Multimodal LLMs =
LLMs and other than text modalities

Dominik Macháček, Josef Vopička (MAMA AI), Peter Polák, Andrei Manea

5.5.2025

Spolufinancováno Evropskou unií

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Supported by the grant CZ.02.01.01/00/23_020/0008518 (Jazykověda, umělá inteligence a jazykové a řečové technologie: od výzkumu k aplikacím).

Charles University

ÚFAL

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Lesson Plan

- Intro ... 10min
- Text-to-Speech Synthesis ... Guest: Josef Vopička, MAMA AI ... 30min
- Audio in LLMs ... Peter Polák
- Vision in LLMs ... Andrei Manea
- Sign Language LLM ... Dominik Macháček

Intro: What are Multimodal LLMs?

What are Multimodal LLMs?

- What is a language?
  - In automata and grammars:
  - In linguistics: a system for communication that associates symbols to meanings, and has certain properties
    - Natural vs. artificial, human vs. animal, language vs. dialect, spoken vs. written vs. signed...
- What is a LM?
- What is a Large LM?
- Modalities of language?
  - Text = digitized, computer-encoded writing
  - Text/Vision: handwriting, typed text in raster images
  - Vision: body language, face expressions, lipreading
    - Sign language
  - Audio: speech, sound, music?
  - Touch? Body and brain sensors?
  - Mix: paralinguistics = non-verbal means of language
    - Prosody, rhythm, stress, intonation
    - Gestures, body language
    - Communication through clothing, hairstyle, make up, perfume, ...
    - Conscious or unconscious (e.g. mocking an accent / having an accent)

Multimodal LLM (for us this lesson):

- Definition of multimodal LLM, for us this lesson:

  Large deep learning model that processes language in other than text modality.

- Examples (to be presented today)
  - Text-to-speech synthesis
  - Speech-to-text transcription and translation
  - Speech+Vision+Text LLM, Vision+Text LLM
  - Sign Language LLM
- Not covered today but relevant:
  - OCR = optical character recognition
  - Image/video/audio generation from natural language
  - Speech-to-speech LLMs

Intro: Wy Multimodal LLMs?

Why Multimodal LLMs?

Why should LLMs work with speech/vision/ … from **user perspective**?

- More natural interaction
- Accessibility
- Primarily non-written languages
- Complementarity with text
- New applications

Speech-to-speech Translation for Unwritten Language

- ~7,000 living languages in the world
- 3,500 are primarily spoken and don’t have a widely used writing system
- Taiwanese Hokkien:
  - ~13.5M speakers
  - No uniform writing system
  - Meta, 2022: “Speech-to-speech translation for a real-world unwritten language”

Why Multimodal LLMs?

Why should LLMs work with speech/vision/ … from **technical perspective**?

- Using information beyond text
  - Gender in voice, lip reading
  - Prosody
    - Intonation, stress, and rhythm
  - Non-verbal language
  - Sentiment
  - Environment
- Avoiding error propagation
- Enable new applications

Challenges of the Modalities

Data Acquisition

Sparse Representations

Difficulties of Data Acquisition

	Text 😇	Speech 😣	Sign Language 😱
anonymization => who gives authorisation?	easy => many	pseudo => volunteers	❌ no – face needed => professionals
parallel corpora use and share	extremely large: Wikipedia, parliaments, Bible, OPUS, …	large: parliaments, CommonVoice, LibriSpeech, …	❌ very limited: How2Sign – 80h, Phoenix
open data (dare to) use but don't share	easy web scraping: media, the entire Web	demanding, but OK: media, Youtube, …	very limited: ❌ identify + download
linguistic resources	annotated corpora, treebanks, lexicons, wordnets, …	small, but OK: use ASR + text resources	❌ very small and limited

How much data space for **one word**?

	Text	**Speech** (audio)	**Sign language** (video)
one word (English)	5 characters	330 ms	~330 ms?
typical representation	UTF-8	16 kHz wav	110 kB/s + codecs → 30 fps, 1280x720, RGB
size of representation	**5 Bytes** 😍	**10 kB** 😨	**37 kB** 😱 + codecs→ **27 MB !!!** 😱 😱 😱

Sparse

Text-to-Speech Synthesis

Josef Vopička, MAMA AI

Audio in LLMs

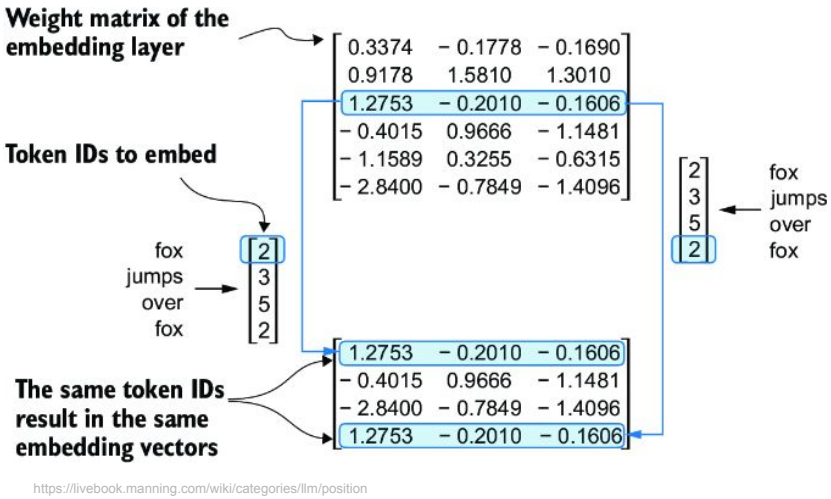
Peter Polák
5.5.2025

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Speech Representation for NNs

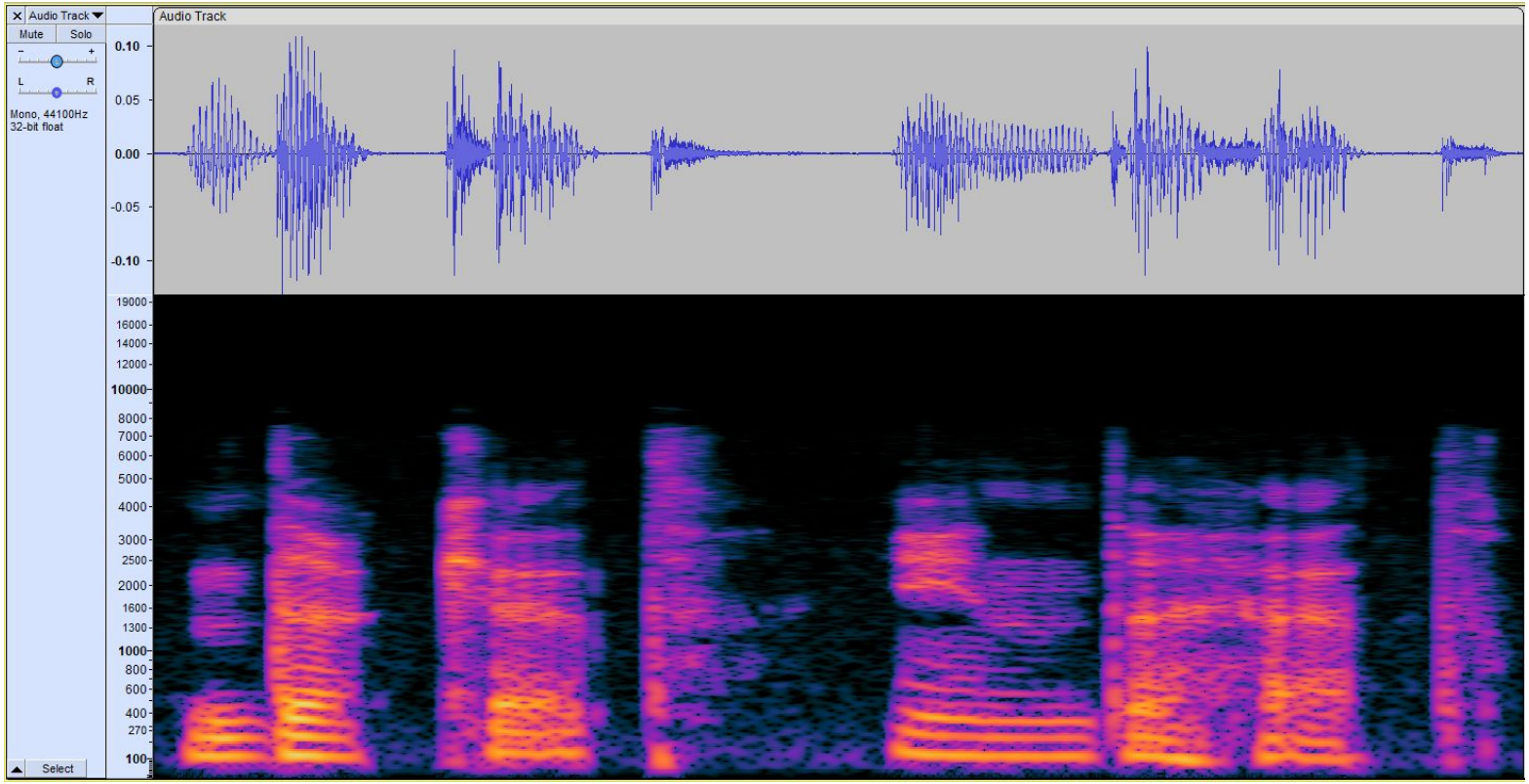
Speech Representations: Recap from Text

- How neural networks read?
- Tokenization
  - Break the text into smaller units = tokens [characters, words, sub-words]
- Translate tokens to indices in a vocabulary
- Embedding Layer
  - Translate each index into a dense vector



