Recent Developments in Sign Language Processing

towards realistic sign language machine translation

Zifan Jiang

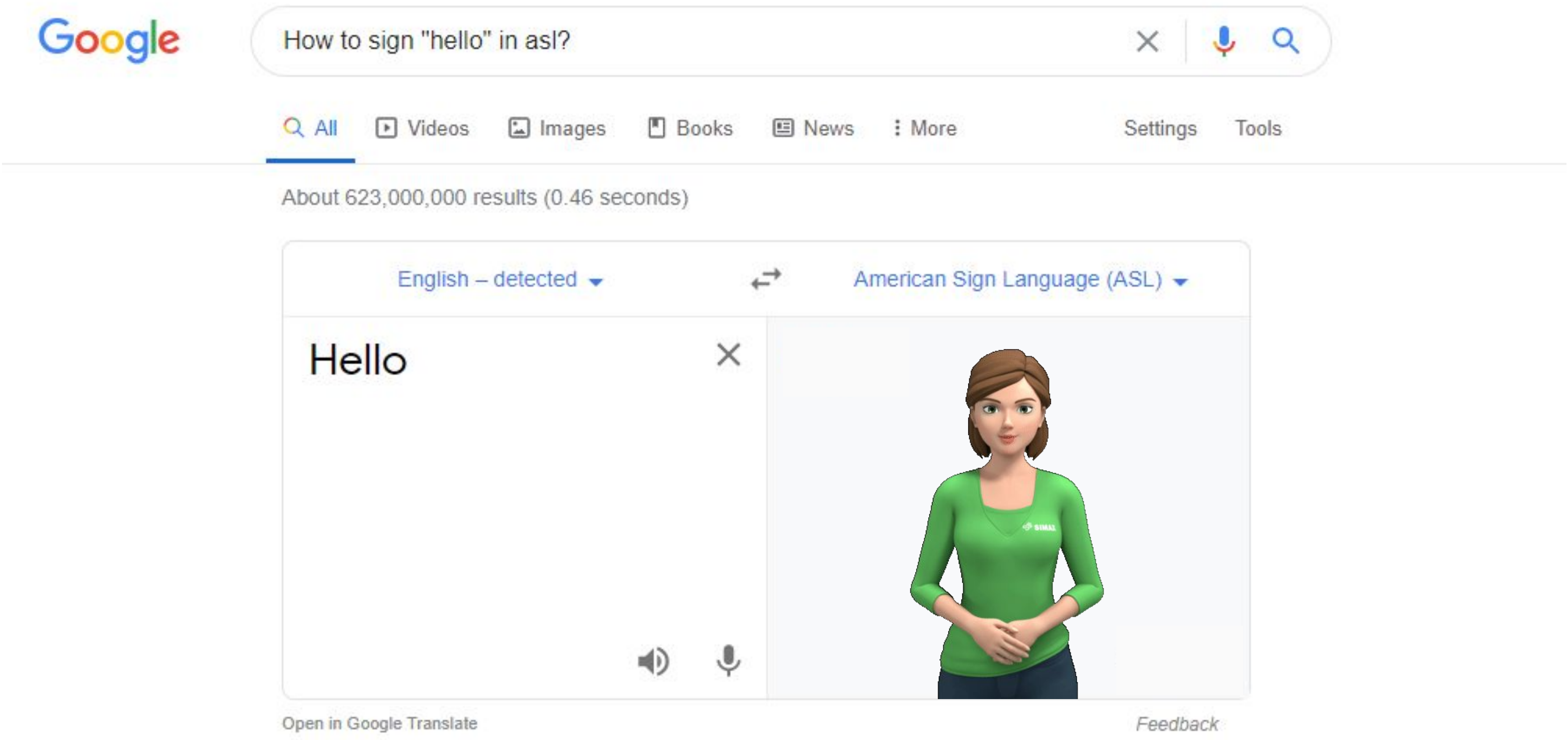
2024.02.20 Zurich

Who am I?

Zifan (子凡) [tsɹ̩^{3} fan^2] Jiang (蔣) [tɕjɑŋ^3]

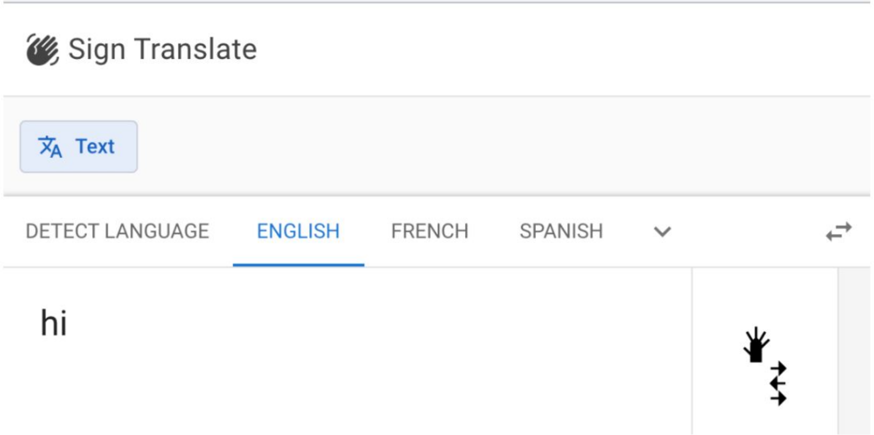
- PhD student at University of Zurich
- Funded by the IICT project
- Computer/data scientist & Web developer
- Computational linguist

