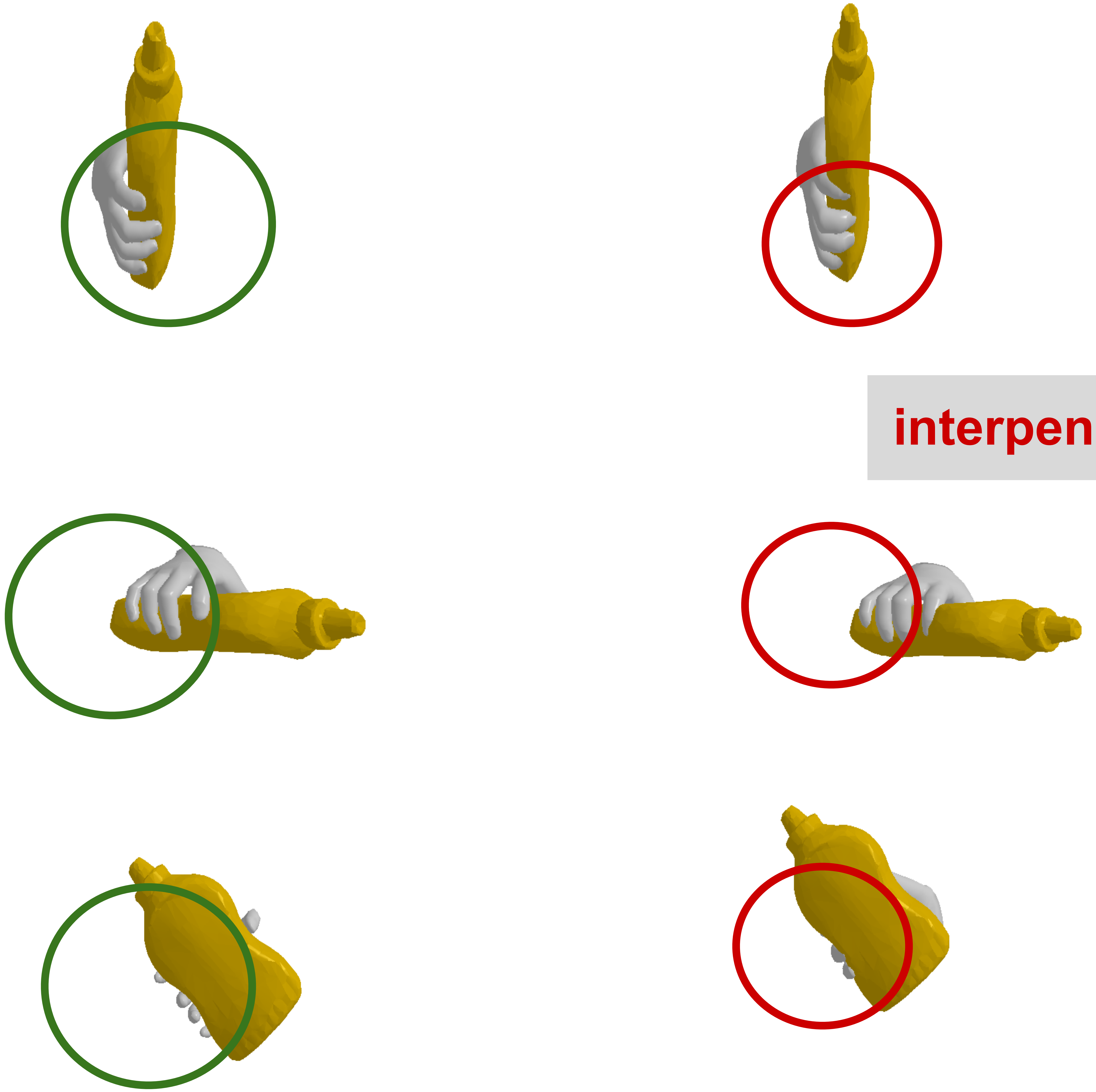
Object model

Joint fitting

Without collision penalization $\mathcal{L}_{col}$

Camera view

Rotated views



**interpenetration**

Input clip