https://livebook.manning.com/wiki/categories/llm/position

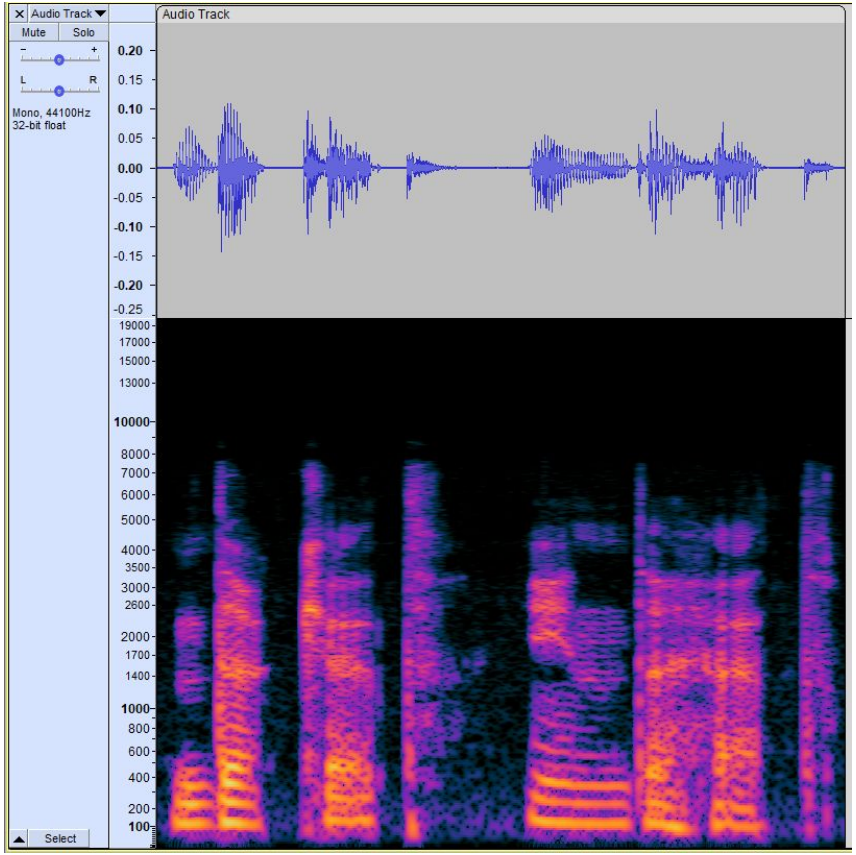
Speech Representations



the cat in the hat

Speech Representations

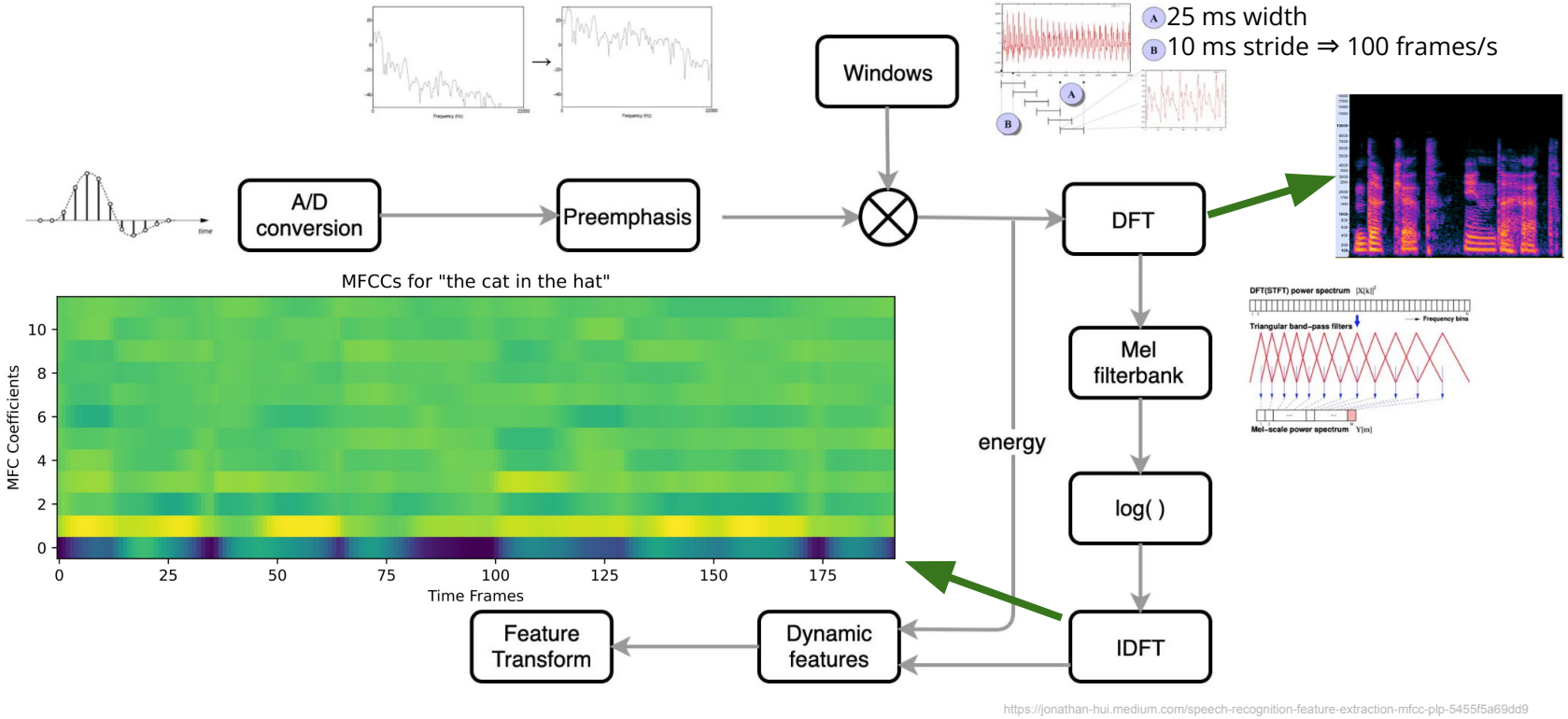
- How do we represent sound in computer?
  - Human speech 100 - 4 kHz, better to 8 kHz
  - 16 kHz wav = 16k floats/s
- How NNs understand speech?
  - 1 word ~ 330 ms ~ 5280 floats
  - Two approaches:
    - Raw audio (some tricks needed)
    - MFCCs



the cat in the hat

Speech Representations: MFCCs

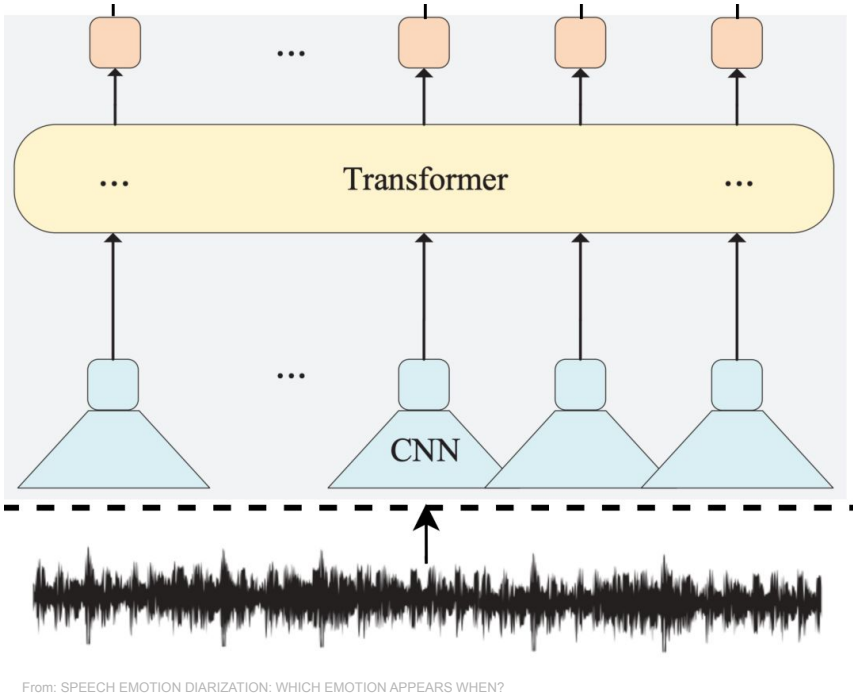
- Mel Frequency Cepstrum Coefficients



https://jonathan-hui.medium.com/speech-recognition-feature-extraction-mfcc-plp-5455f5a69dd9

Speech Representations: Direct Approach

- Feed directly 16 kHz to NN
  - We want to use Transformers - where is the problem?
    - Complexity of self-attention - $O(n^2)$
  - Solution
    - Downsample long input with CNNs
    - CNN serves as feature encoder
      - Similar to MFCCs
      - But the representation is **learned from data**
  - Typically a part of pre-trained models
    - Wav2vec 2.0, HuBERT, WavLM



From: SPEECH EMOTION DIARIZATION: WHICH EMOTION APPEARS WHEN?

Speech Representations: Comparison

- MFCCs
  - 👍 Efficient
  - 👍 Interpretable
  - 👎 Limited features
    - Ideal for ASR, not ER
  - 👎 Not robust to noise

- Direct approach
  - 👍 More complex features
    - speaker characteristics, emotions, and environmental noise
  - 🤔 Depends on training data
    - Can be robust
  - 👎 Computational cost
  - 👎 Interpretability

Plugging Speech to LLMs

Plugging Speech to LLMs: Typical Approach

- Embed audio with some encoder
  - Conformer, HuBERT, ...
- Shrink audio representations
  - Length Adapter
  - A few Transformer layers + CNN
- Interleave audio and prompts
- Examples:
  - Llama 3
  - Qwen2-Audio
  - Phi-4-Multimodal

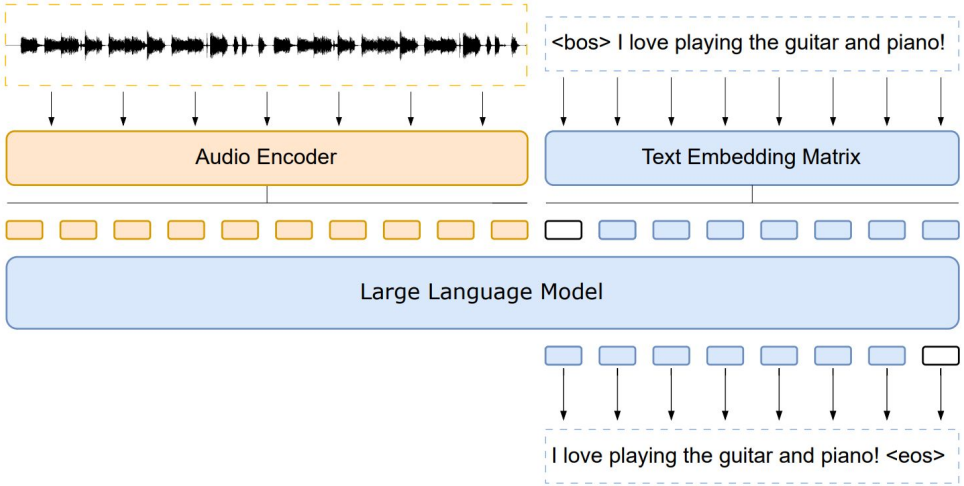


Figure 2: Model architecture. The embedding sequence generated from the audio encoder is directly prepended to the text embeddings sequence. This is directly fed into the decoder-only LLM, tasked with predicting the next token. The LLM can be frozen, adapted with parameter efficient approaches such as LoRA or fully finetuned. This work will investigate the former two.

From: Prompting Large Language Models with Speech Recognition Abilities

Plugging Speech to LLMs: Other Approaches

- Dual Encoders
  - ASR, ST Encoder
  - Non-speech Audio Encoder
  - SALMONN
- Cross-attention
  - Audio Flamingo

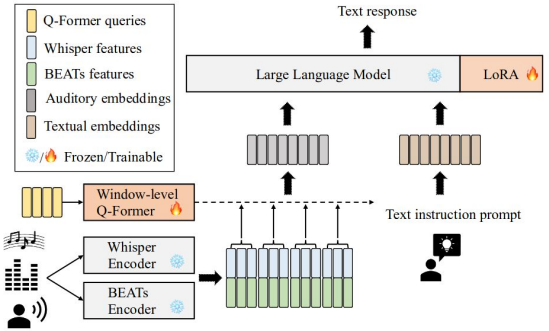


Figure 1: The model architecture of SALMONN. A window-level Q-Former is used as the connection module to fuse the outputs from a Whisper speech encoder and a BEATs audio encoder as augmented audio tokens, which are aligned with the LLM input space. The LoRA adaptor aligns the augmented LLM input space with its output space. The text prompt is used to instruct SALMONN to answer open-ended questions about the general audio inputs and the answers are in the LLM text responses. The LLM and encoders are kept frozen while the rest can be updated in training.