(Goal?) of Existing Sign Language Works



Recap: my first work

Machine Translation between Spoken Languages and Signed Languages Represented in SignWriting



https://sign.mt/

University of Zurich[UZH]

Zifan Jiang, Amit Moryossef, Mathias Müller, Sarah Ebling
Department of Computational Linguistics

preprint code demo

May 2023 @ EACL & LoResMT

Outline

- WMT shared task on sign language translation
- Data for sign language processing
- Methodology for sign language processing
  - Segmentation
  - Alignment
  - Representation
- (interlude) Sign language processing 2024 and future
  - In the era of LLMs and deep pretrained models

WMT shared task on sign language translation

WMT-SLT

Home Motivation Schedule Participate Data Tools Calls Organizers FAQ Previous Versions

WMT-SLT 23

Second WMT shared task on sign language translation

News

01/08/2023	Participation instructions are now live.
28/07/2023	Our test set can now be downloaded.
26/06/2023	We shifted our remaining deadlines by two weeks, to give participants more time. See updated schedule.
22/06/2023	Our training data SRF can now be downloaded.
06/06/2023	Our training data Signsuisse can now be downloaded.
16/05/2023	Delayed release of training data for one more week
02/05/2023	Schedule is updated due to delays in data preparation.
22/03/2023	2023 Website is up. Last year's site can be found here.

https://www.wmt-slt.com/

WMT shared task on sign language translation

all		
Rank	Ave.	System
1	87.051	HUMAN
2-3	2.075	MSMUNICH
2-3	2.008	SLATTIC
4-5	0.520	UZH (baseline)
4-8	0.437	DFKI-MLT
5-8	0.339	DFKI-SLT
5-8	0.207	UPC
5-8	0.041	NJUPT-MTT

2022 edition

both domains		
Rank	Ave.	System
1	83.829	HUMAN
2	0.669	TTIC
3-5	0.024	CASIA
3-5	0.008	BASELINE
3-5	0.005	KNOWCOMP

2023 edition

What’s wrong?

- Data
  - Number of parallel examples: 10k << 1m
  - Quality: alignment, parallel vs. *comparable* data (>> 10k)

wmt-slt-data 23

	SRF training data 22	SRF training data 23
Number of episodes	29	771
Time span of episodes	March 2020 to March 2021	July 2014 to May 2021
Total duration videos	16 hours	437 hours
Total number of subtitles (before/after sentence segmentation)	14265 / 7071	354901 / 231834
Number of signers	3	4

train test

- Methodology
  - Transformers + ?
  - Tokenization/segmentation

Data

Swiss TV broadcast data

- https://www.wmt-slt.com/data
- https://sites.google.com/view/wmt-slt-v2022/data?authuser=0