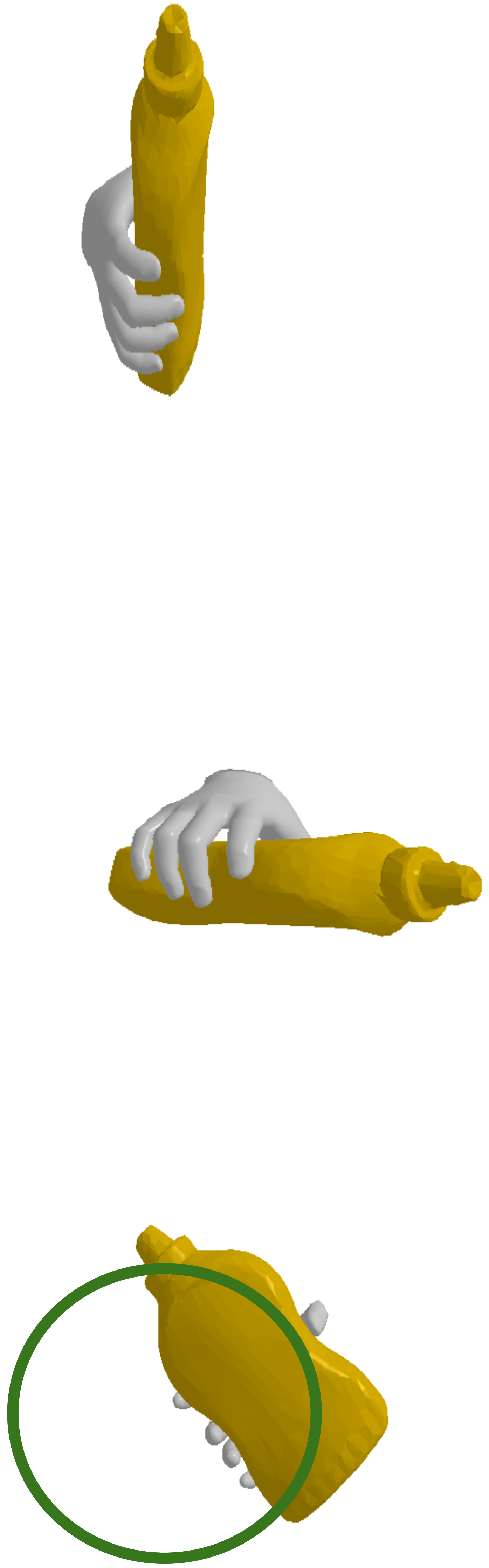
Object model

Joint fitting

Without local interactions $\mathcal{L}_{local}$

Camera view

Rotated views



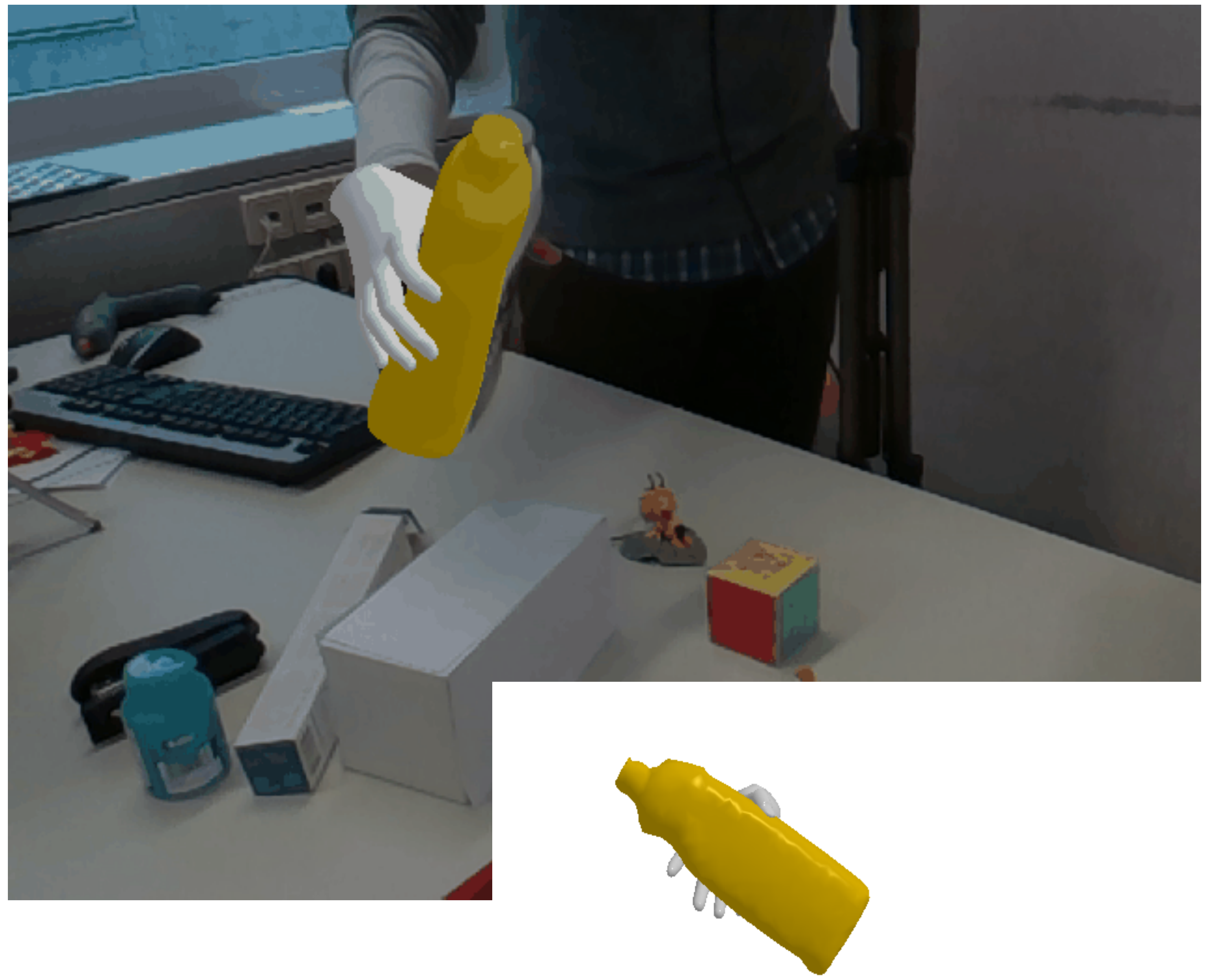
