Learning to understand meaningful gestures

Andrew Zisserman

Visual Geometry Group (VGG)

University of Oxford

June 2026

Two Types of Semantic Gestures for Communication

Sign Language

Co-Speech Gestures

so this **cycle** of perception and action

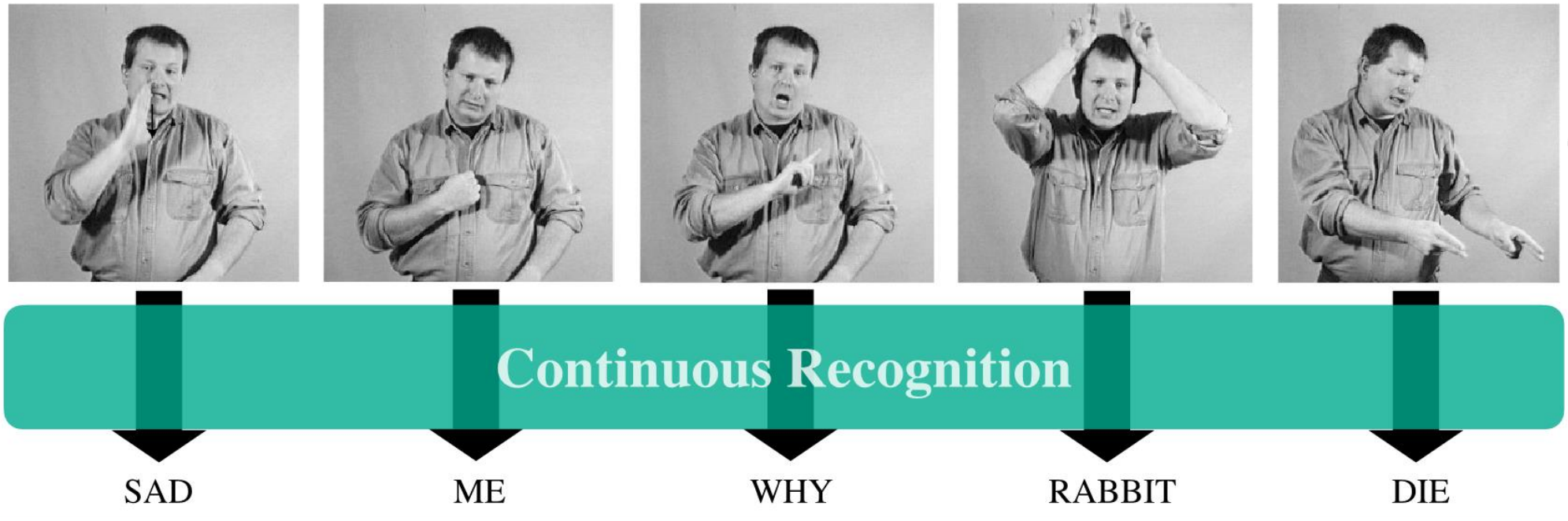
Part I: Sign language Translation

Objective: Sign Language Translation

- **Translate BSL to English**

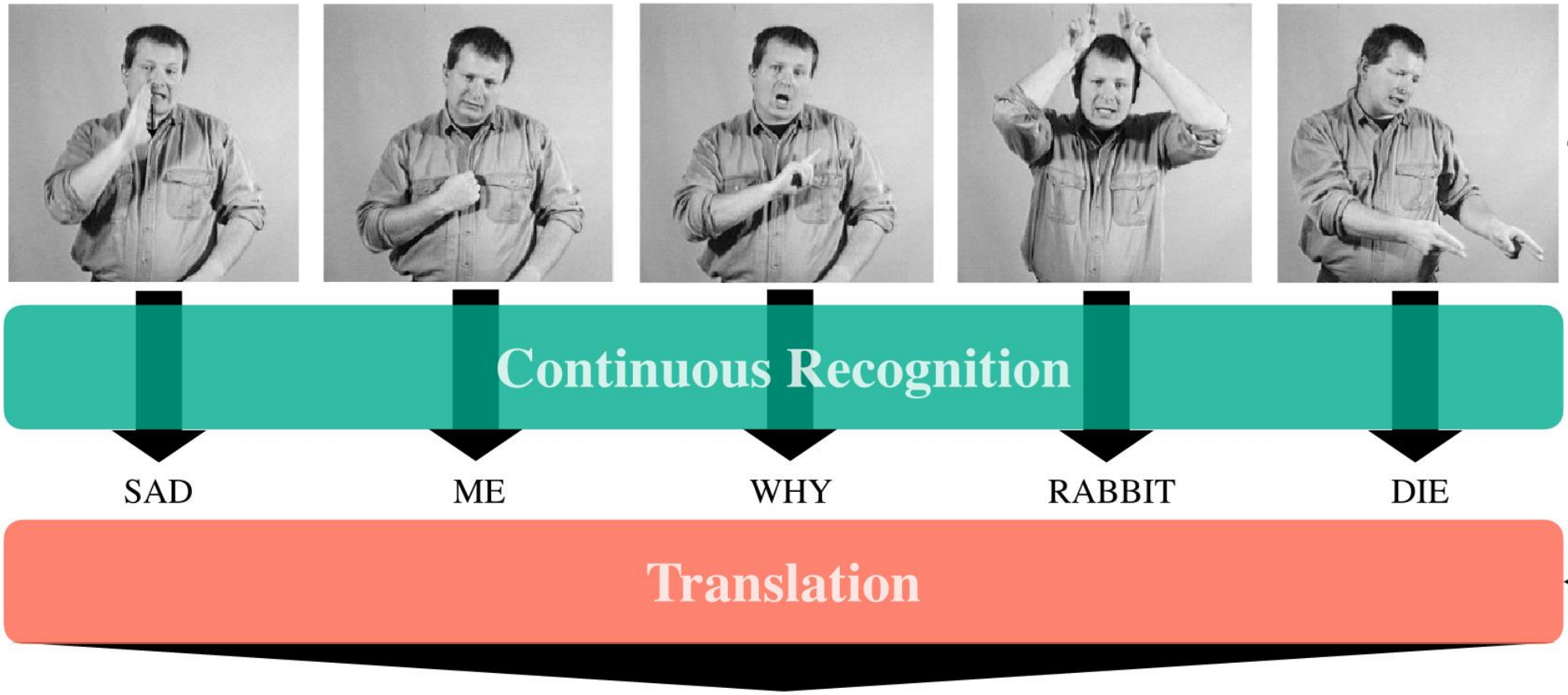
Two `problems’:

1. Transform the video to a gloss sequence (a sequence of sign-words)
2. Translate the gloss sequence to English



Two `problems’:

1. Transform the video to a gloss sequence (a sequence of sign-words)
2. Translate the gloss sequence to English

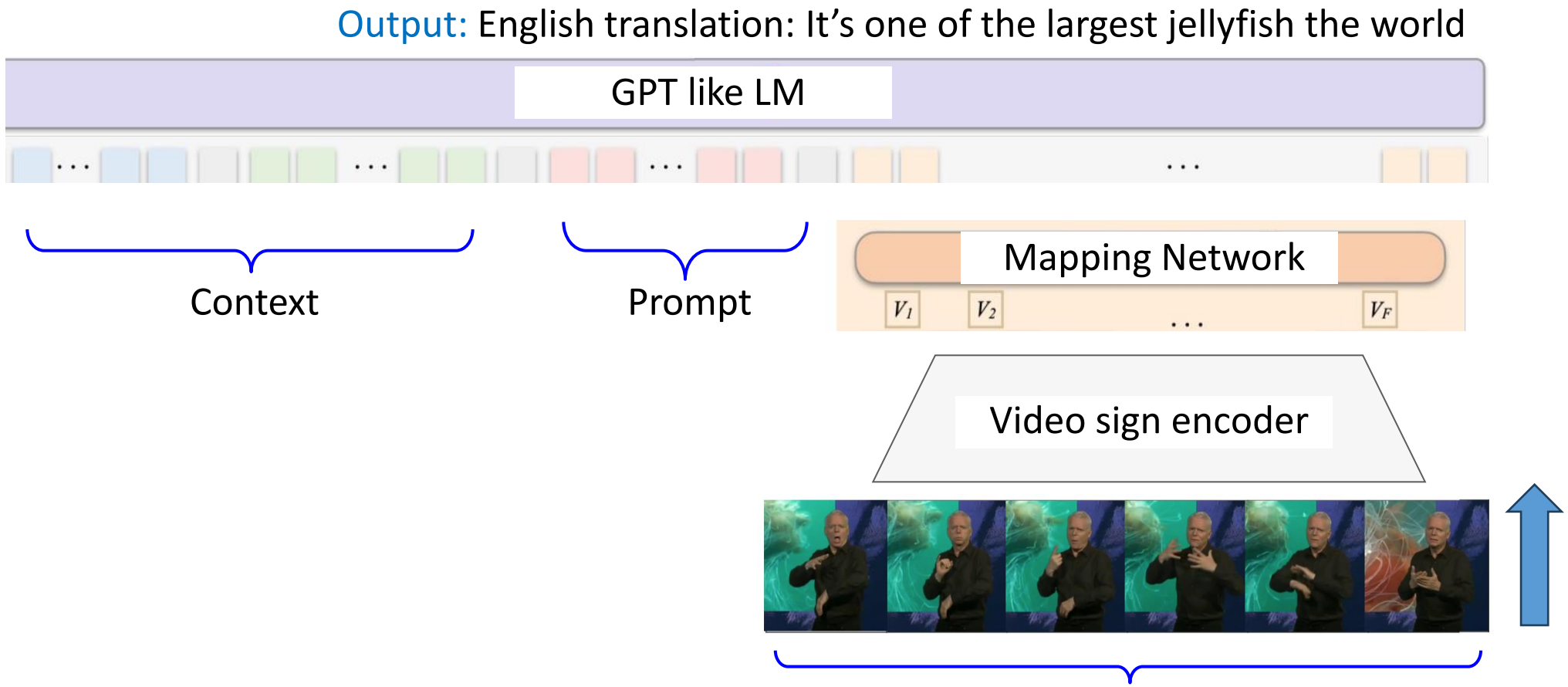


English language output: *I am sad because the rabbit died.*

Two `problems':

1. Transform the video to a gloss sequence (a sequence of sign-words)
2. Translate the gloss sequence to English

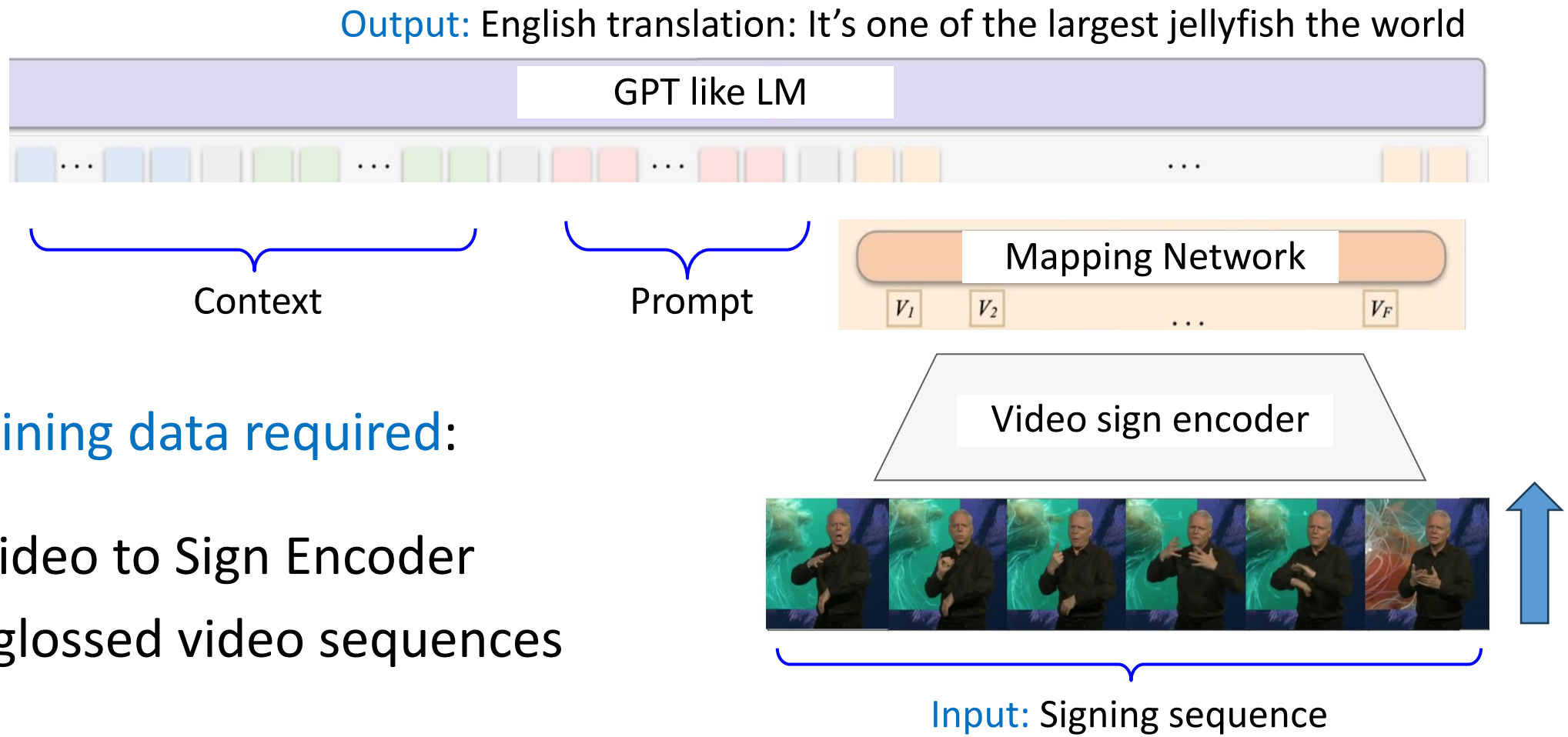
Sign Language Translation – Standard Model



Two `problems':

1. Transform the video to a gloss sequence – Video to Sign Encoder
2. Translate the gloss sequence to English – LLM (pre-trained, e.g. Llama3-8B)

Sign Language Translation – Standard Model



Large-scale training data required:

1. Train the Video to Sign Encoder
   - Requires glossed video sequences

2. Train the mapping network and fine-tune the LLM
   - Requires paired sign-sequence English-sentences

Overview

1. How to obtain large scale training data?
2. Improving performance with context
3. Improving performance with a better video sign encoder
4. Beyond signer interpreted video

Obtaining large scale training data

Approach: learn signs from signer-interpreted TV

Signer-interpreted TV dataset

1. Obtain sign annotations

1a. Spotting individual signs using mouthings

BSL-1K: Scaling up sign language recognition using mouthing cues, T. Afouras, S. Albanie, J. S. Chung, N. Fox, L. Momeni, Prajwal K. R., G. Varol, A. Zisserman, ECCV 2020

Localizing mouthings using Visual Keyword Spotting

- Leverage subtitles to perform visual key word spotting

- Result: localized sign

Search within subtitled window for “happy”

Mouthing detected:

IMPORTANT

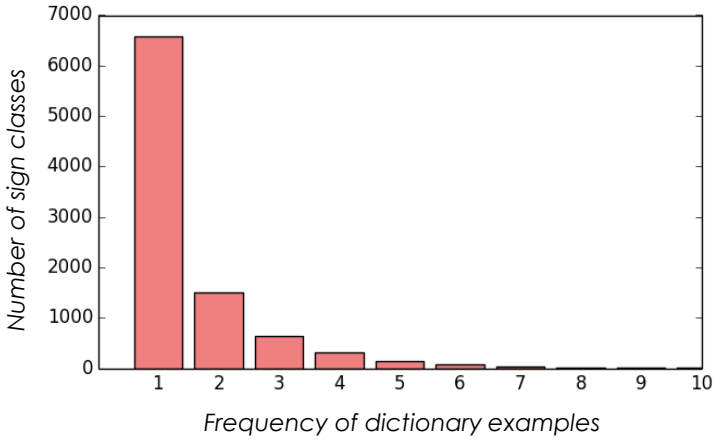
1. Obtain sign annotations

1b. Spotting individual signs using a sign dictionary

Watch, read and lookup: learning to spot signs from multiple supervisors, Momeni, Varol, Albanie, Afouras, Zisserman, ACCV 2020

BSLDict: Dictionary dataset

- Collected from signbsl.com
- 14K video clips
- 9K signs
- 1-10 example clips per sign
- 148 signers