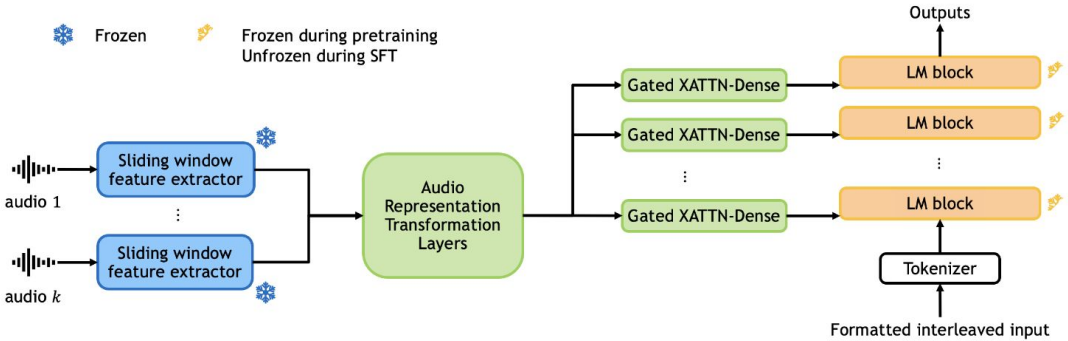
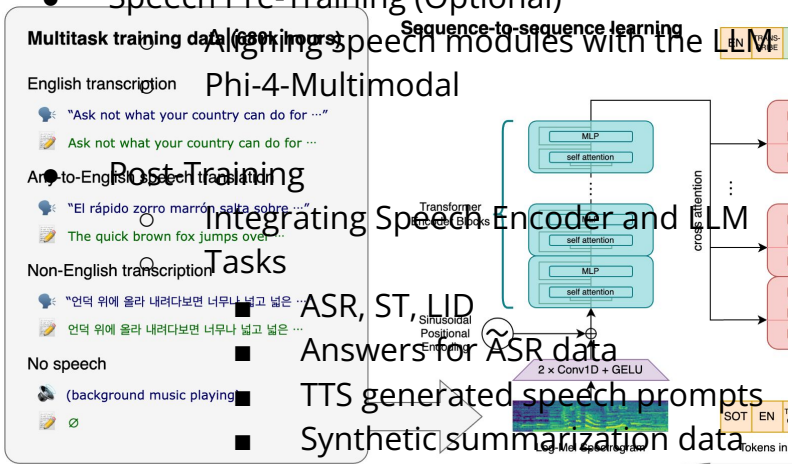


Figure 2. Neural architecture of Audio Flamingo. It takes interleaved audio and text as input and outputs free-form text.

Speech in LLMs: Training Pipeline

- Speech Encoder Pre-Training
  - Supervised
    - ASR, ST, …
  - Unsupervised
    - Bert-like, …
- Speech Pre-Training (Optional)
  - Aligning speech modules with the LLM
  - Phi-4-Multimodal
- Post-Training
  - Integrating Speech Encoder and LLM
  - Tasks
    - ASR, ST, LID
    - Answers for ASR data
    - TTS generated speech prompts
    - Synthetic summarization data
    - Audio understanding
    - DPO



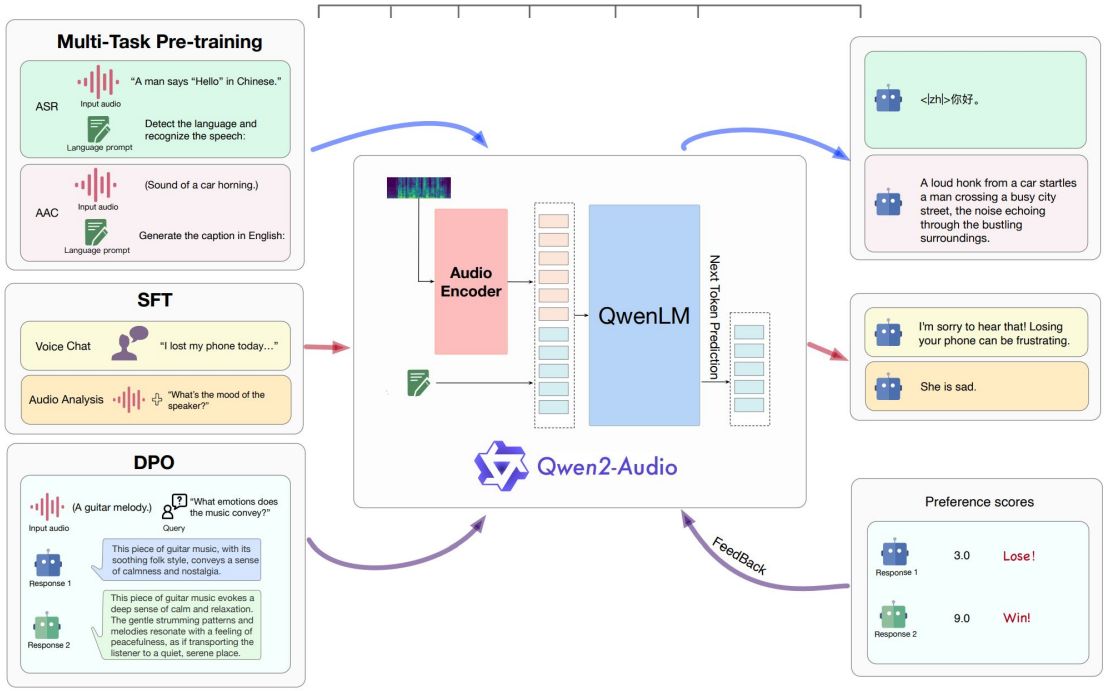


Figure 2: The overview of three-stage training process of Qwen2-Audio.

the nearest vector in the codebook. The pre-training objective is for the ASR encoder to take the masked input signals and predict the labels corresponding to the masked part provided by the random-projection quantizer.

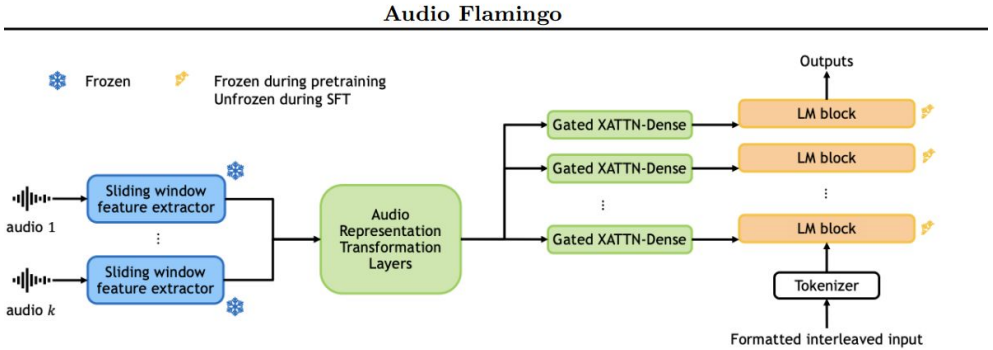
Speech in LLMs: What to Train?

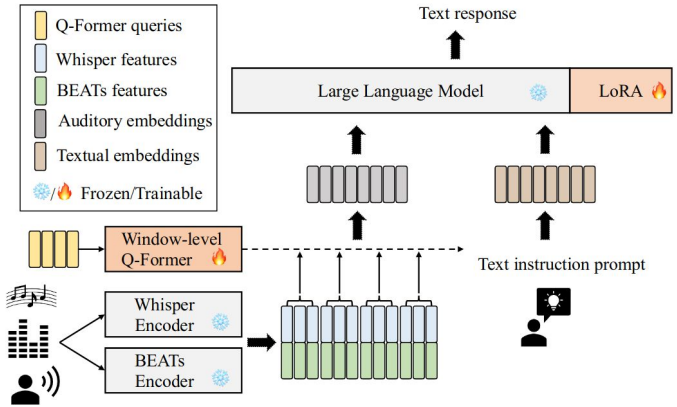
- Which parts to train?
  - LLMs are huge and difficult to train
  - Training with speech is even more demanding

- Options
  - Train entire model

  - Train a portion only
    - Encoder + Length Adapter
      - Llama 3
    - Audio Representation Layers + Cross-attention + LLM
      - Audio Flamingo

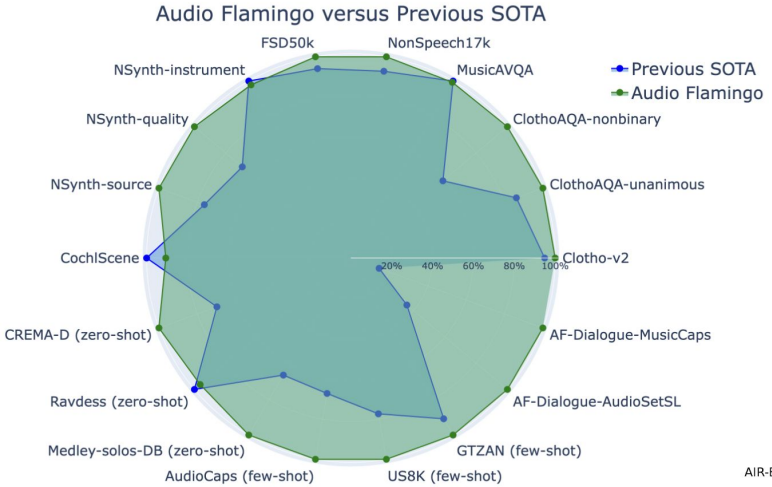
  - Other tricks
    - LoRA
      - Phi-4-Multimodal
      - SpeechVerse

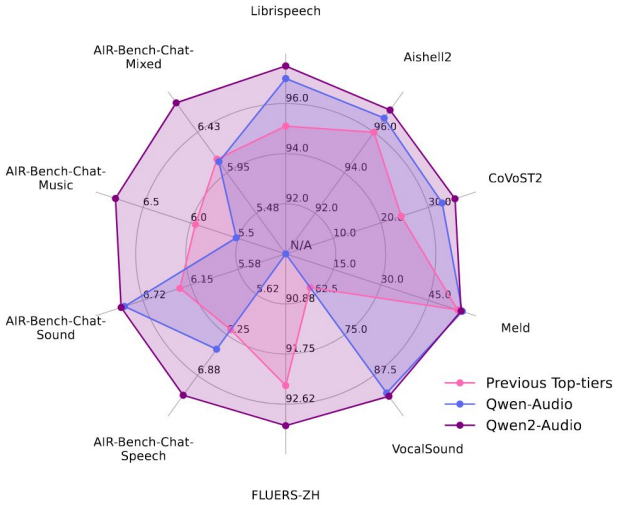




Speech in LLMs: Results

- Achieve SOTA results
- Generalize to untrained tasks
- Reasoning
- Better integration of speech is still needed