The Signsuisse Lexicon

- https://www.sgb-fss.ch/signsuisse/

SwissSLi: the Multi-parallel Sign Language Corpus for Switzerland

- Under review @lrec-coling-2024

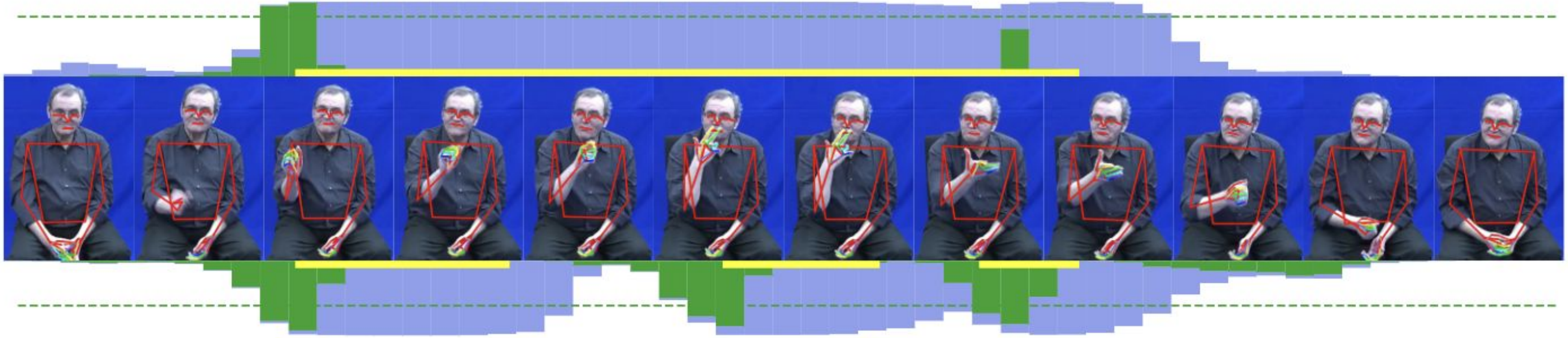
Methodology

More basic tools

- **Segmentation**
- Alignment
- Representation

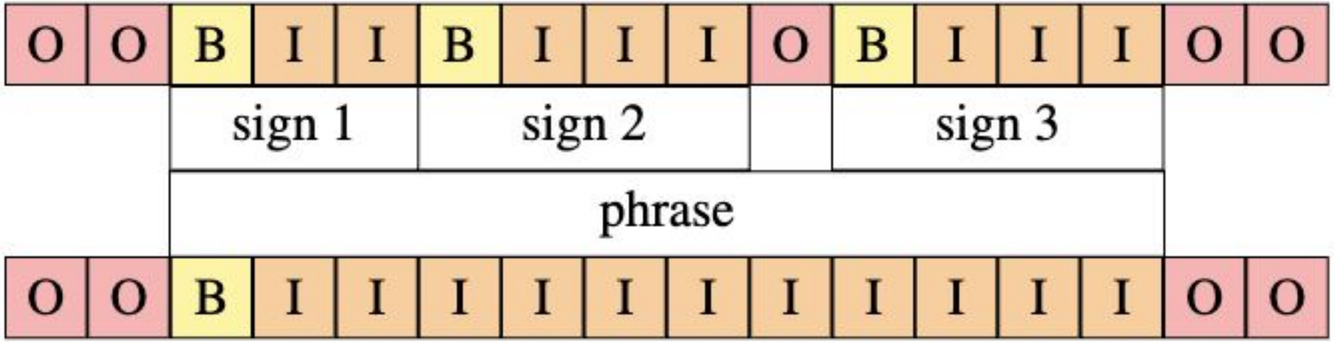
Linguistically Motivated Sign Language Segmentation

- Phrase-level
- Sign-level



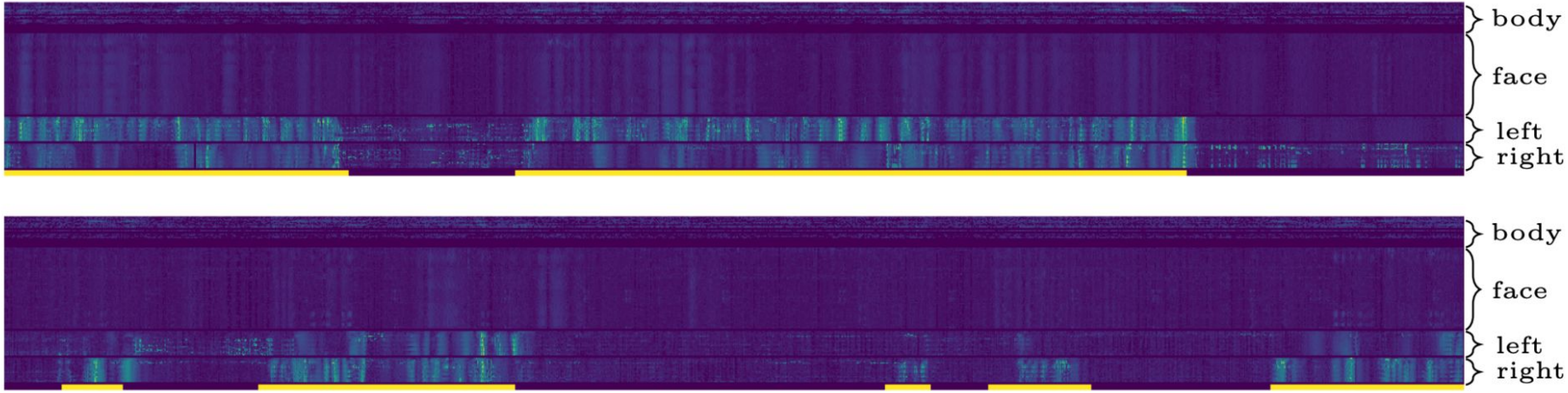
Linguistically Motivated Sign Language Segmentation

Labelling Strategy: 0/1 vs. BIO Tagging



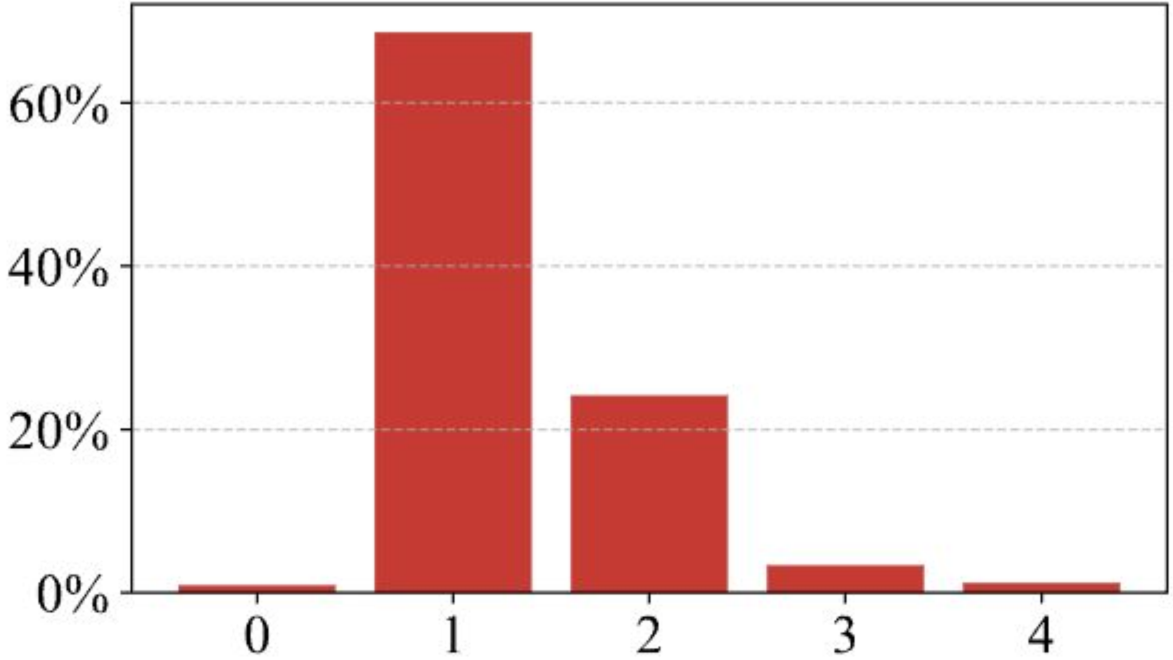
Per-frame classification of a sign language utterance following a BIO tagging scheme