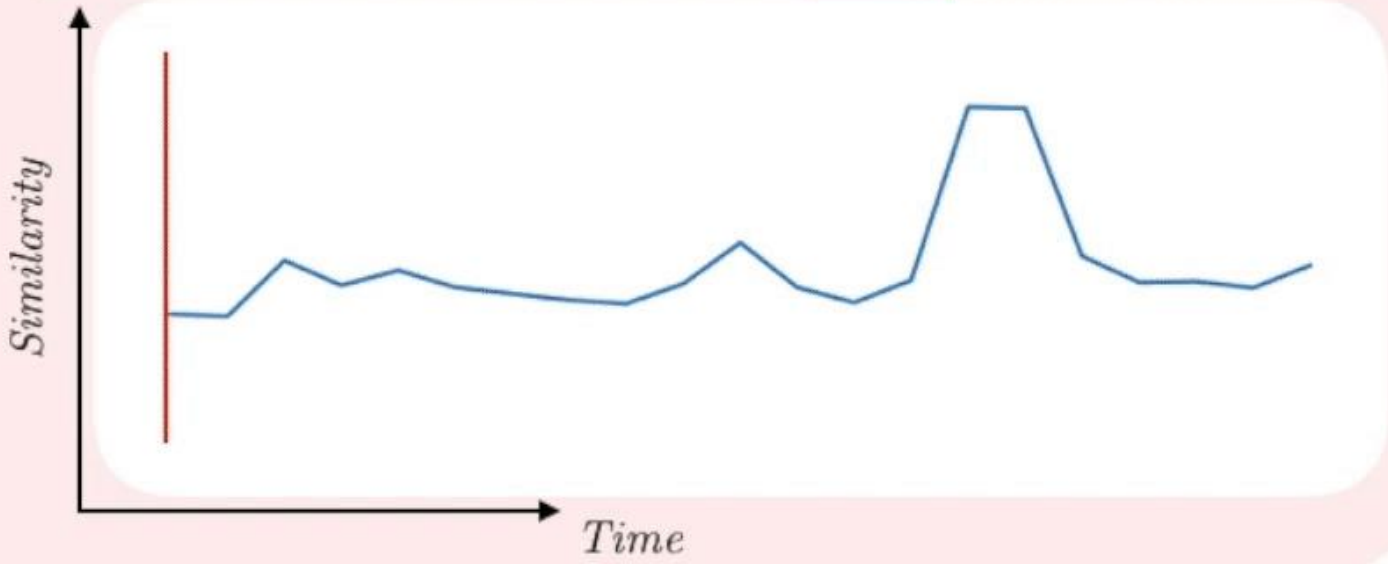


Objective: temporal similarity map

Isolated sign "Apple" from dictionary

Continuous signing

"We have some reasonable-sized apple trees in the garden."



New Annotations: isolated sign annotations (pseudo-glosses)

Generated automatically

Example: PERFECT

2M isolated sign annotations, 8K vocabulary, available on BOBSL website

Obtained from different sources such as mouthings, dictionary matching, …

Training Data

BOBSL

- BSL only
- 1200 hrs
- 750k training sentences
- 37 signers

YouTube-SL-25

YouTube-SL-25 is a corpus of video IDs corresponding to >3200 hours of sign language videos with seemingly well-aligned captions, across >25 sign languages. The dataset was curated using coarse human filtering, so a small fraction of content may be irrelevant, incorrect, or misaligned. **You can find the list of video IDs here.**

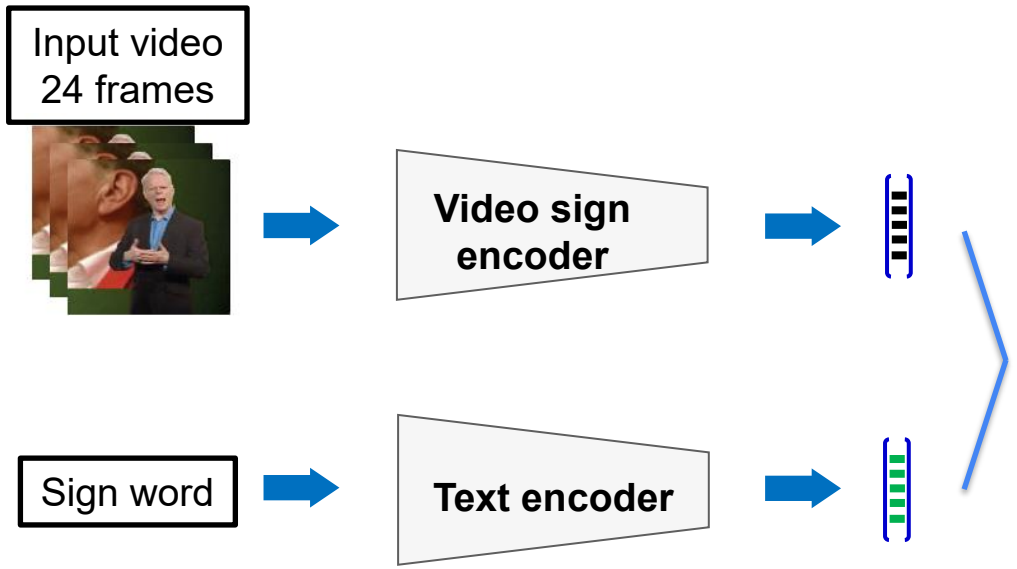
- ASL & BSL subsets
- 1500 hrs
- 450k training sentences
- **2500 signers**

YouTube-SL-25: A largescale, open-domain multilingual sign language parallel corpus.
Garrett Tanzer and Biao Zhang, ICLR 2025

Training

1. Pre-train video sign encoder

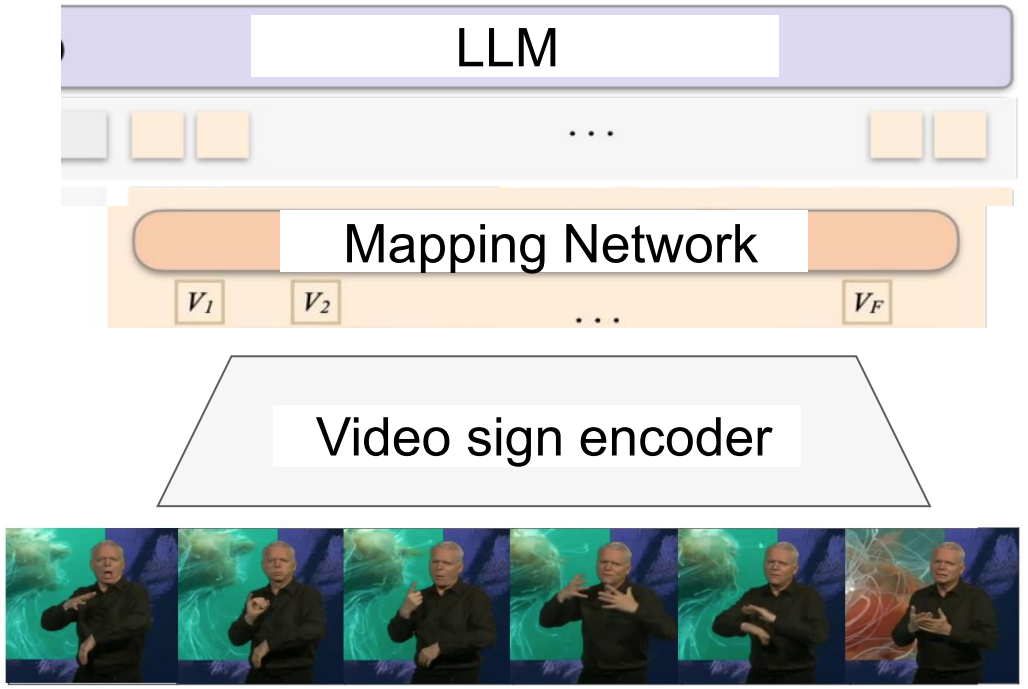
- Word (gloss) level contrastive training
- Train on pseudo-glosses from BOBSL



- Swin Video encoder

2. Train entire network

- Sentence translation prediction
- Train on BSL from BOBSL and BSL & ASL from YouTube-SL-25