Vision Representations in LLMs

Andrei Manea
5.5.2025

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Overview

- Patch Embeddings: ViT, CLIP-ViT, SigLIP
  - Q-Former extension: BLIP-2, mBLIP
- Discrete tokens: TiTok
- Integrating visual tokens into LLMs
- Vision-Language Tasks:
  - ImageText-to-Text Generation
  - Visual Grounded Reasoning

Patch Embeddings (ViT)

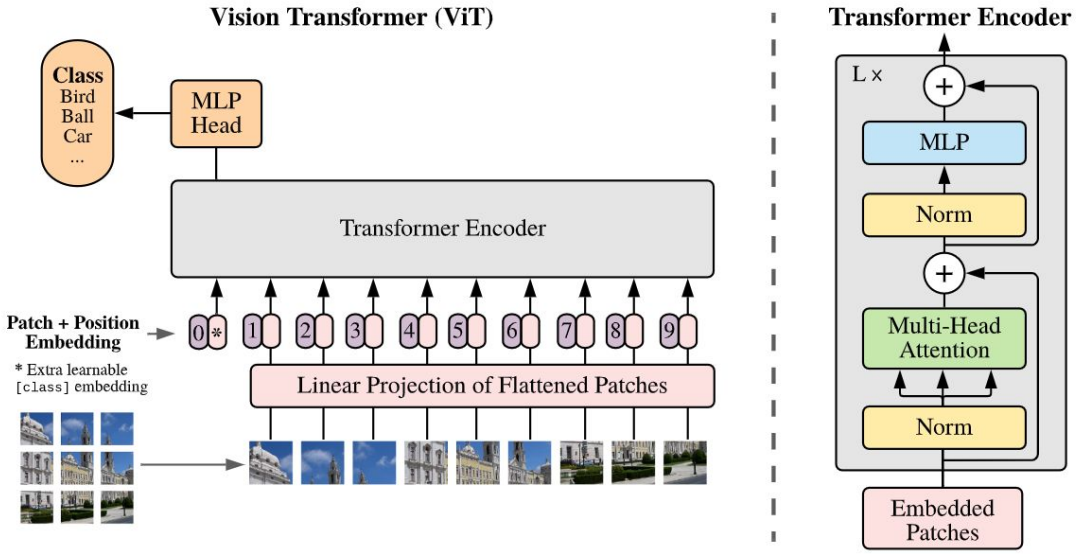


Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification token" to the sequence. The illustration of the Transformer encoder was inspired by Vaswani et al. (2017).

Patch Embeddings (ViT)

- (P, P) is the resolution of each image patch $\rightarrow$ $N = HW/P^2$
- Similar to BERT, **[class]** token serves as the image representation

- Objectives:
  - Pre-training: **Masked Patch Prediction** - predicting 3-bit RGB, mean color (i.e., 512 colors in total) of every corrupted patch.
  - Fine-tuning: **Image classification**

Patch Embeddings (CLIP-ViT)

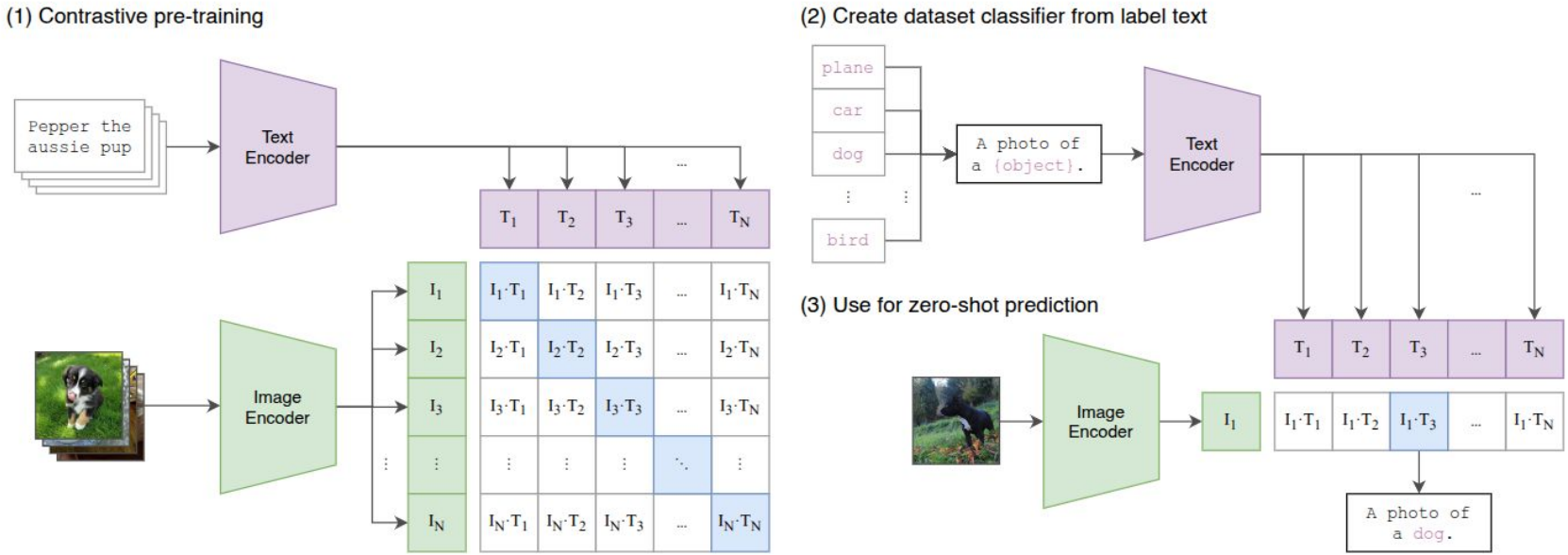


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the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

Patch Embeddings (CLIP-ViT)

- Contrastive pre-training: Considering cosine similarity, the goal is:
  - Getting closer the positive samples
  - Pushing away the negative samples
- Cross Entropy with normalization from both directions -> avg

```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Figure 3. Numpy-like pseudocode for the core of an implementation of CLIP.

Patch Embeddings (SigLIP)

- Alternative to contrastive pre-training
- The sigmoid loss operates solely on image-text pairs and does not require a global view of the pairwise similarities for normalization.

- “We find that the sigmoid loss performs significantly better than the softmax loss when the batch size < 16k.”

Algorithm 1 Sigmoid loss pseudo-implementation.

```
# img_emb      : image model embedding [n, dim]
# txt_emb      : text model embedding [n, dim]
# t_prime, b   : learnable temperature and bias
# n            : mini-batch size