**contact errors**

# Results on HO-3D

Input clip

2D segmentations

Reconstruction

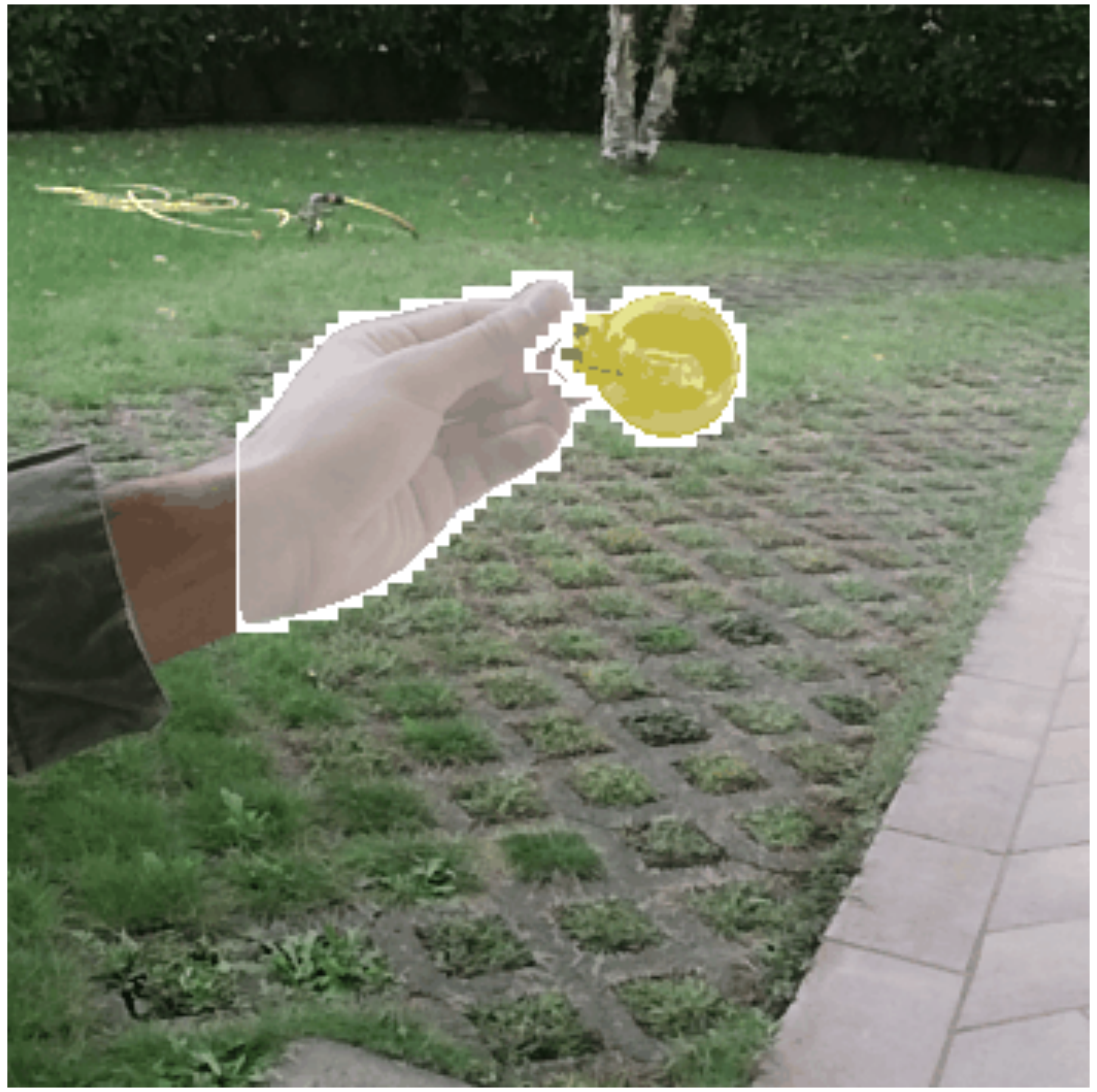


# Ablation study for effect of 2D noise in pseudo-labels for HO-3D dataset

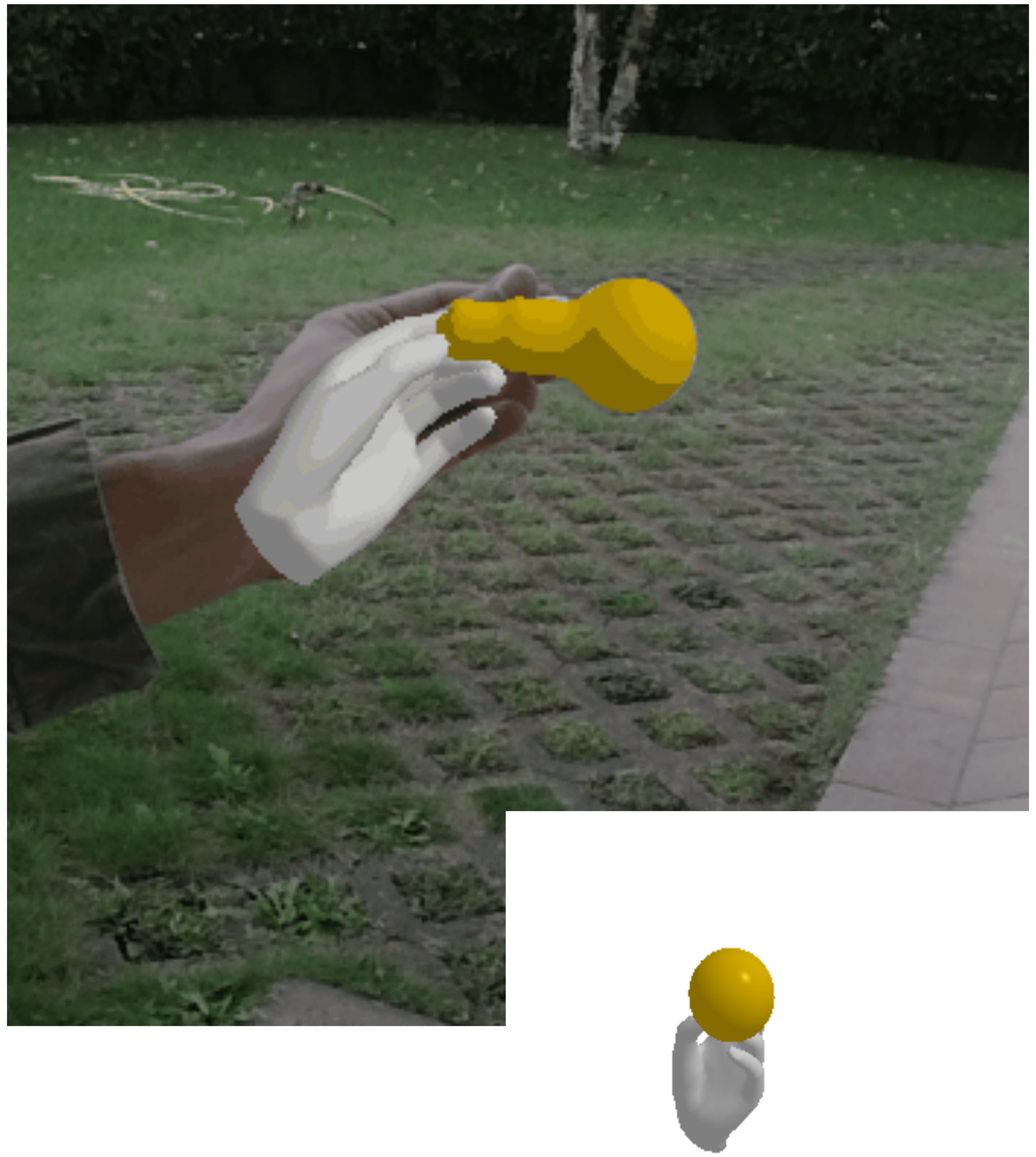