- DoRA fine-tune LLM
- End-to-end fine-tuning

Sign Language Translation: Results

Sign Language Translation: Results

Ground Truth (GT): It’s one of the largest jellyfish in the world

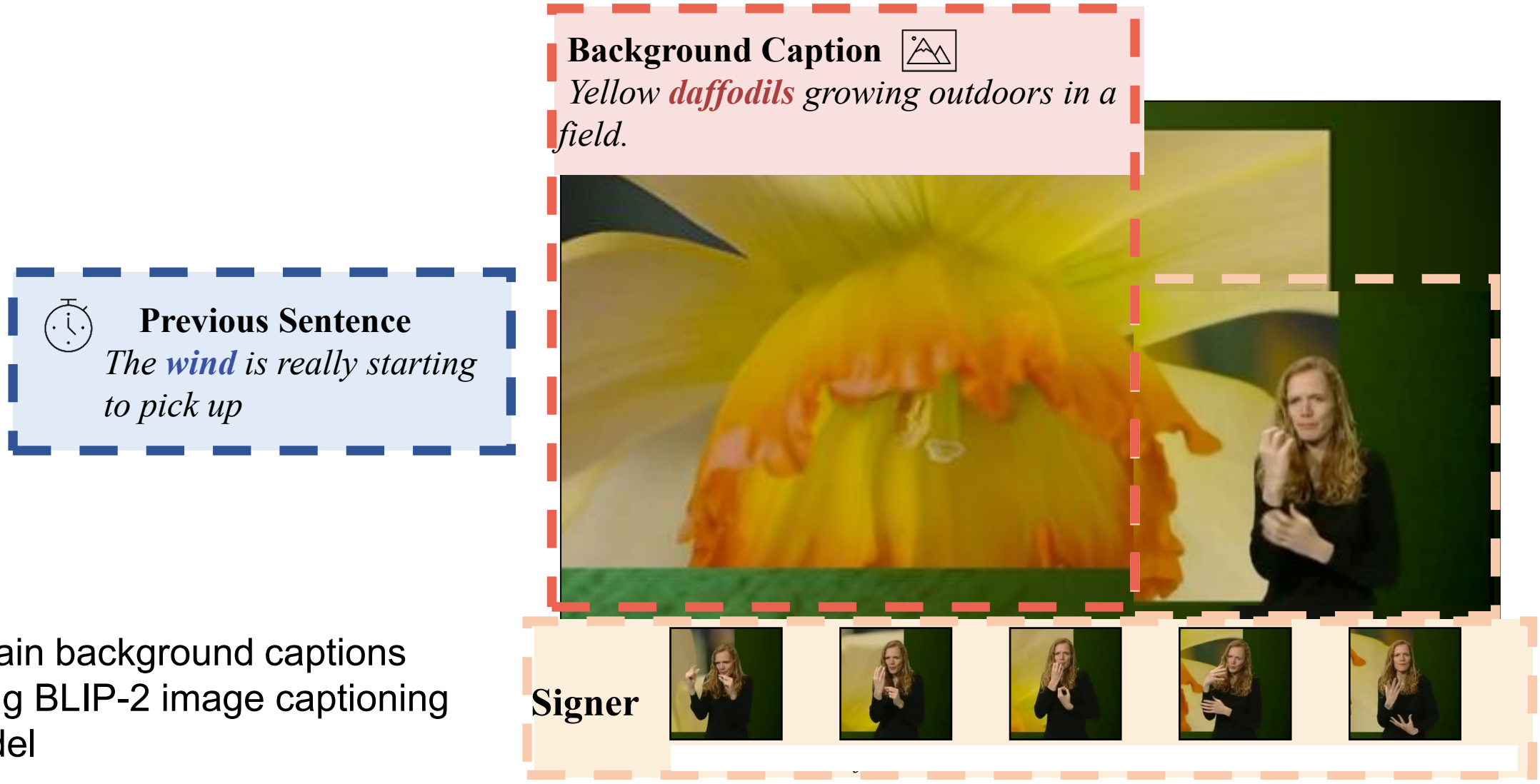
Predicted: It’s one of the biggest animals in the world

Improving performance with context

Lost in Translation, Found in Context: Sign Language Translation with Contextual Cues

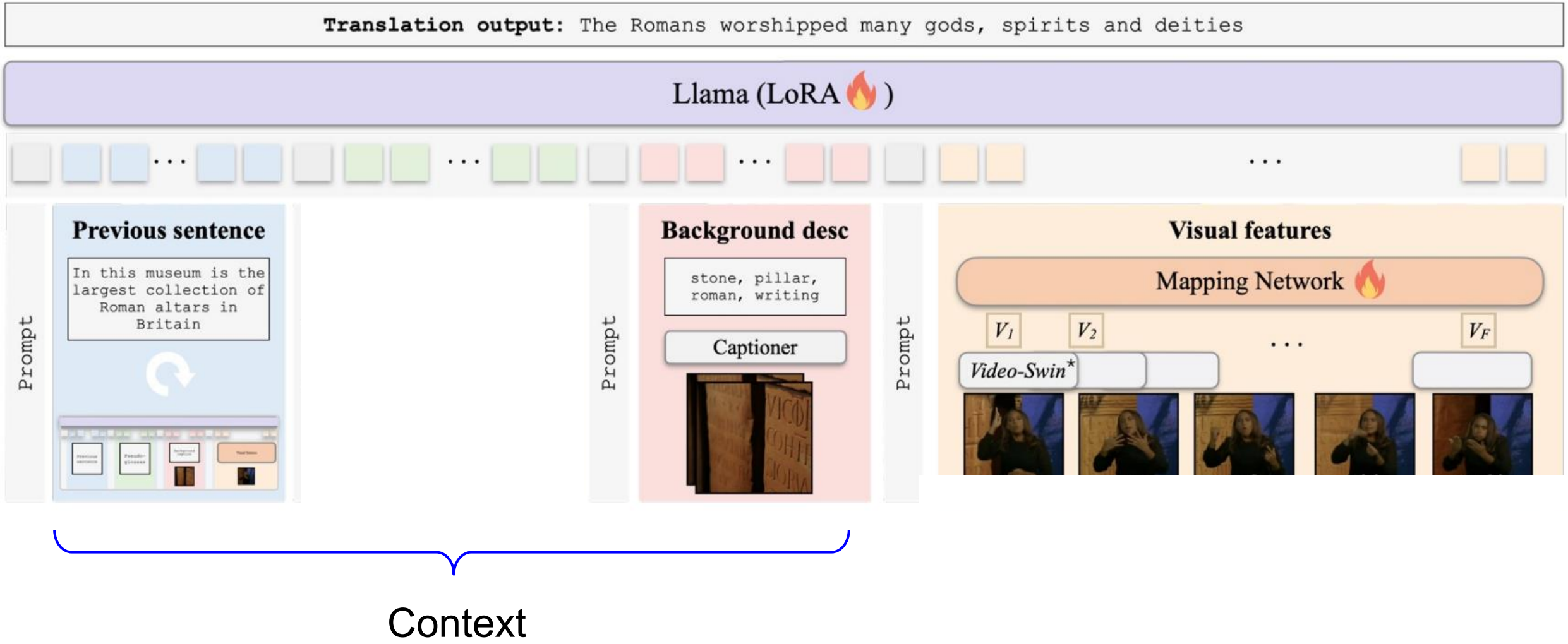
Youngjoon Jang, Haran Raajesh, Liliane Momeni, Gül Varol, Andrew Zisserman, CVPR 2025

Translation for Signer-Interpreted Data



- Obtain background captions using BLIP-2 image captioning model
- Annotate 1 frame per second

Sign Language Translation with Contextual Cues



Sign Language Translation: Results

BG Image

Sign

Prev : It looks like a lion's mane, it's about two and half half meters wide and 50 meters long, and it's got 50 meters of tentacles hanging underneath it.

BG : man, holding, net, water, jellyfish, floating, swimming

GT : It's one of the largest jellyfish in the world.

Vid only : It's one of the biggest animals in the world

Predicted (with Prev & BG) : It's one of the largest jellyfish in the world

Lost in Translation, Found in Context: Sign Language Translation with Contextual Cues,
Youngjoon Jang, Haran Raajesh, Liliane Momeni, Gül Varol, Andrew Zisserman, CVPR 2025

Improving performance with a better video sign encoder

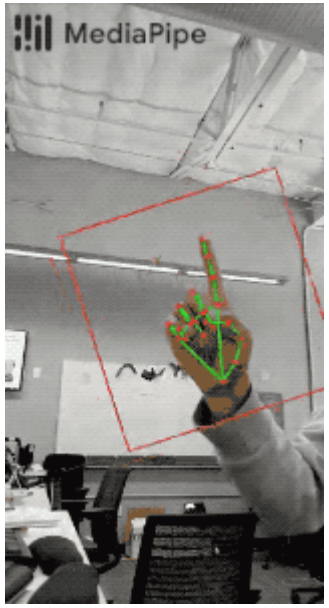
Lost in Translation, Found in Embeddings: Sign Language Translation and Alignment

Youngjoon Jang, Liliane Momeni, Zifan Jiang, Joon Son Chung, Gül Varol, Andrew Zisserman, arXiv 2026

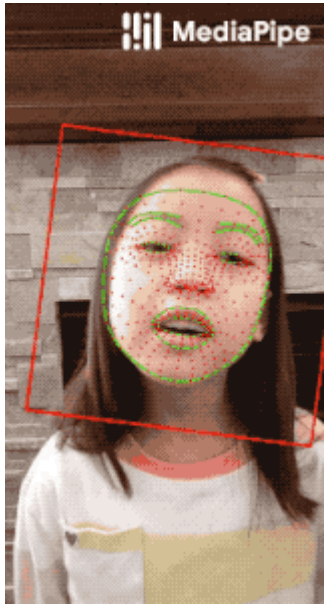
Build on pre-trained human pose and lip features