t = exp(t_prime)
zimg = l2_normalize(img_emb)
ztxt = l2_normalize(txt_emb)
logits = dot(zimg, ztxt.T) * t + b
labels = 2 * eye(n) - ones(n) # -1 with diagonal 1
l = -sum(log_sigmoid(labels * logits)) / n
```

Q-Former (mBLIP)

- An additional encoder-only transformer with 32 learned query tokens as input. It contextualizes the query tokens – via the cross-attention mechanism – with the Patch Embeddings encoded by a large (frozen) ViT.

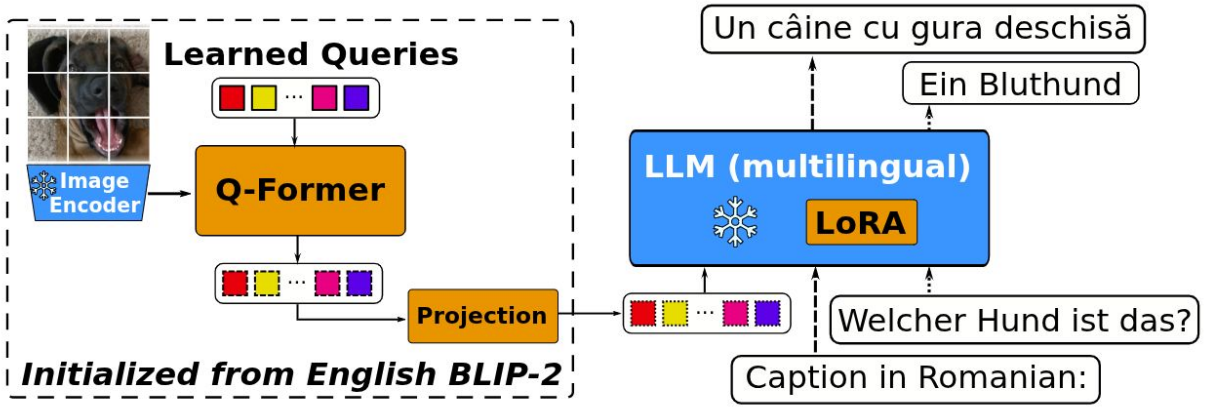
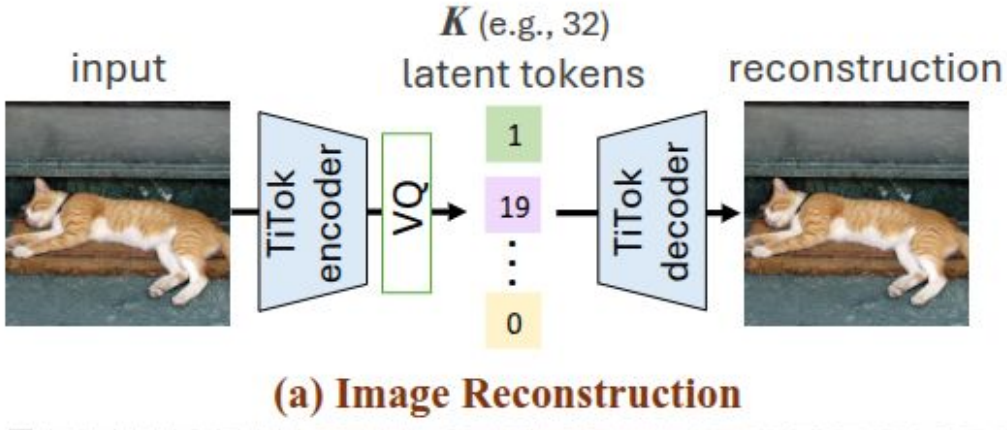


Figure 1: The mBLIP architecture: A Q-Former encodes the image in learned query tokens which are projected to the LLM space. We initialize the Q-Former from a BLIP-2 model and *re-align* it to the multilingual LLM with a multilingual task mix. The image encoder and LLM (aside from LoRA weights) are frozen during training.

Discrete Tokens (TiTok)

Latent representation is learned via MaskGIT, in 2 stages:

1. Tokenization Stage: compress images into discrete latent space, using Encoder, Codebook Quantization and Decoder



(a) Image Reconstruction

Discrete Tokens (TiTok)

- Codebook Quantization (KNN-like):

f. Subsequently, each embedding $z \in \mathbb{R}^D$ is mapped (via the vector quantizer $Quant$) to the nearest code $c_i \in \mathbb{R}^D$ in a learnable codebook $\mathbb{C} \in \mathbb{R}^{N \times D}$, comprising N codes. Formally, we have:

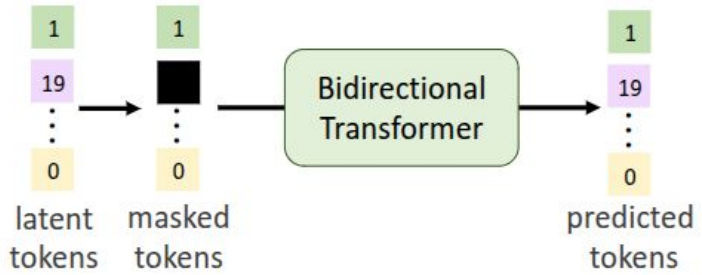
$$Quant(z) = c_i, \text{ where } i = \underset{j \in \{1,2,\ldots,N\}}{\operatorname{argmin}} \|z - c_j\|_2. \quad (1)$$

Discrete Tokens (TiTok)

Latent representation is learned via MaskGIT, in 2 stages:

2. Masked-Generation Stage: A random ratio of the latent tokens are replaced with mask tokens. Then, a bidirectional transformer predicts the corresponding discrete token ID of those masked tokens.

The latent tokens are also updated in this training.



(b) Image Generation

Discrete Tokens (TiTok)

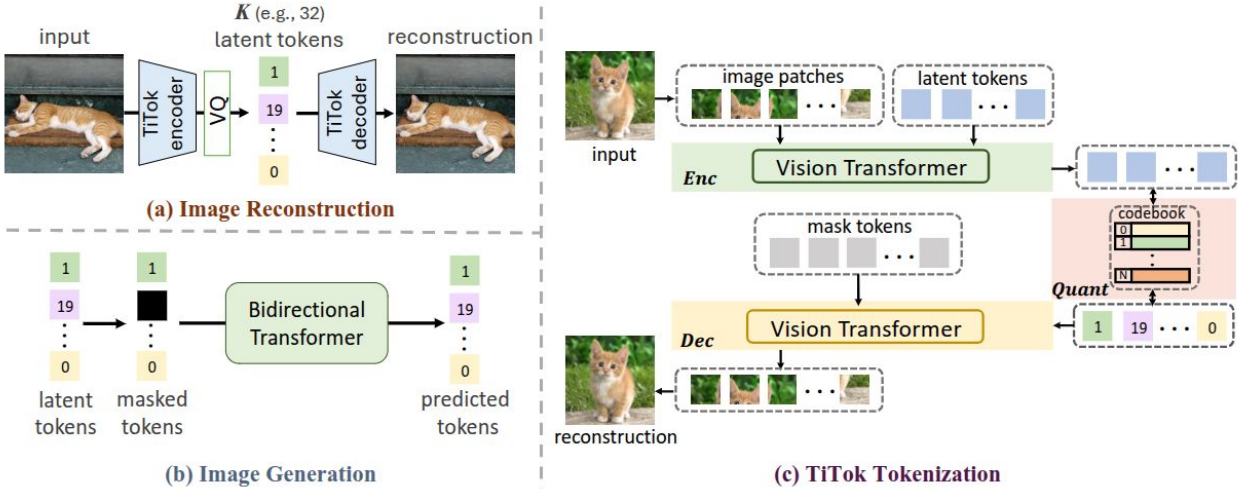


Figure 3: **Illustration of image reconstruction (a) and generation (b) with the TiTok framework (c).** TiTok contains an encoder Enc, a quantizer $Quant$, and a decoder Dec. Image patches, along with a few (*e.g.*, 32) latent tokens, are passed through the Vision Transformer (ViT) encoder. The latent tokens are then vector-quantized. The quantized tokens, along with the mask tokens [15, 24], are fed to the ViT decoder to reconstruct the image.