# Results on Core50

Input clip

2D segmentations



Reconstruction



# Results on Epic-Kitchens

Input clip

2D segmentations

Reconstruction

# Summary

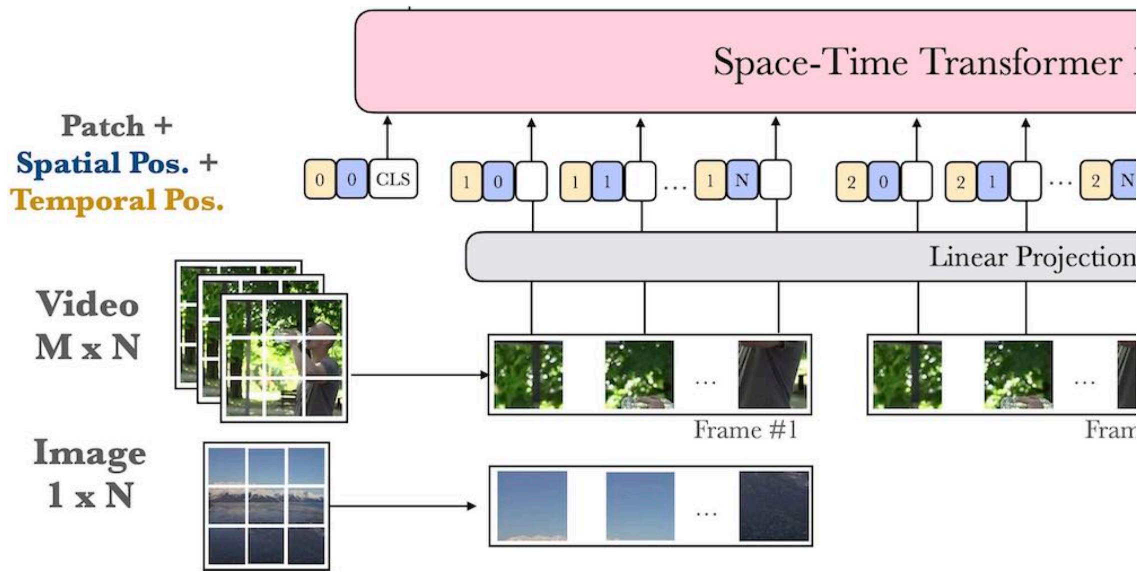