Google MediaPipe

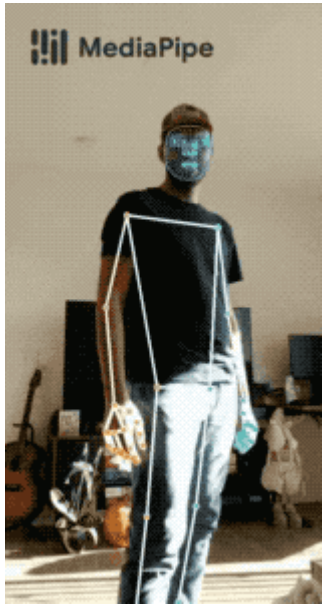
Input: video
Output: frame-wise human pose keypoints



hand



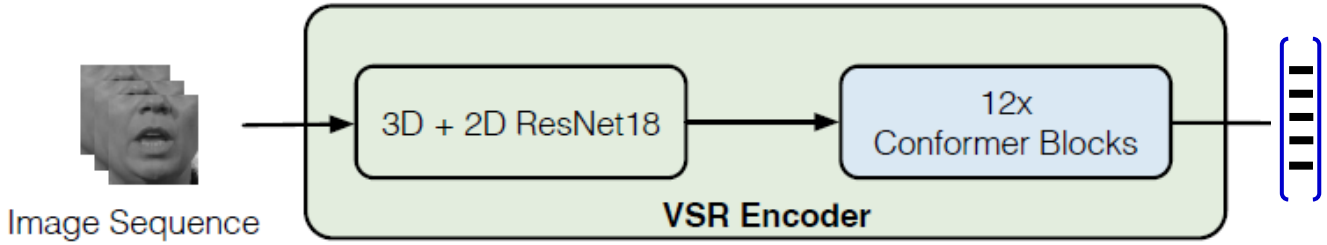
face



body

AUTO-AVSR Lip Reading Features

Input: video
Output: frame-wise lip features



mouth morphemes/mouthings

AUTO-AVSR: Audio-visual speech recognition with automatic labels.
Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic, ICASSP, 2023

Sign Frames

Ours: The Yorkshire Dales are home to 30 times as many sheep as people - half a million of them.

GT: In fact, there are 30 times more sheep in the Yorkshire Dales than people - more than half a million of them.

Quantitative Results

On BOBSL sentence translation test set

Encoder	Context used	BLEURT score (max 100)
Swin	No	37.8
Swin	Yes	40.3
Pose/mouth	No	47.1

Beyond signer-interpreted data

BSL Corpus

- BSL only
- 125 hours of spontaneous linguistically curated data
- 249 deaf BSL users
- Conversations/interviews
- https://bslcorpusproject.org/

Generalization to BSL Corpus?

Signer Interpreted Data

BSL Corpus Conversational

Translation of unseen BSL Corpus conversation clips

Sign Frames

Ours: If I am signing to someone and they don't understand me, I have to change my signs.

GT: If I am signing to someone, and they don't understand me, I have to make changes to my signing.

Translation of unseen BSL Corpus conversation clips

Sign Frames

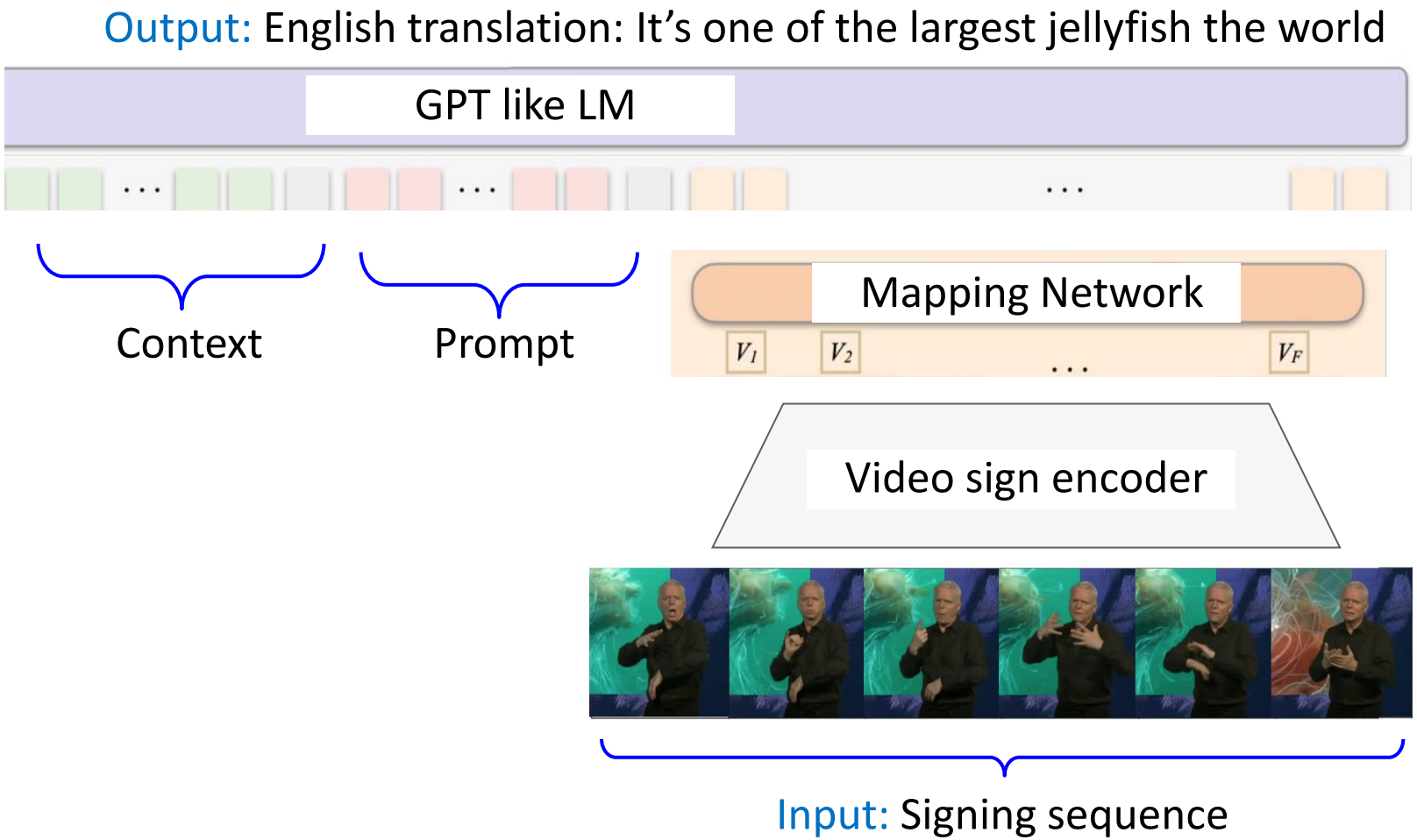
Ours: In Scotland, their signs are completely different.

GT: In Scotland, their signs are completely different.

Summary & take home messages

Sign Language Translation

- Good progress
- Use of LLMs with large-scale paired and glossed data
- Next:
  - Scale-up
  - Multiple-languages
  - Real-time translation



Part II:Recognizing Co-Speech Gestures

What are Co-speech Gestures?

Natural hand and body movements; help to emphasise and illustrate the spoken content

so this **cycle** of perception and action

you cannot **open** the newspapers without …

What are Co-speech Gestures?

Natural hand and body movements; help to emphasise and illustrate the spoken content

(time) the past …

Categories of co-speech gestures

1. Iconic, e.g. miming a rising quantity
2. Metaphoric, more abstract like time
3. Beat, matching rhythm of speech
4. Deictic, e.g. pointing to a particular person, or object of reference

“Hand and mind: What gestures reveal about thought”, David McNeill, 1994

Challenges in Recognizing Co-speech Gestures 1

- Most of the spoken words are not gestured
- Presence of random and beat gestures
- Unavailability of labelled datasets

Beat gestures

Random gesture

Challenges in Recognizing Co-speech Gestures 2

- No one-to-one mapping between gestures and words

Same word but different gestures

ZOOM

ZOOM

ZOOM

Similar gestures but different words

GROW

EXPAND

HEAVY

Co-Speech Gesture Research

Most previous and current work is on **generating** co-speech gestures, e.g.

- Learning individual styles of conversational gesture.

Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., Malik, J., CVPR 2019

- ExpressGesture: Expressive gesture generation from speech through database matching.

Ferstl, Y., Neff, M., McDonnell, R., Computer Animation and Virtual Worlds, 2021

- GestureDiffuClip: Gesture diffusion model with clip latents.

Ao, T. , Zhang, Z. , Liu, L., ACM Transactions on Graphics (TOG), 2023

- Joint Co-Speech Gesture and Expressive Talking Face Generation using Diffusion with Adapters,

Hogue, S., Zhang, C., Tian, Y., Guo, X., WACV 2025

Co-Speech Gesture Research

Very little on **understanding** co-speech gestures (perception), e.g.