Boundary of Phrases



Optical flow of a conversation between two signers in the Public DGS Corpus

Boundary of Signs



Number of hand shapes per sign in SignBank

3D Hand Normalization

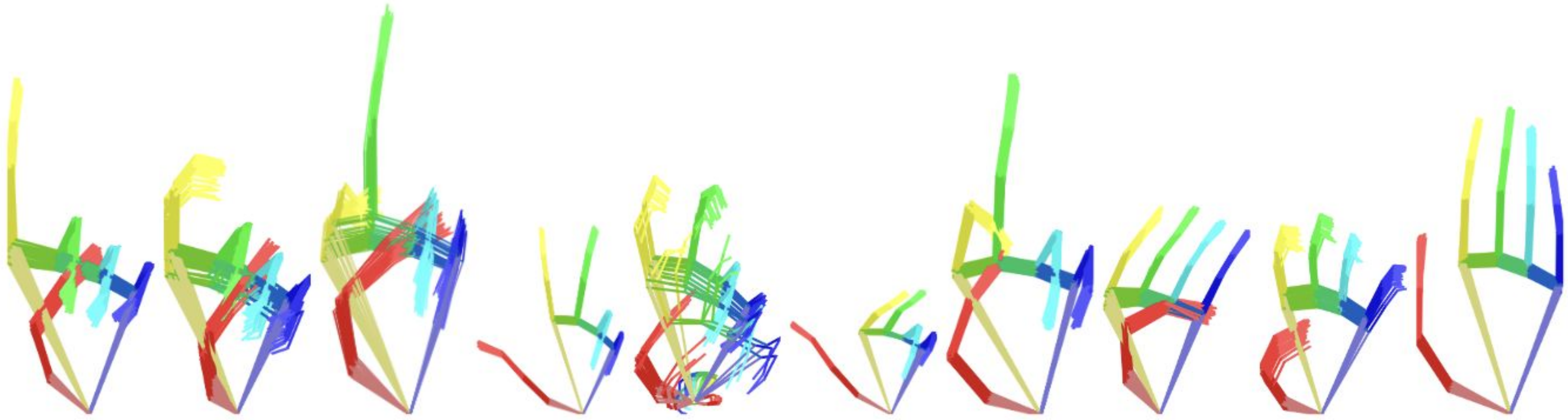


Figure 14: Visualizations of 10 hand shapes, each with 48 crops overlayed.

Segmentation + (isolated) recognition = translation?

- https://colab.research.google.com/drive/1CKIXVI3vP0NKZDZZ_I-Qb_wSHt2cw4VT#scrollTo=u3NuOI9PYx7h
- Limitation of glosses: word order, information loss ->

Considerations for meaningful sign language machine translation based on glosses

Mathias Müller[1], Zifan Jiang[1], Amit Moryossef[1,2], Annette Rios[1] and Sarah Ebling[1]
[1] Department of Computational Linguistics, University of Zurich, Switzerland
[2] Bar-Ilan University, Israel
{mmueller,jiang,rios,ebling}@cl.uzh.ch, amitmoryossef@gmail.com

Abstract

Automatic sign language processing is gaining popularity in Natural Language Processing (NLP) research (Yin et al., 2021). In machine translation (MT) in particular, sign language translation based on *glosses* is a prominent approach. In this paper, we review recent works on neural gloss translation. We find that limitations of glosses in general and limitations of specific datasets are not discussed in a trans-

Glosses (DSGS) KINDER FREUEN WARUM FERIEN NÄHER-KOMMEN
Translation (DE) Die Kinder freuen sich, weil die Ferien näher rücken.
Glosses (EN) (‘CHILDREN REJOICE WHY HOLIDAYS APPROACHING’)
Translation (EN) (‘The children are happy because the holidays are approaching.’)