**This talk:**

- ✦ 1) Text-to-video retrieval
- ✦ 2) Temporal localisation in sign language videos
- ✦ 3) Hand-object reconstruction from RGB videos

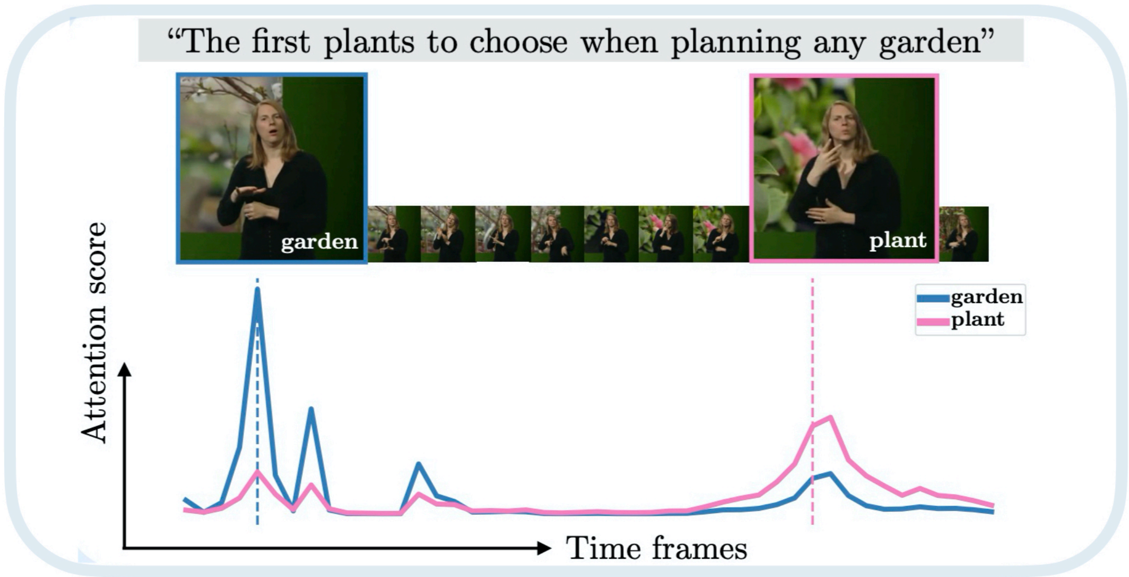
**What next? Open problems?**

- Long-term video modeling
- Using more knowledge from image models for video modeling
- Sign language translation
- Bridging the gap between 3D and semantics