- Co-Speech Gesture Detection through Multi-Phase Sequence Labeling

Esam Ghaleb, Ilya Burenko, Marlou Rasenberg, Wim Pouw, Peter Uhrig, Judith Holler, Ivan Toni, Asli Özyürek, Raquel Fernández, WACV 2024

- SocialGesture: Delving into Multi-person Gesture Understanding,

Xu Cao, Pranav Virupaksha, Wenqi Jia, Bolin Lai, Fiona Ryan, Sangmin Lee, James M. Rehg, CVPR 2025

- Understanding Co-speech Gestures in-the-wild,

Sindhu B Hegde, K R Prajwal, Taein Kwon, Andrew Zisserman, ICCV 2025

Progress requires a large-scale dataset for training and evaluation

Overview

1. Introduce a new large-scale dataset for co-speech gestures

2. Classify semantic co-speech gestures (iconic, metaphoric, deictic) from others (beat, random, no gestures)

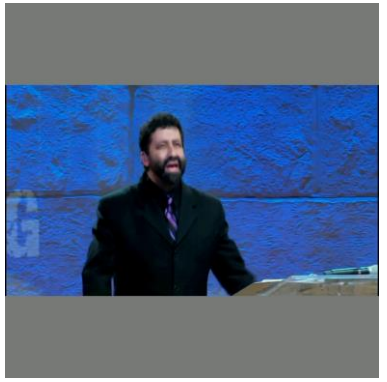
3. Recognize and localize the co-speech gesture

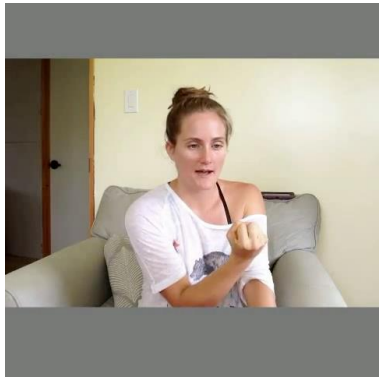
The Gesture Recognition in the Wild (GRW) Dataset

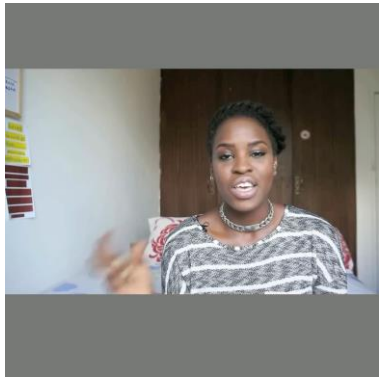
Type	# Words	# Video clips	Duration
Non-semantic	150	139,215	154.68 hours
Semantic	150	17,473	19.41 hours

Manual annotation of the AVSpeech* dataset

- *Word and gesture localization labels*
- *Diverse, in-the-wild speaker environments*
- *Train/test splits – 4000 clips*







Recognizing Co-speech Gestures in-the-wild,
Sindhu B Hegde, K R Prajwal, Andrew Zisserman, arXiv 2026