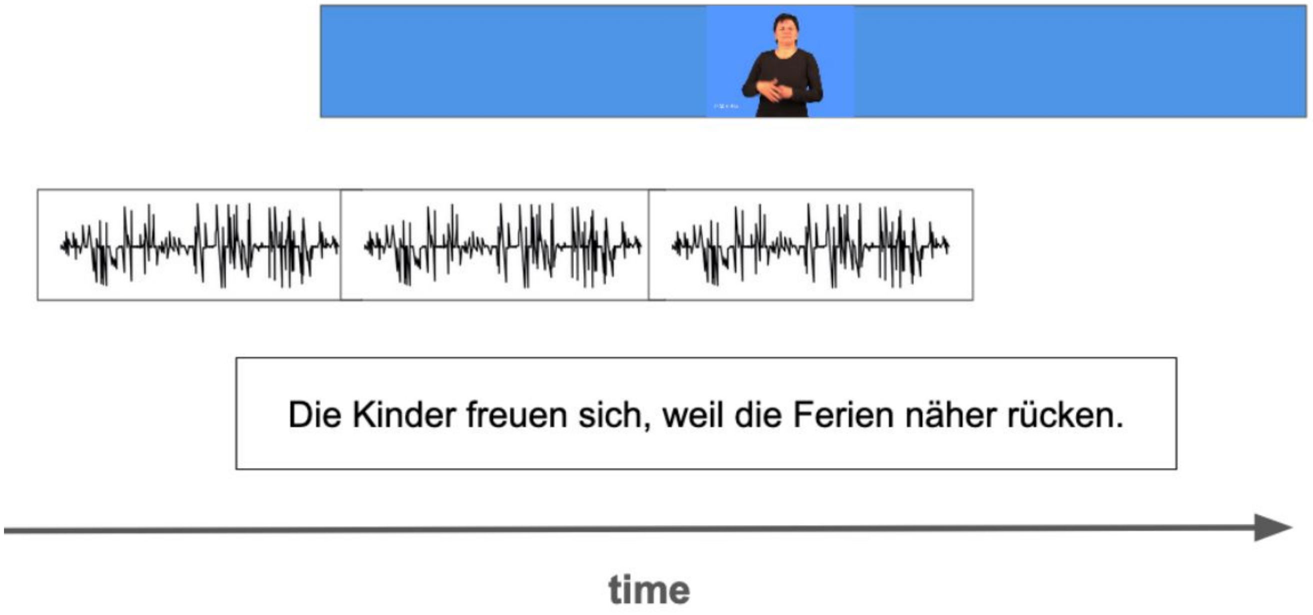
@ACL2023

Methodology

More basic tools

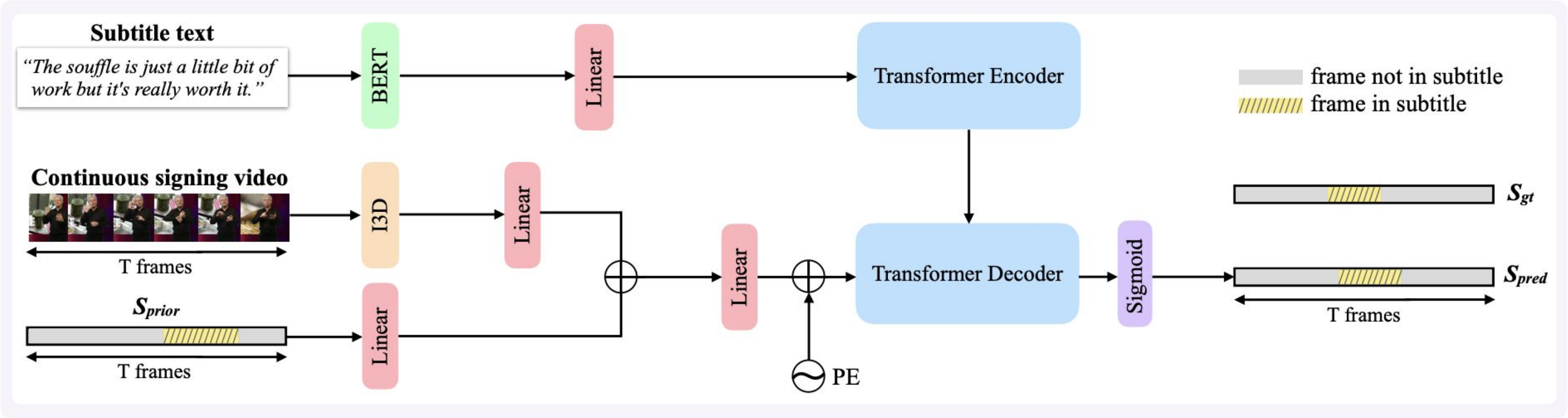
- Segmentation
- **Alignment**
- Representation

Alignment



https://www.wmt-slt.com/data

Alignment



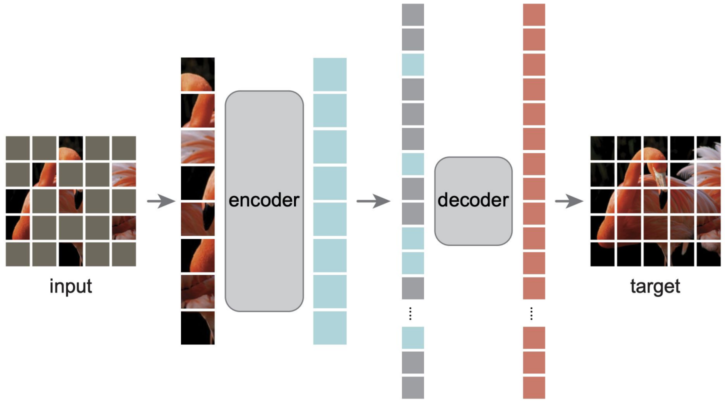
https://www.robots.ox.ac.uk/~vgg/research/bslalign/