Integrating Visual Representations

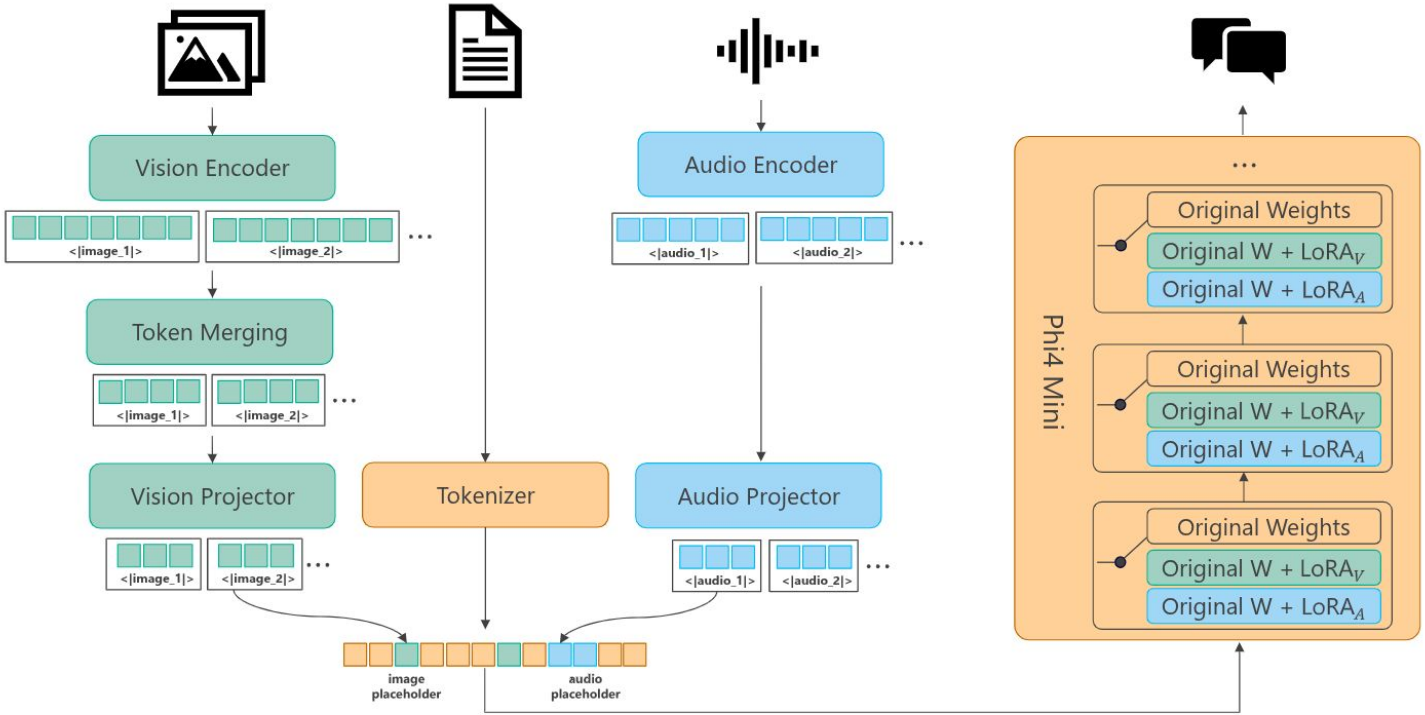


Figure 1: A overview of the Multimodal architecture for `Phi-4-Multimodal`

Vision-Language Tasks

- ImageText-to-Text Generation
- Evaluation: MT metrics

98 Most_SNP

Q: Kde sa tento most nachádza? *Where is this bridge located?*
A: V Bratislave. *In Bratislava.*

36 Sokolov

Q: V jakém městě se nachází tento zámek? *In which city is this castle located?*
A: V Sokolově. *In Sokolov.*

---

37 Laterna_magika

Q: Jak se jmenuje divadlo na obrázku? *What is the name of the theater in the picture?*
A: Laterna magika. *Laterna magika.*

Vision-Language Tasks

- Visual Grounded Reasoning - classification based

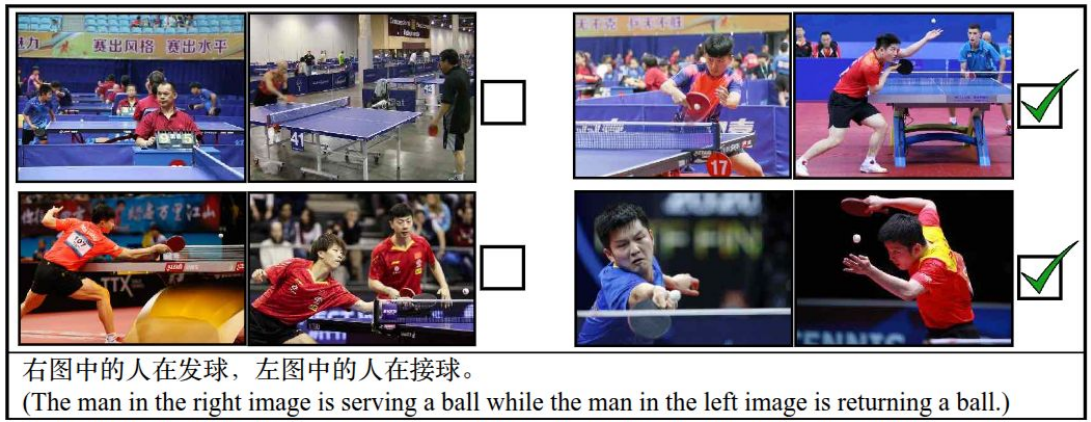


Figure 3: **Left**: For each annotation instance, eight images are randomly picked from the image set of a concept and randomly paired into four pairs. Annotators then write a caption that is True for two pairs but False for the other two. **Right**: Labels are hidden and a different set of annotators will relabel them.

Sign Language Translation and LLM

Dominik Macháček
5.5.2025

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

(American) Sign Language Translation + LLM Assistant

ASL

Translate

So for example she would say:
“Today’s sign is ‘what’ ?”
That’s teaching.

English text

(American) Sign Language Translation + LLM Assistant

ASL

Translate

+

Answer

So for example she would say:
"Today's sign is 'what' ?"
That's teaching.

USER: *What does she want?*
ASSISTANT: ***Nothing, she explains something about American Sign Language.***

English text

Sign Languages

- **9M** (2.8% of population) **in USA** reports using sign lang./signs to communicate

  [ref. Mitchell and Young, 2022]

  - Deaf
  - Hard of hearing
  - Family members
  - Professionals, linguists
  - “Foreign lang.” learners
- ASL: appx **1M** L1+L2 users

  [ref: En wikipedia]

How: Interdisciplinary Team

- 17 members, 5 continents, 6 institutes

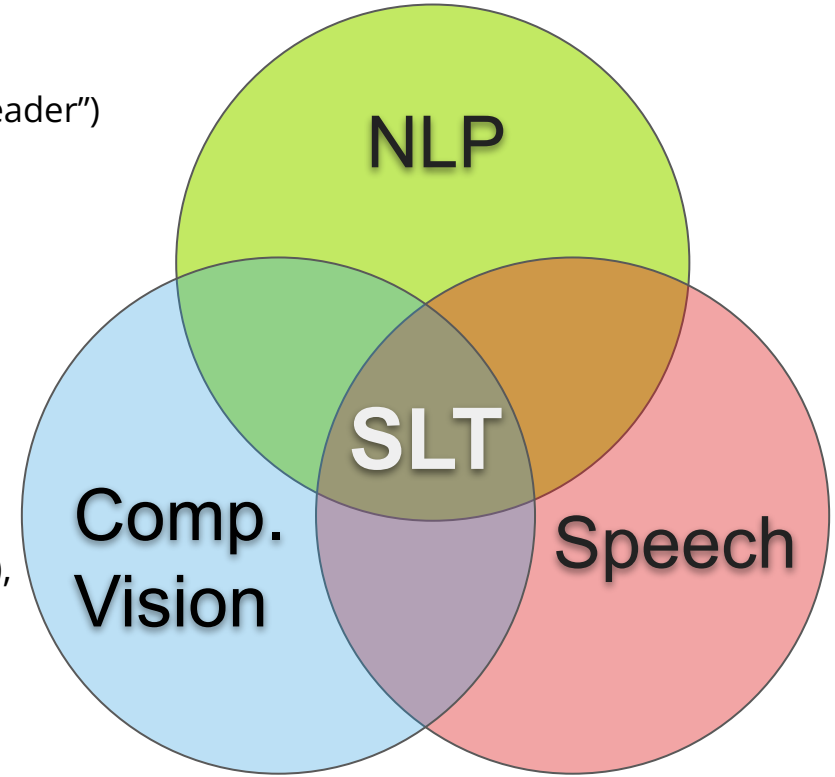
  5 seniors, 3 fresh PhDs, 5 grads, 4 undergr. (+1 “Meta-leader”)
- University of West Bohemia

  CV: **Marek Hrúz**, Ivan Gruber + 4
- Johns Hopkins University, USA

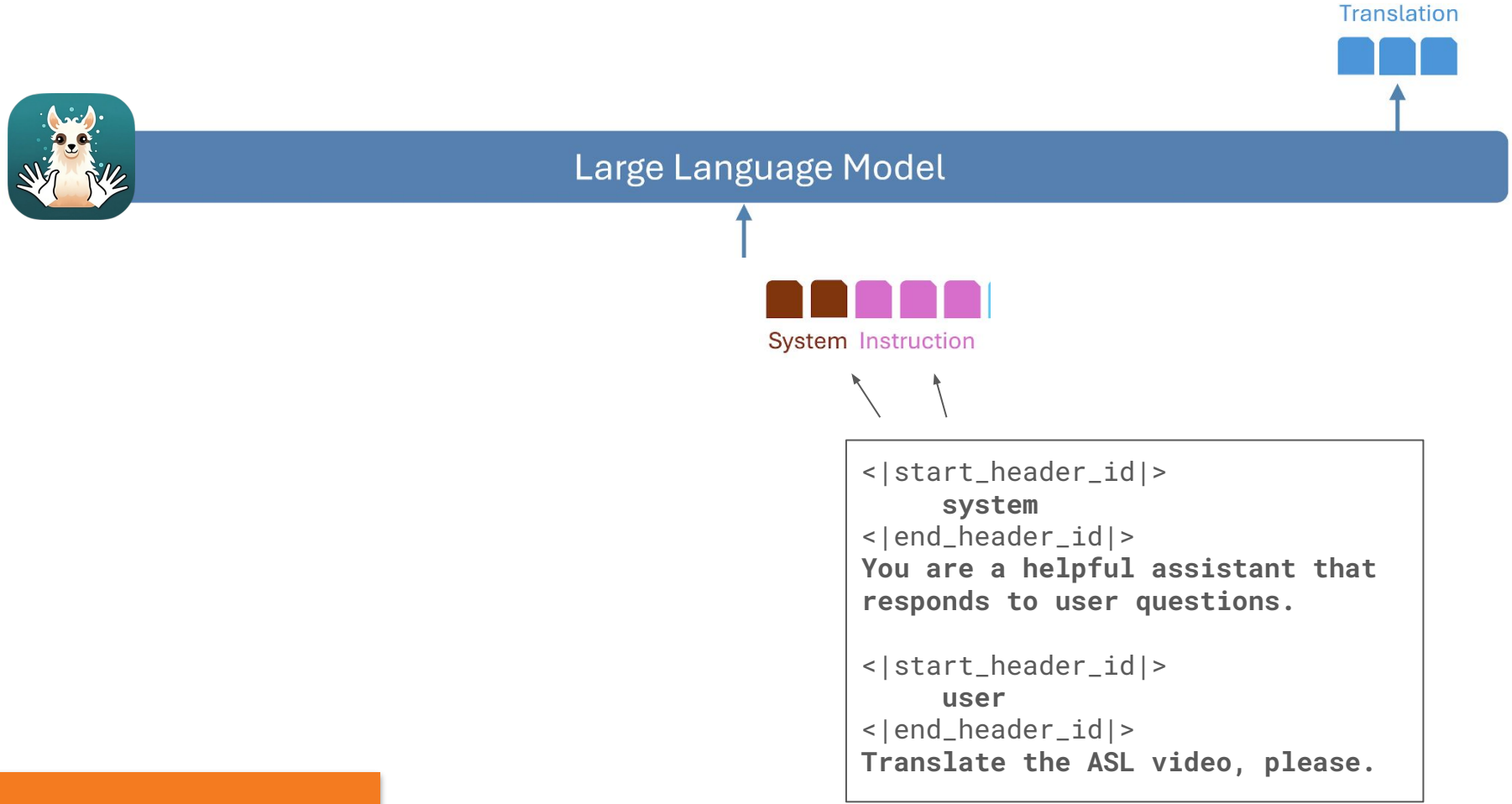
  NLP (MT, LLM): Kevin Duh, **Xuan Zhang** + 1