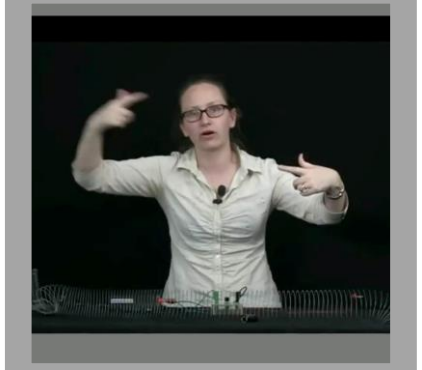
* Ephrat et al, AVSpeech, SIGGRAPH 2018

GRW: Data Samples

CONNECT



SPIRAL

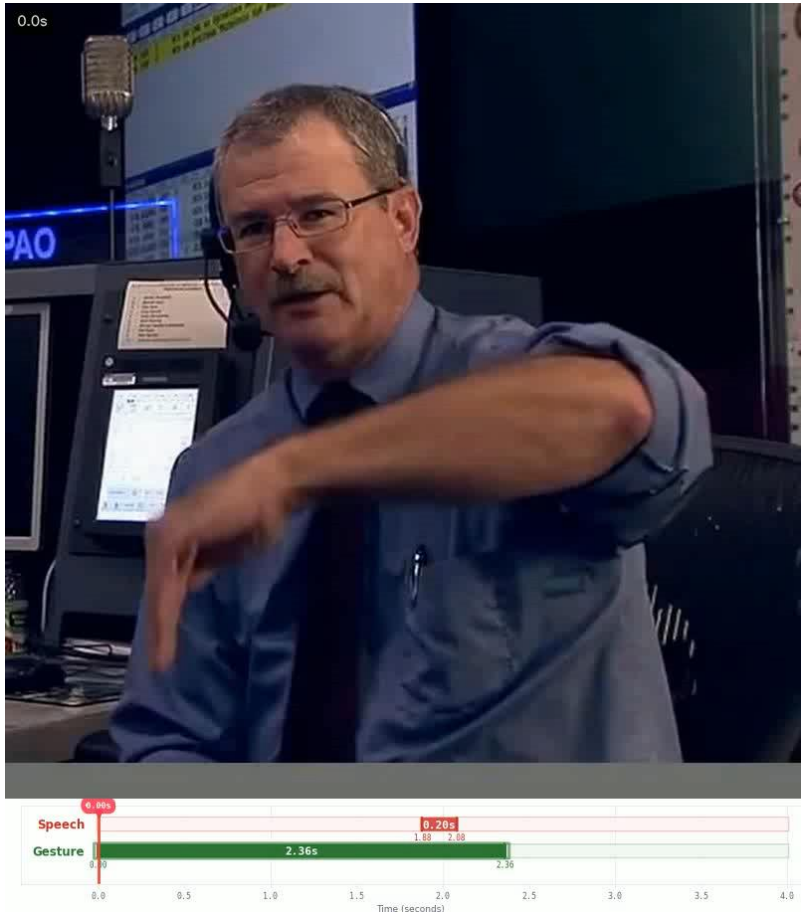



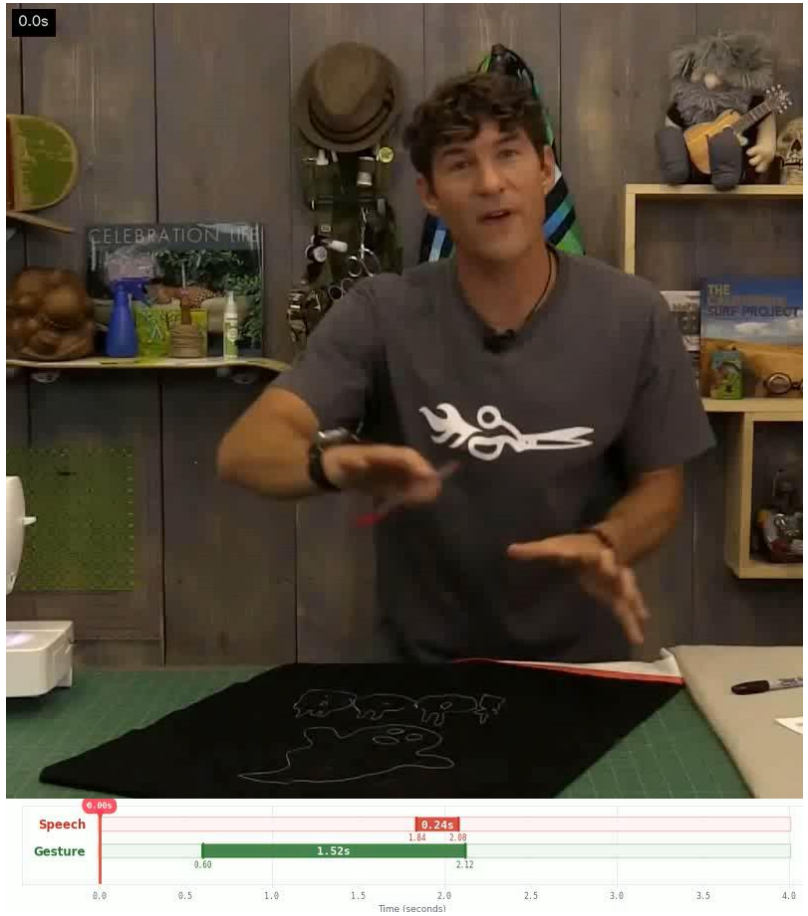

GRASP

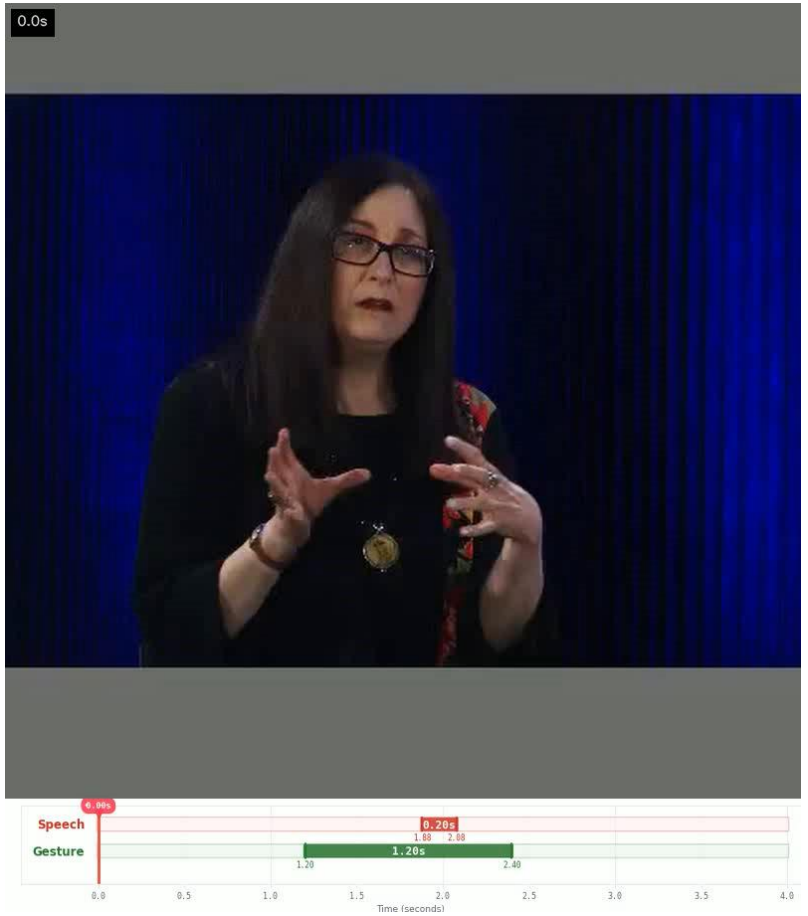



Temporal offset between gesture and spoken word

Target word: **Arc**





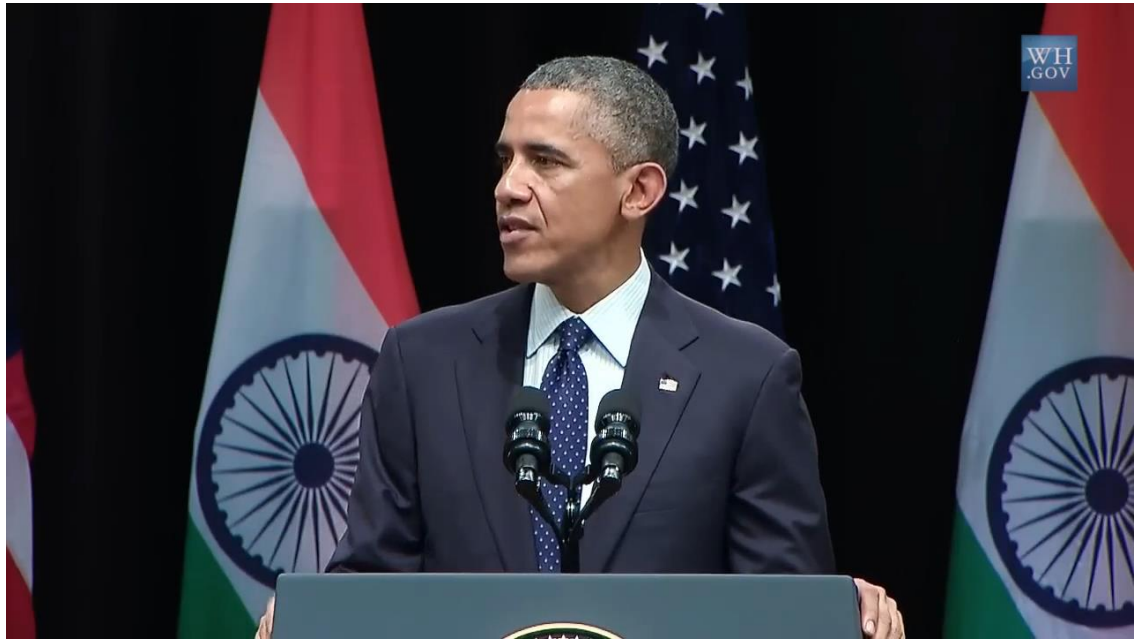


Gesture starts well before the word is spoken

Task 1: Semantic vs. Non-semantic gestures

Semantic gestures

Iconic/ Metaphoric / Deictic



Text: *"Common purpose, if India, as **massive** as it is"*

✔ Clear "iconic" gesture present

✔ Possible to recognise words

Non-semantic gestures

Beat / Random

Text: *"like up close, hear it well, see it well"*

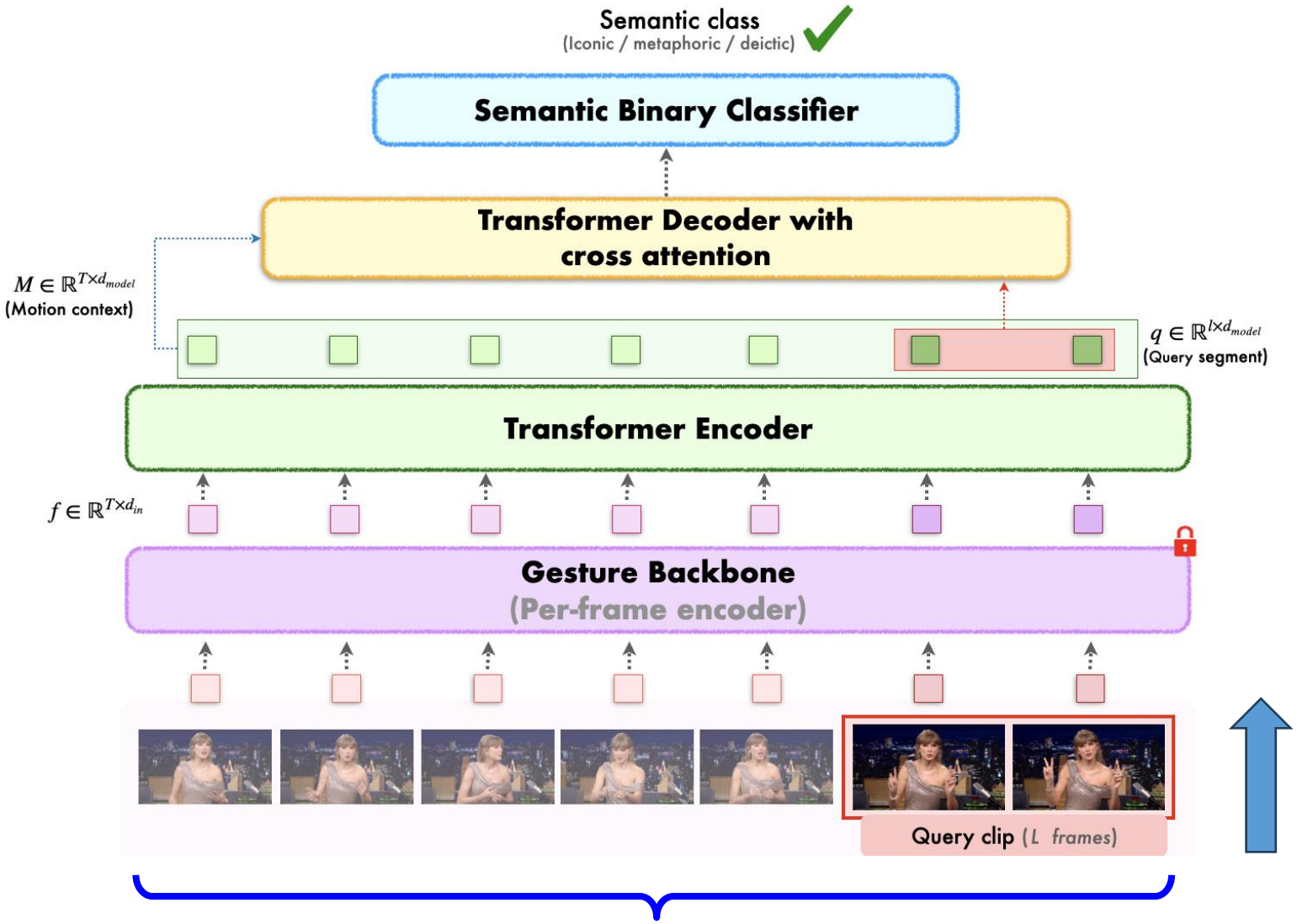
✘ No "semantic" gestures present

✘ Cannot recognize the word from the gesture

✔ Random hand movements / beat gestures

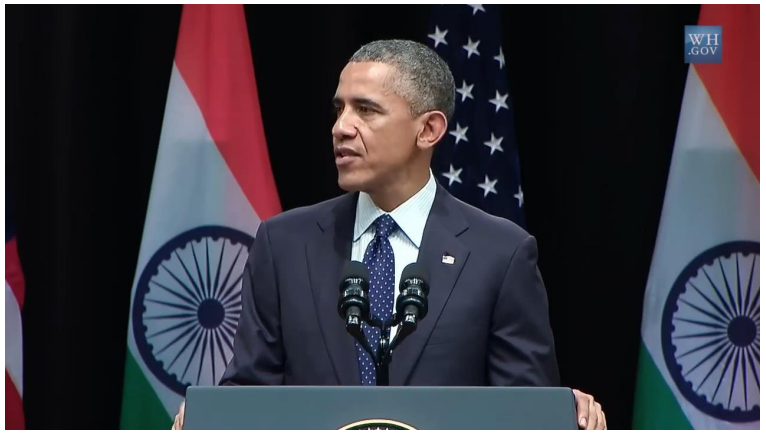
Semantic Gesture Classification: Model

Input: Video sequence



long video for context

Semantic Classification: Results



Pred: Semantic
GT: Semantic

Pred: Semantic
GT: Semantic

Pred: Semantic
GT: Semantic

Pred: Non-semantic
GT: Non-semantic

Pred: Non-semantic
GT: Non-semantic

Pred: Non-semantic
GT: Non-semantic

Note: These videos are of unseen people (in-the-wild testing), not part of our dataset

Quantitative Results