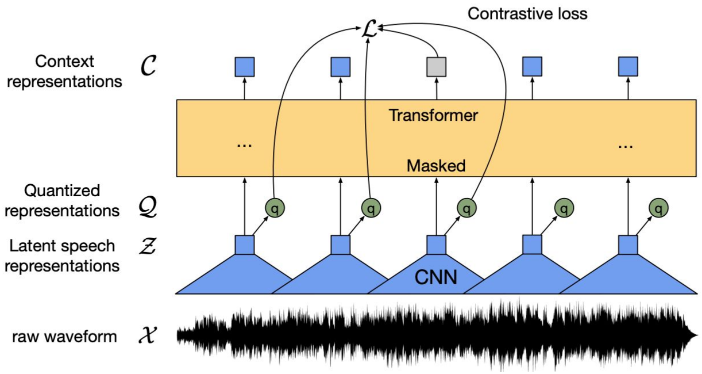
(interlude) Sign Language Processing 2024

In the era of LLMs and deep pretrained models

Self-supervised deep pretrained models (on huge data)

- Text: BERT, GPT, etc.
- Image: masked autoencoders (MAE)
    - based on ViT
- Speech: wav2vec 2.0
    - Quantization
- Video: InternVideo
    - Too expensive to train?





Weak supervision from text - CLIP

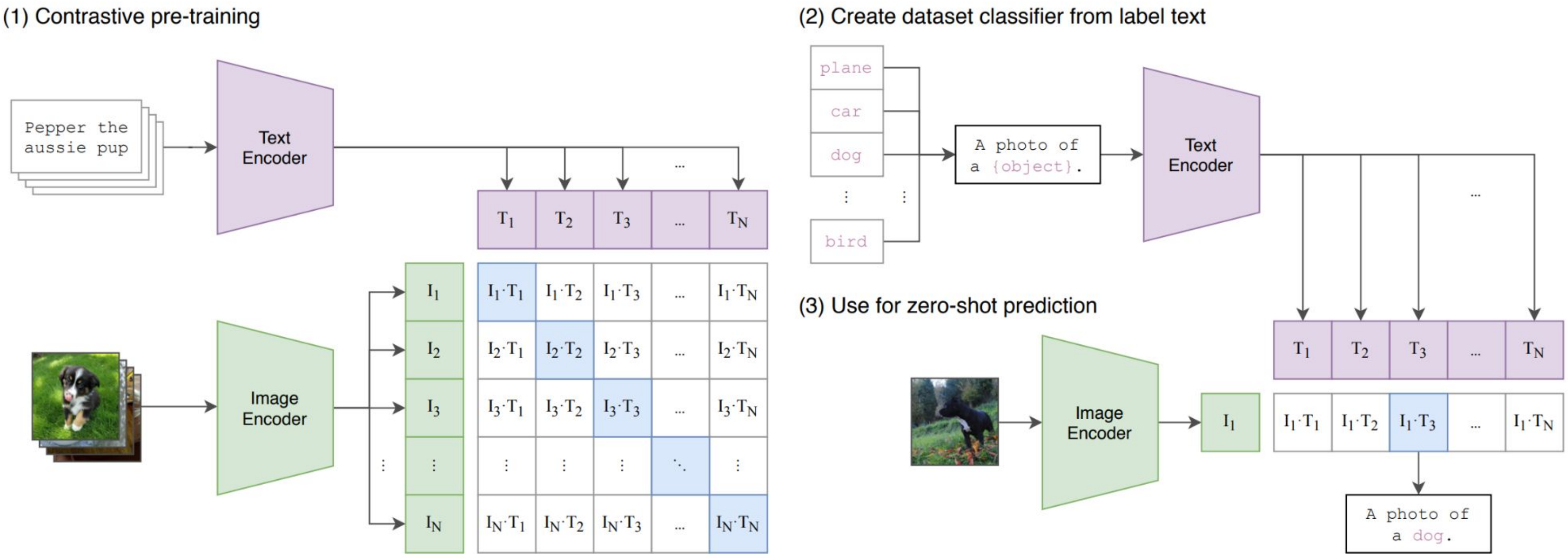


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

SignCLIP: our solution to alignment (sign language to text)

- Adapted from a VideoCLIP model
- Data scale
  - HowTo100M videos (duration of each is ~6.5 minutes with ~110 clip-text pairs)
  - Now collecting a few hundred thousand isolated ASL sign examples
  - Spreadthesign: 600k
- Data representation
  - 10-second video
  - Dimension reduction
  - Spatial vs. temporal
- Usage
  - Language identification
  - Recognition/retrieval
  - Segmentation/alignment
  - Glossed-based translation
  - Quality estimation

