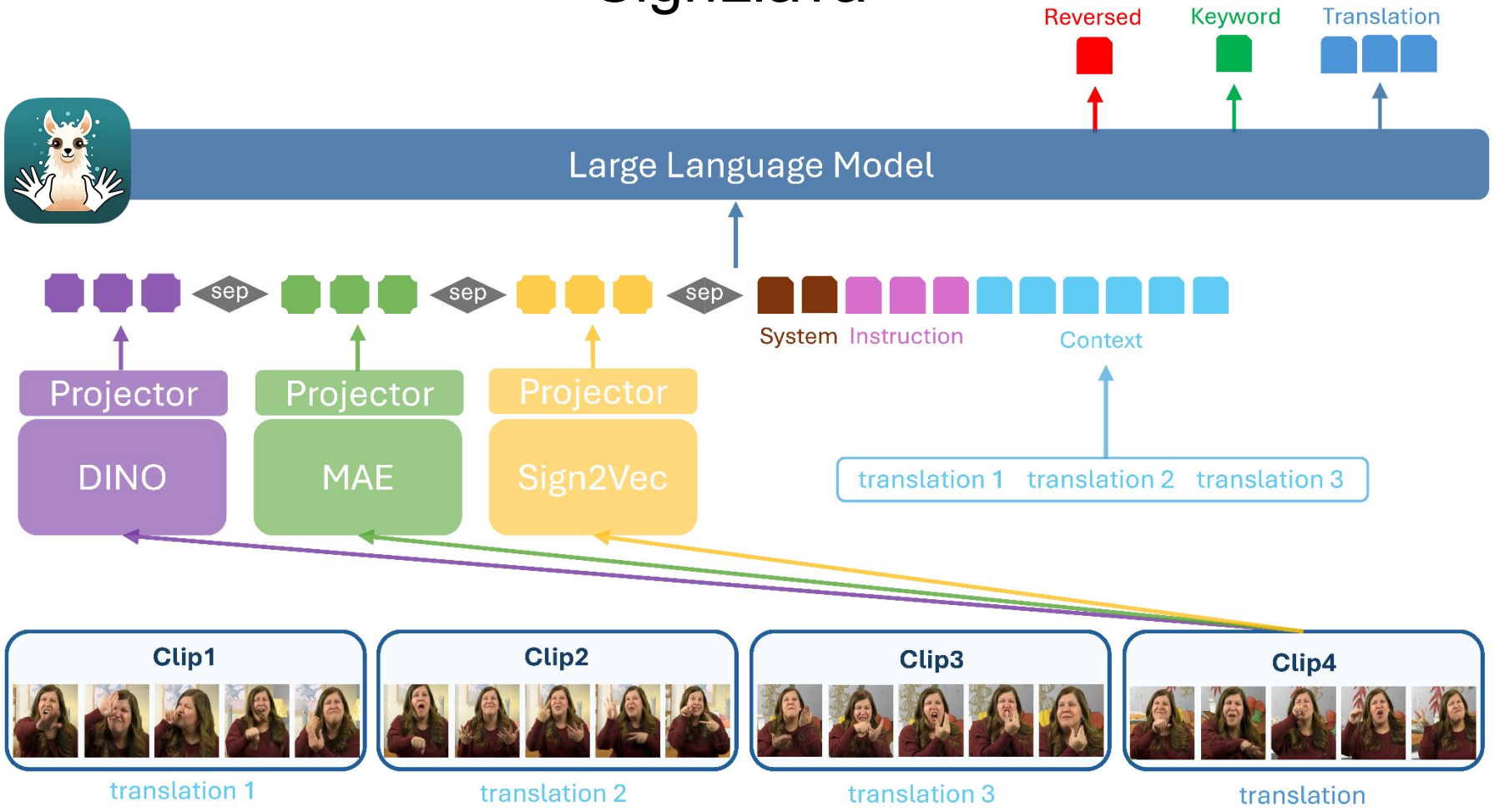
SignLlava summary for LUMI hackaton 2024

Tomas Zelezny, Jakub Straka, Matej Sieber

SignLlava


Reversed
Keyword
Translation
Large Language Model
sep
sep
sep
System
Instruction
Context
Projector
Projector
Projector
DINO
MAE
Sign2Vec
translation 1
translation 2
translation 3
Clip1
Clip2
Clip3
Clip4
translation 1
translation 2
translation 3
translation

Code Status

https://github.com/JSALT2024/Sign_LLaVA

- AMD support
- ZeRO2 code parallel, ZeRO3 experimental
- Running on single node with 8 GPUs
- We focussed on feature implementation so far
- Preliminary training experiments done (on LUMI)
- Wandb logging
- h5 data format

Hackaton goals

- Optimization of our model based on Llama3
- Multi-node training
- Using Llama-70b instead of Llama-8b
  - Correctly and efficiently use the model distribution
- Dealing with some existing problems
  - Correctly use flash-attention (1Torch was not compiled with flash attention)
  - Correctly use memory efficient attention (1Torch was not compiled with memory efficient attention)
  - SDPA attention implementation warnings for multi-node ROCM
- Moving our codes to latest Llama3.1 version