On GRW semantic gesture classification test set

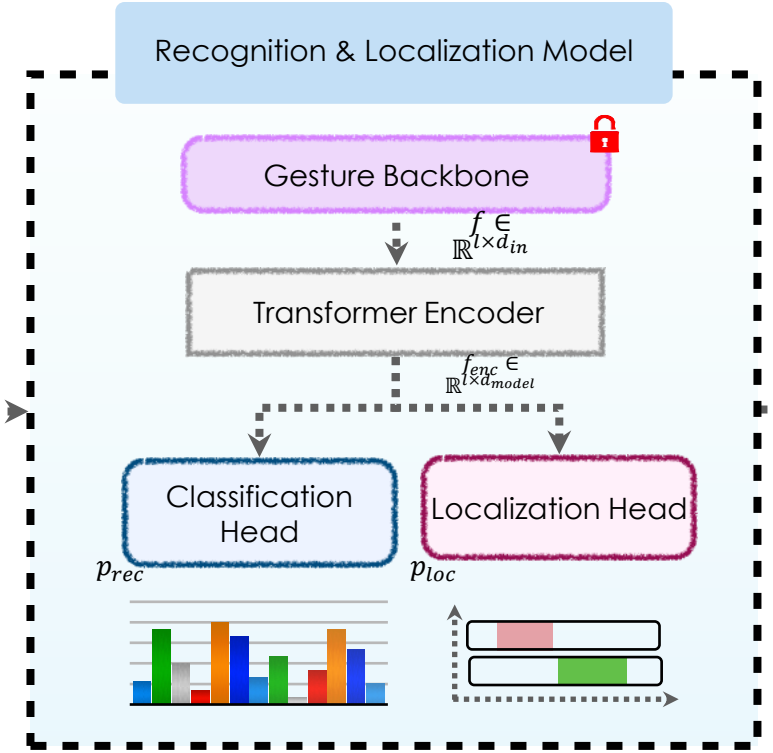
Scenario	Accuracy	Accuracy (high confidence)
Unseen test set	70.3	91.6
Unseen words	63.8	85.1

Task 2: Word Recognition & Localization: Task Description

Given a speaker video with a co-speech semantic gesture, predict:

- the word, and
- its temporal boundaries

Input: Video sequence

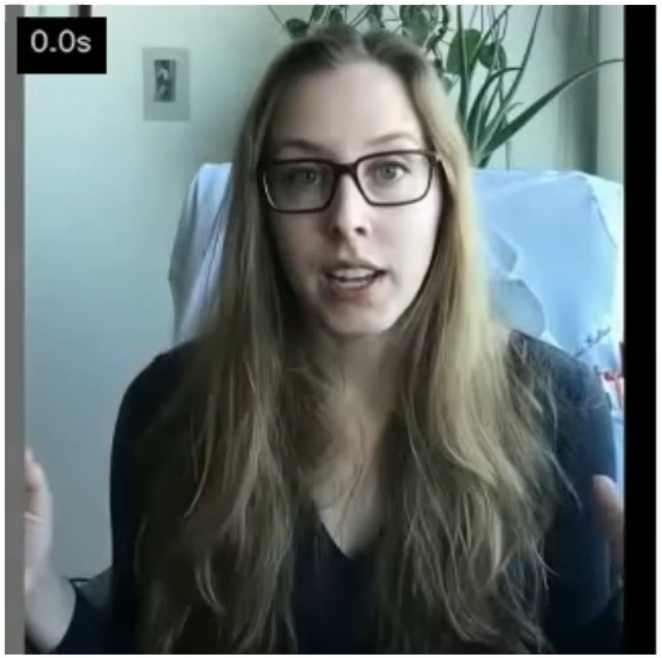


Output:

Word class: "round"

Word boundaries:
- start: 0.6sec
- end: 1.7 sec

Word Recognition & Localization: **Results** (GRW test set samples)



Note: Speech is played here only for reference, it is NOT used as input in any of the models.

Quantitative Results

On GRW Word Recognition and Localization test set (100 classes)

	Recognition			Localization
Scenario	**Rec@1**	**Rec@5**	**Rec@10**	**mIoU**
Random	1.00	5.00	10.00	0.20
Trained Model	18.6	37.5	52.2	0.66

Summary & take home messages

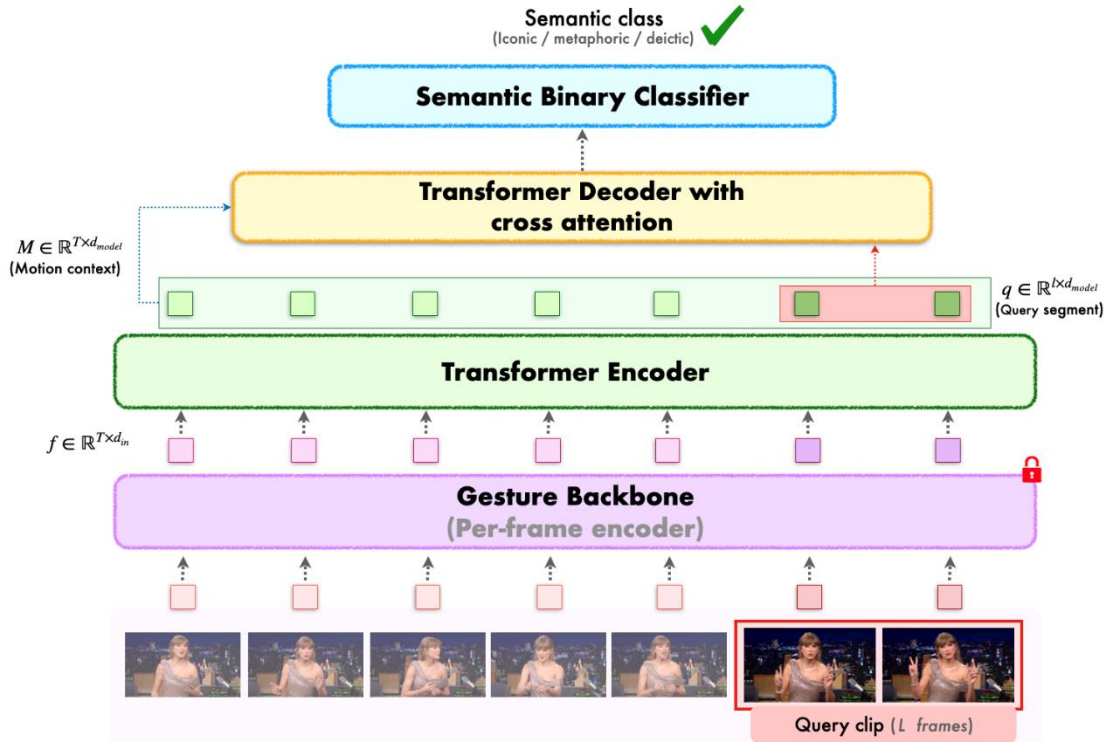
Recognizing co-speech gestures

- Less explored

- Gesture Recognition in the Wild large-scale dataset for training and evaluation

- Classify semantic gestures (iconic, metaphoric, deictic) from others (beat, random, no gestures)

- Recognize and localize the co-speech gesture



Recognizing Co-speech Gestures in-the-wild,
Sindhu B Hegde, K R Prajwal, Andrew Zisserman, arXiv 2026

Where next for co-speech Gestures?