- Bogazici University, Turkey

  prof. Lale Akarun (CV), Murat Saraclar (Speech/NLP),

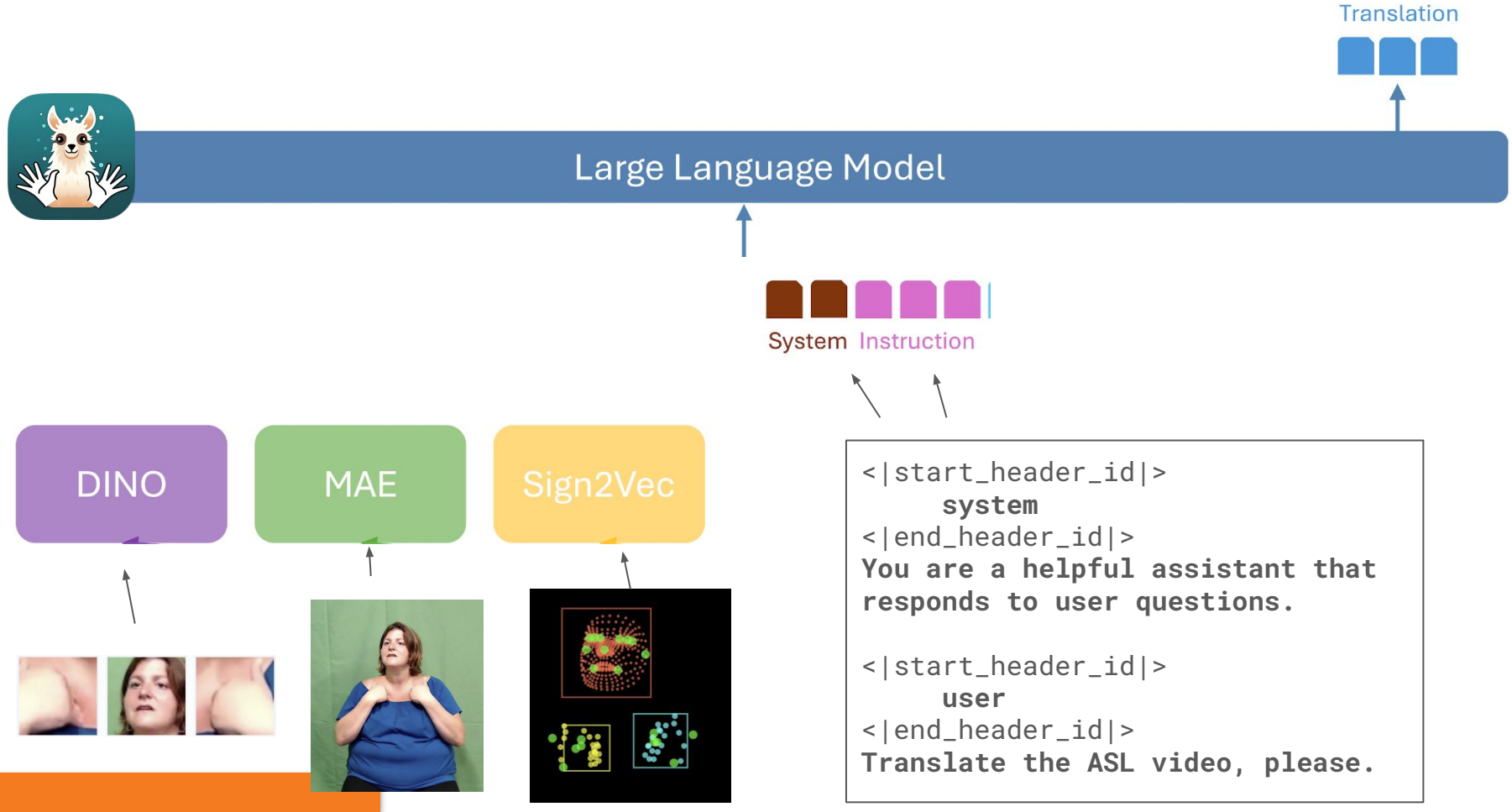
  Karahan Sahin, Bolaji Yusuf (Speech)
- + ÚFAL + 2, incl. ASL user



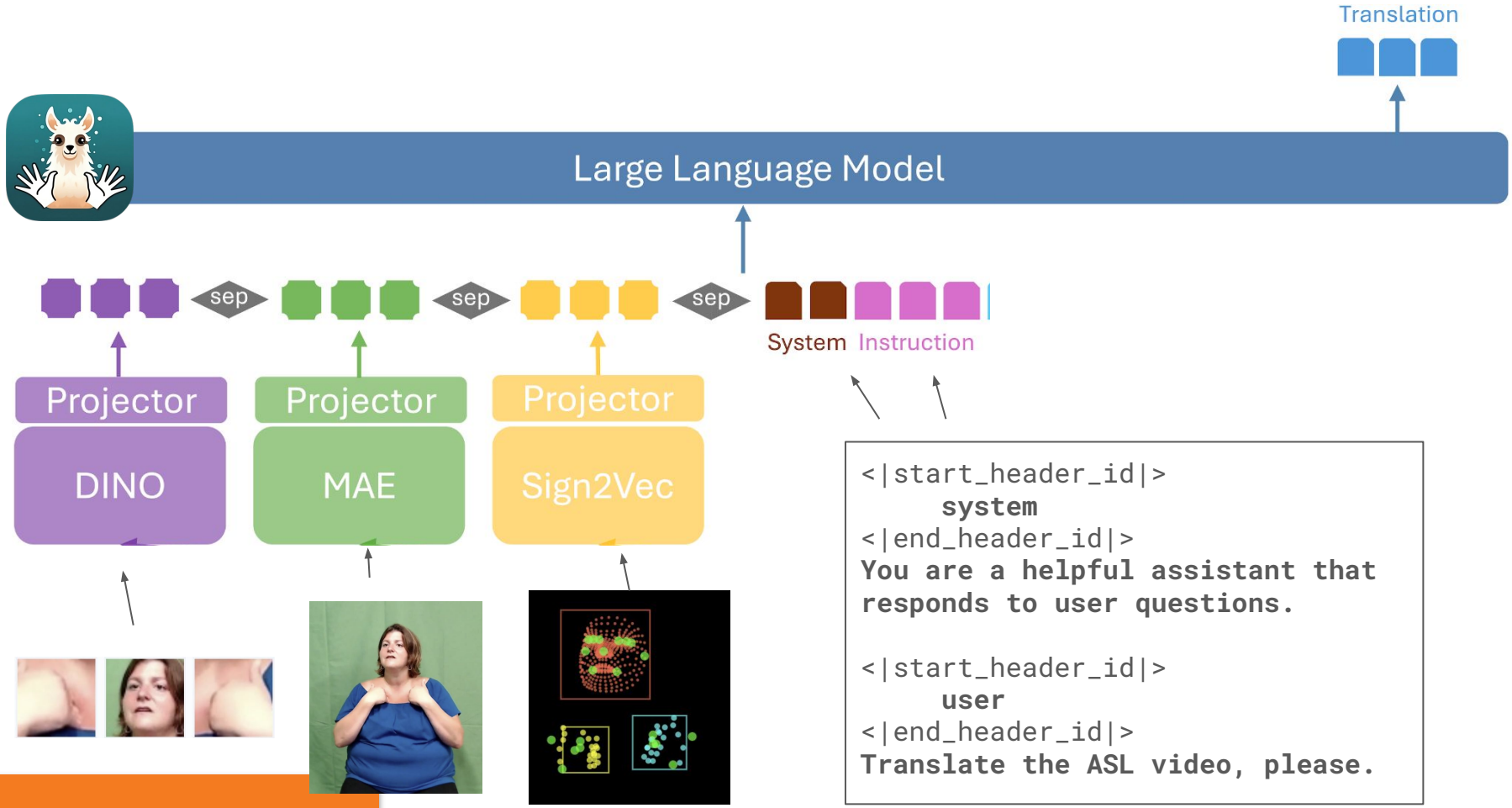
How: SignLaVa



How: SignLaVa

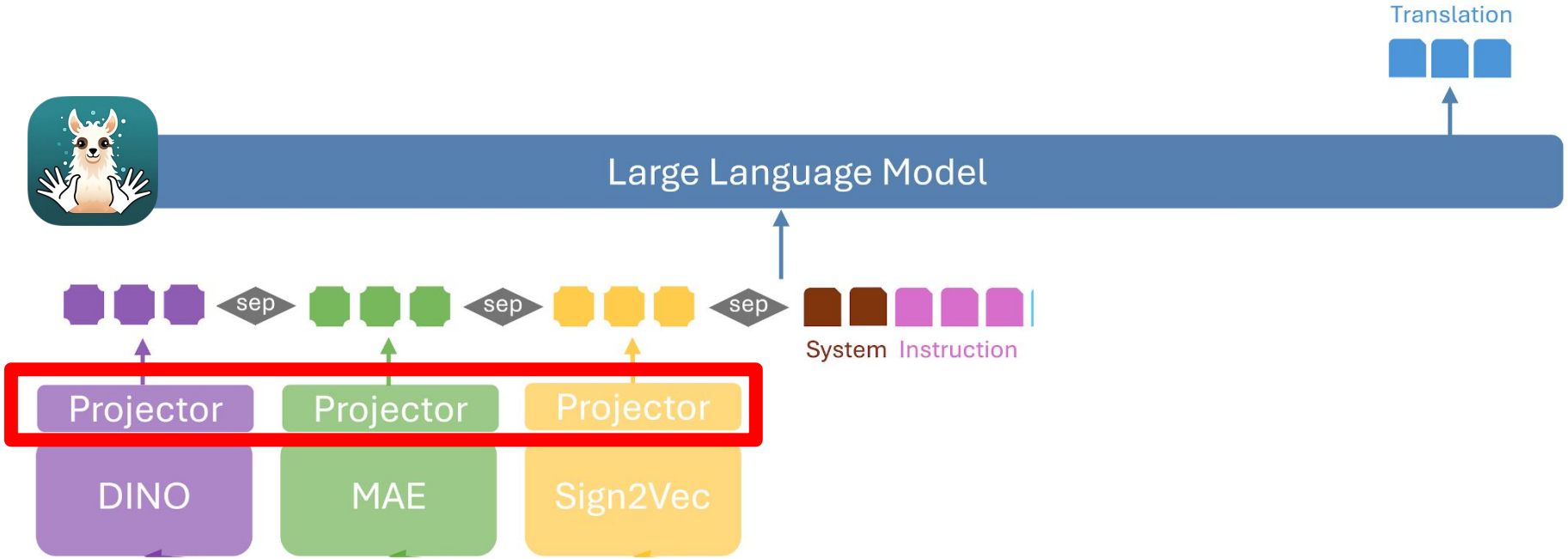


How: SignLaVa



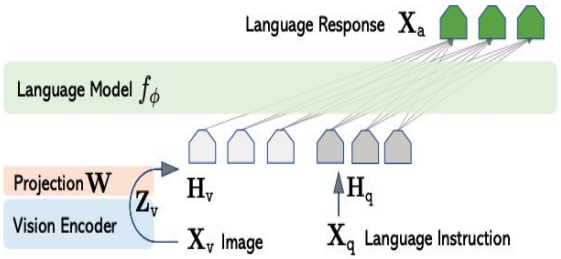
How to Train

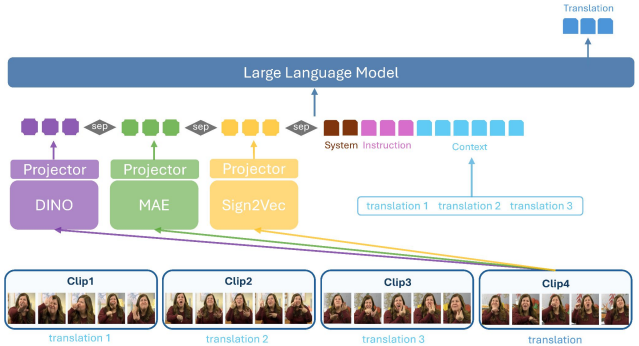
- Use pre-trained LLM (= Llama)
- Train sign lang. representation models separately
- Freeze LLM and repr. models. **Train MLP projector layers.**



Similarly for other modalities. E.g. LLaVA for images:

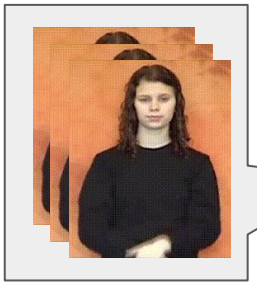
	LLaVA	SignLLaVA
Inputs	Image	Sign language video
Task	General purpose language and image understanding.	Sign language translation and understanding.
Visual Encoder	CLIP image encoder	Dino, MAE, Sign2Vec
Language prompt	w/o context	context: (1) preceding sentences; (2) beginning sentences.





Sign Language Representations

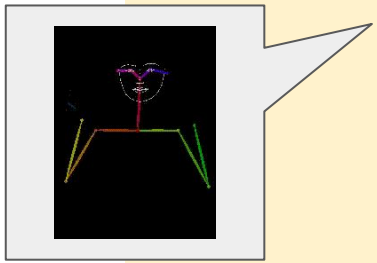
Sign Language Representations



video

What is your name?

text



poses

glosses

YOU NAME-[what]?

notation

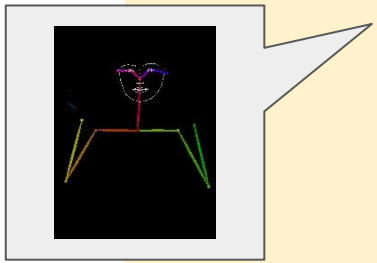
https://research.sign.mt/#sign-language-recognition-translation-and-production

Past: Gloss-based approaches

video

text

What is your name?



poses

glosses

YOU NAME-[what]?

notation

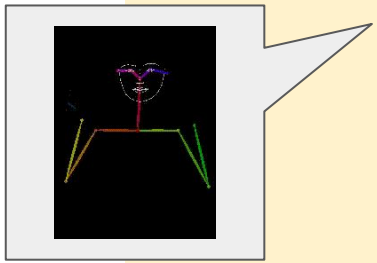
https://research.sign.mt/#sign-language-recognition-translation-and-production

Today: Gloss-free

What is your name?

video

text



poses

glosses

YOU NAME-[what]?

notation

https://research.sign.mt/#sign-language-recognition-translation-and-production

SignLLaVa (Ours)

What is your name?

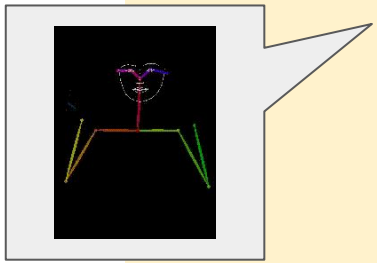
video

DINO, MAE

text

MediaPipe

Sign2Vec



poses

glosses

YOU NAME-[what]?

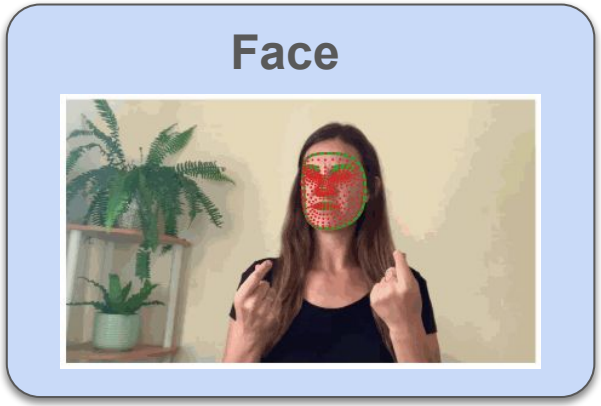
notation

https://research.sign.mt/#sign-language-recognition-translation-and-production

MediaPipe for Pose Estimation

- Small deep learning model (~5M params.)
- by Google
- Colab demo:

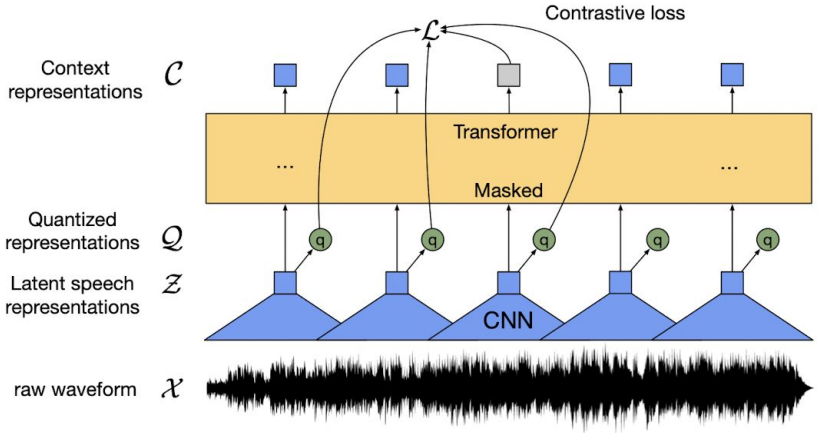
https://colab.research.google.com/drive/16EaS_132dMjlipot8vy28Ak0HjuYRZDV https://colab.research.google.com/drive/1aAV7leNH5QZJFni_pGMBQEQ8xvODM4nW#scrollTo=QDYoswfhL452
https://colab.research.google.com/drive/1aAV7leNH5QZJFni_pGMBQEQ8xvODM4nW?usp=sharing