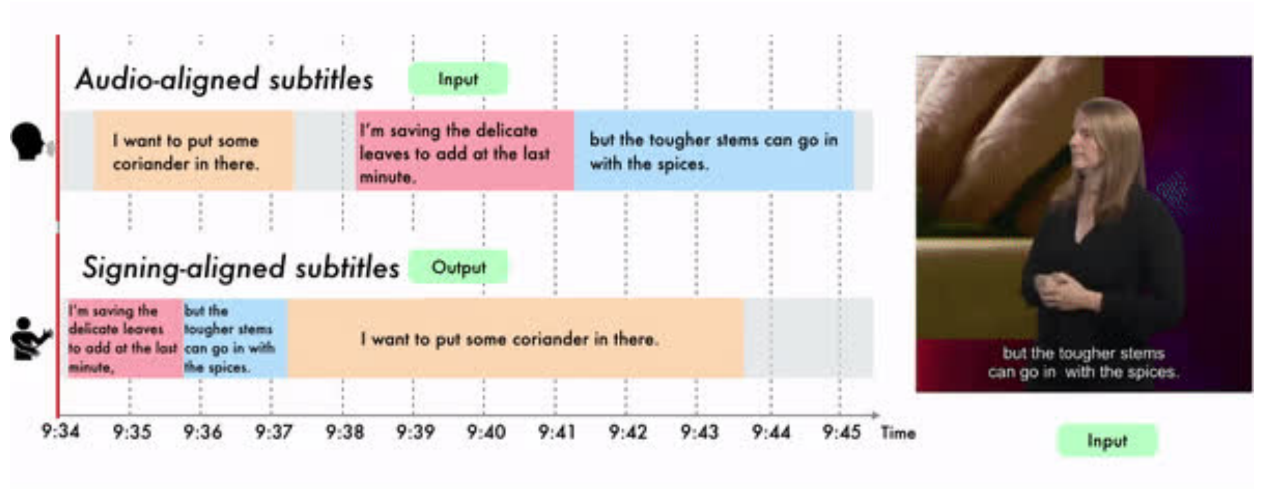
# Summary



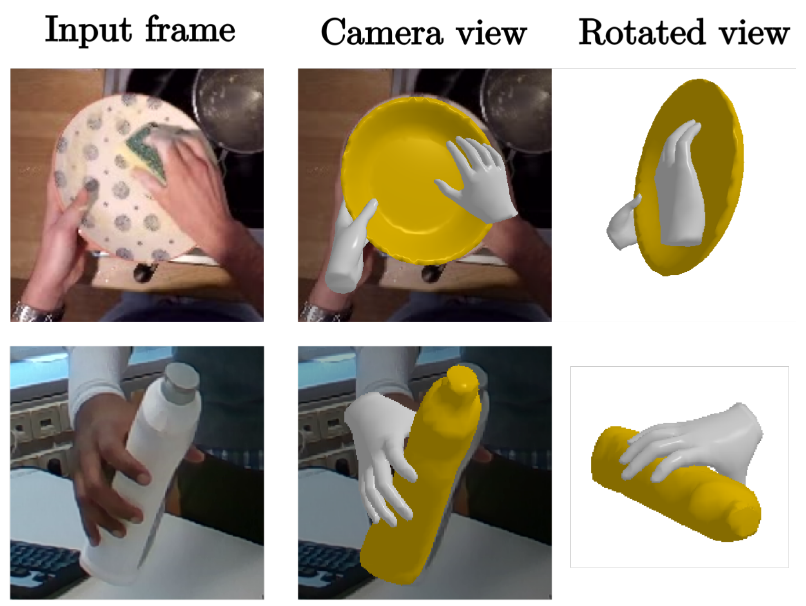

**Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval**
Max Bain, Arsha Nagrani, Gül Varol and Andrew Zisserman
*ICCV* 2021.




**Read and Attend: Temporal Localisation in Sign Language Videos**
Gül Varol*, Liliane Momeni*, Samuel Albanie*, Triantafyllos Afouras* and Andrew Zisserman
*CVPR* 2021.




**Aligning Subtitles in Sign Language Videos**
Hannah Bull*, Triantafyllos Afouras*, Gül Varol, Samuel Albanie, Liliane Momeni and Andrew Zisserman
*ICCV* 2021.




**Towards unconstrained joint hand-object reconstruction from RGB videos**
Yana Hasson, Gül Varol, Cordelia Schmid and Ivan Laptev
*3DV* 2021.