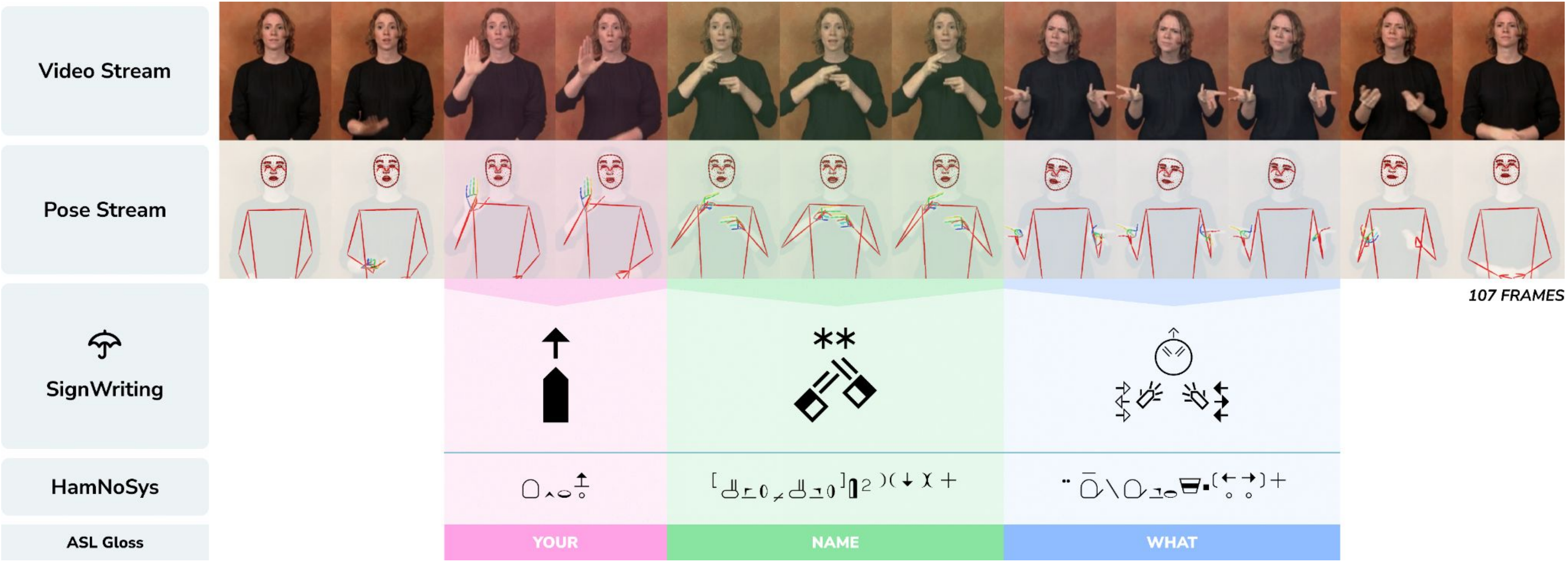
Encoder	Temporal dim.	Spatial dim.
Original video	10x30	640×480x3
S3D (pretrained on HowTo100M)	10	512
I3D (pretrained on BSL)	10	1024
MediaPipe Holistic	10x30	543
SignVQNet	10	1024

Methodology

More basic tools

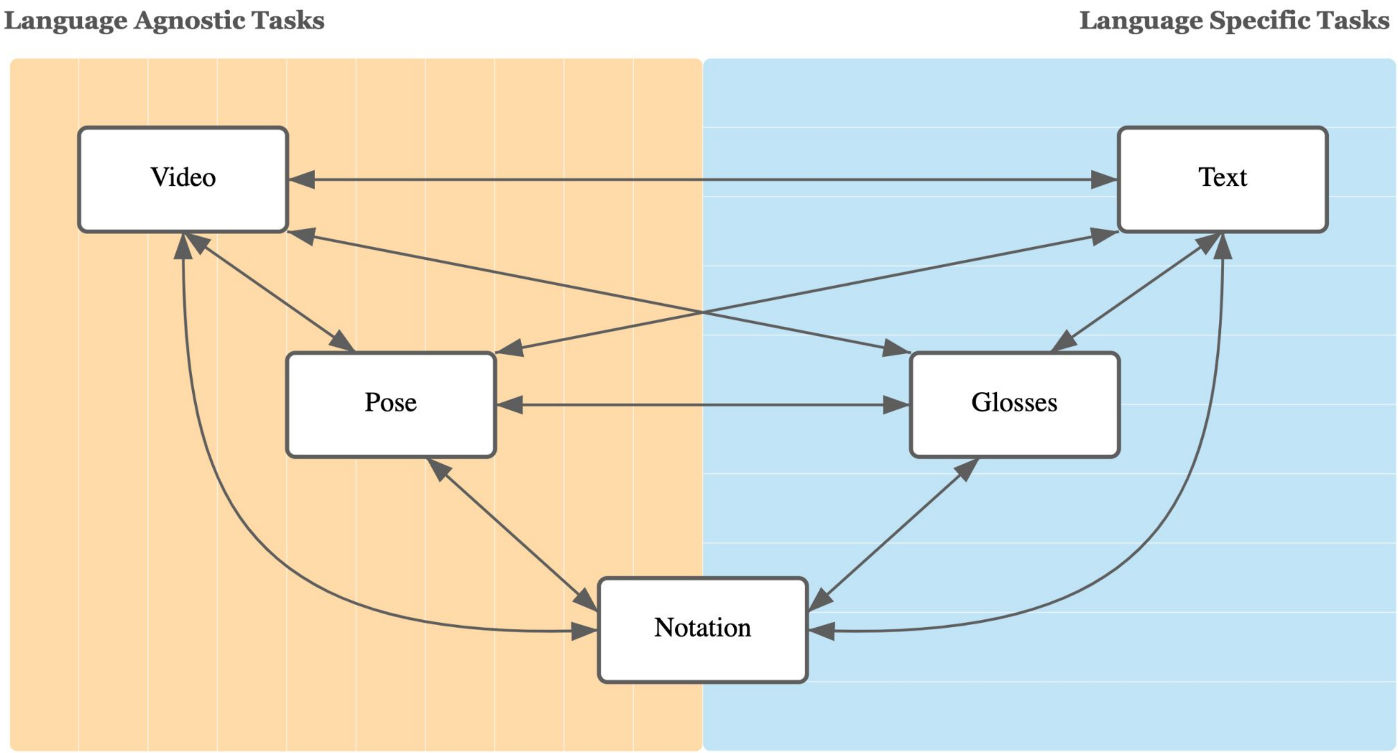
- Segmentation
- Alignment
- **Representation**