https://ai.google.dev/edge/mediapipe/solutions/guide

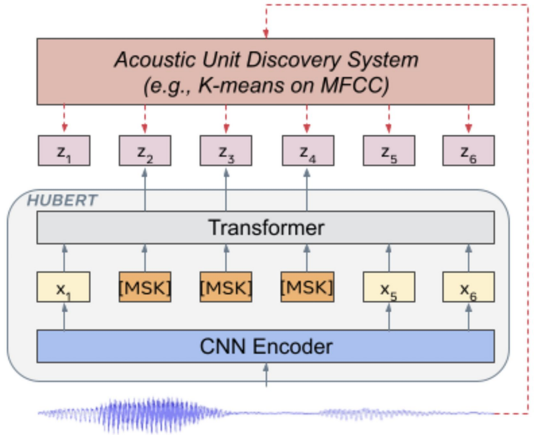
BERT-like Self-Supervised Representations

wav2vec2.0



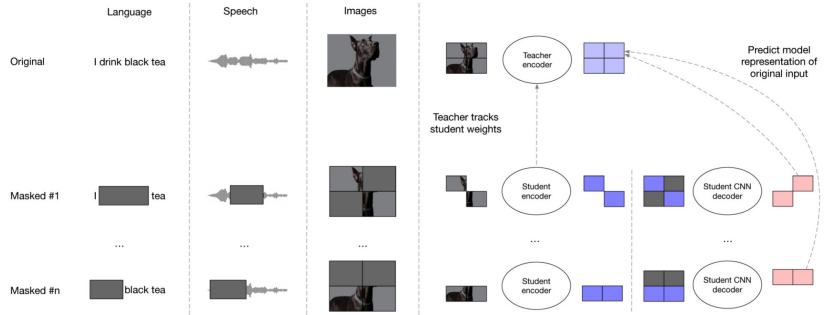
Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." Advances in neural information processing systems 33 (2020): 12449-12460.

HuBERT



Hsu, Wei-Ning, et al. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units." IEEE/ACM transactions on audio, speech, and language processing 29 (2021): 3451-3460.

data2vec2

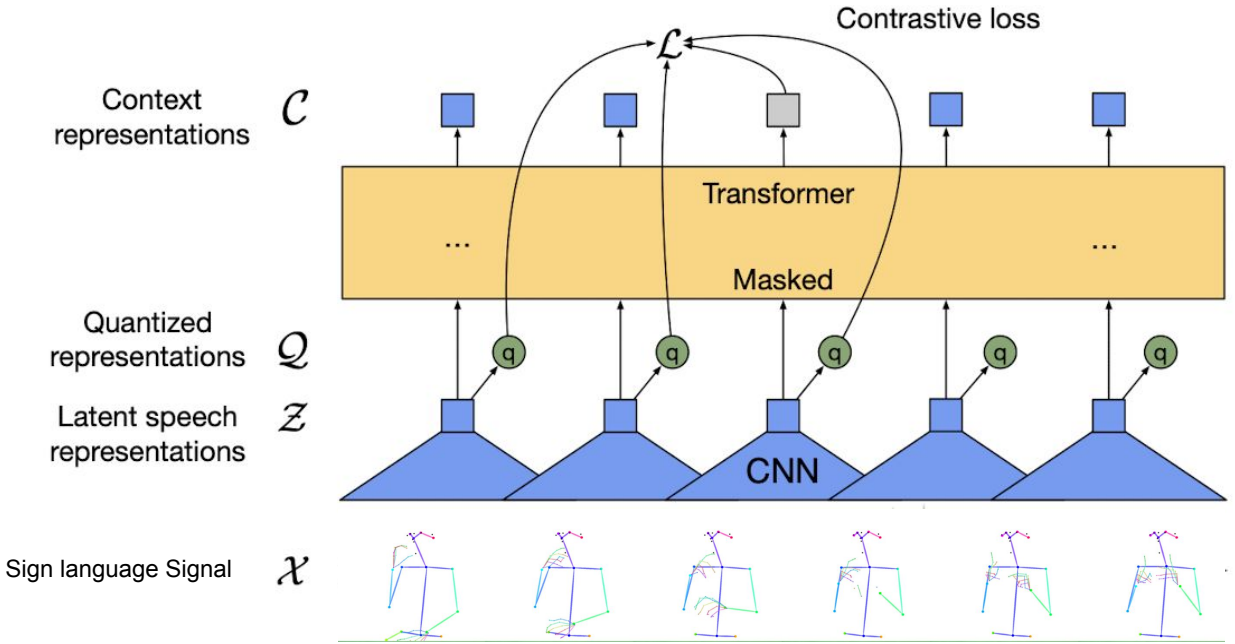


Baevski, Alexei, et al. "Efficient self-supervised learning with contextualized target representations for vision, speech and language." International Conference on Machine Learning. PMLR, 2023.

Sign2Vec

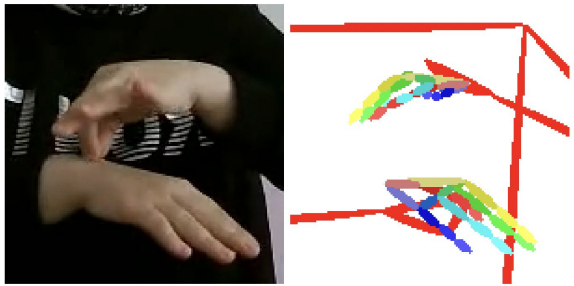
BERT-like: Self-supervised learning of sign language representations

GOAL: Given a pose sequence, produce a vector (sequence of contextual representations) specific to sign language content



The Problem of keypoints:

Often inaccurate or hallucinate.



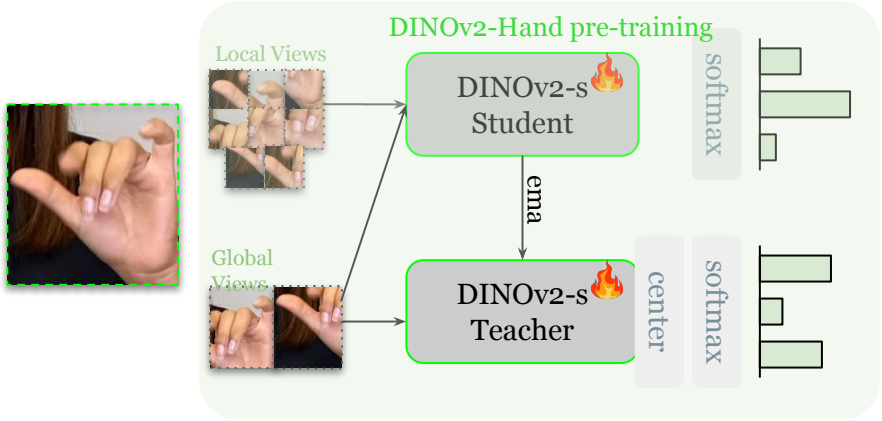
References:

Amit Moryossef, Ioannis Tsochantaridis, Joe Dinn, Necati Cihan Camgoz, Richard Bowden, Tao Jiang, Annette Rios, Mathias Muller, and Sarah Ebling. Evaluating the immediate applicability of pose estimation for sign language recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021.

Anna Kuznetsova and Vadim Kimmelman. Testing mediapipe holistic for linguistic analysis of nonmanual markers in sign languages. arXiv preprint arXiv:2403.10367, 2024.

DINO: Self-Distillation with No Labels

- Self-supervised model for image vector representation
- Student: local view, teacher: global view
- The student tries to match the teacher's output
- The teacher is updated slowly over time (through a moving average of the student).

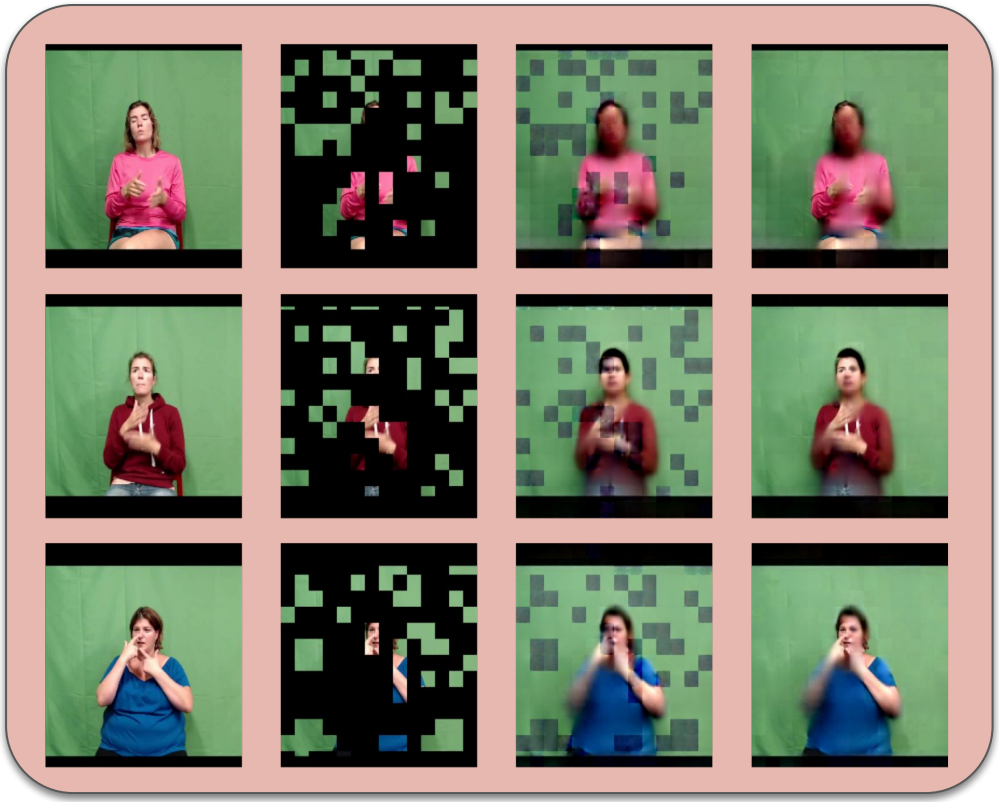


Masked Autoencoder

Self-supervised learning of sign representations

- Representation extractor

  for Sign Language
  - Global information
  - Full pose
  - Full texture
  - Representation of whole scene



The biggest problems in Sign Lang. Translation

- Sign. L. Generation
- Sparse representation
- Data acquisition
- Handshape modelling

Summary

Multimodal LLMs

- What is it, why, what is difficult
- How:
  - Text-to-Speech
  - Audio and Video
  - Sign Language