Representations of Signed Languages



https://research.sign.mt/

Representations of Signed Languages



https://research.sign.mt/

SignVQ: our solution to representation

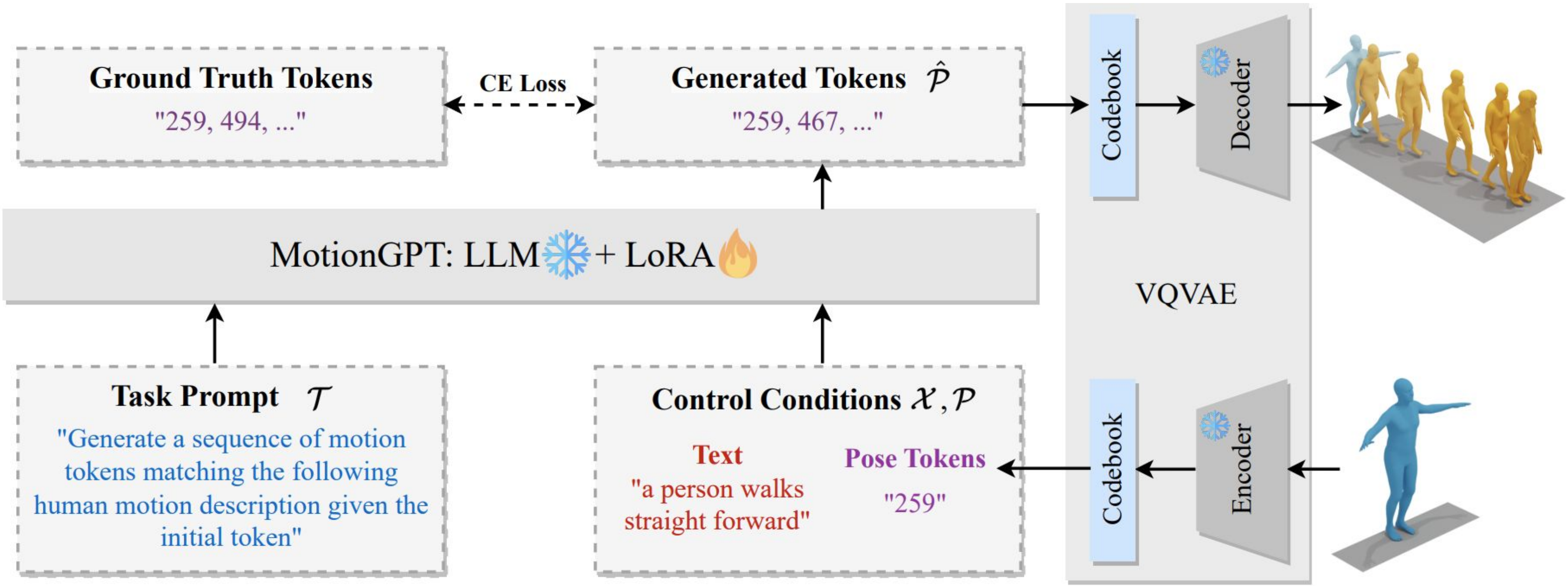
Existing work

- [Autoregressive Sign Language Production: a Gloss-Free Approach with Discrete Representations](http://nlpcl.kaist.ac.kr/~projects/signvqnet)
- [SignAvatars: A Large-scale 3D Sign Language Holistic Motion Dataset and Benchmark](https://signavatars.github.io/)
- Lee's new work: Learning Sub-Lexical Components to Represent Sign Language

Ours

- Sign MediaPipe VQ

MotionGPT



Ground Truth Tokens
"259, 494, ..."
CE Loss
Generated Tokens $\hat{\mathcal{P}}$
"259, 467, ..."
Codebook
Decoder
MotionGPT: LLM + LoRA
VQVAE
Task Prompt $\mathcal{T}$
"Generate a sequence of motion tokens matching the following human motion description given the initial token"
Control Conditions $\mathcal{X}, \mathcal{P}$
Text
"a person walks straight forward"
Pose Tokens
"259"
Codebook
Encoder