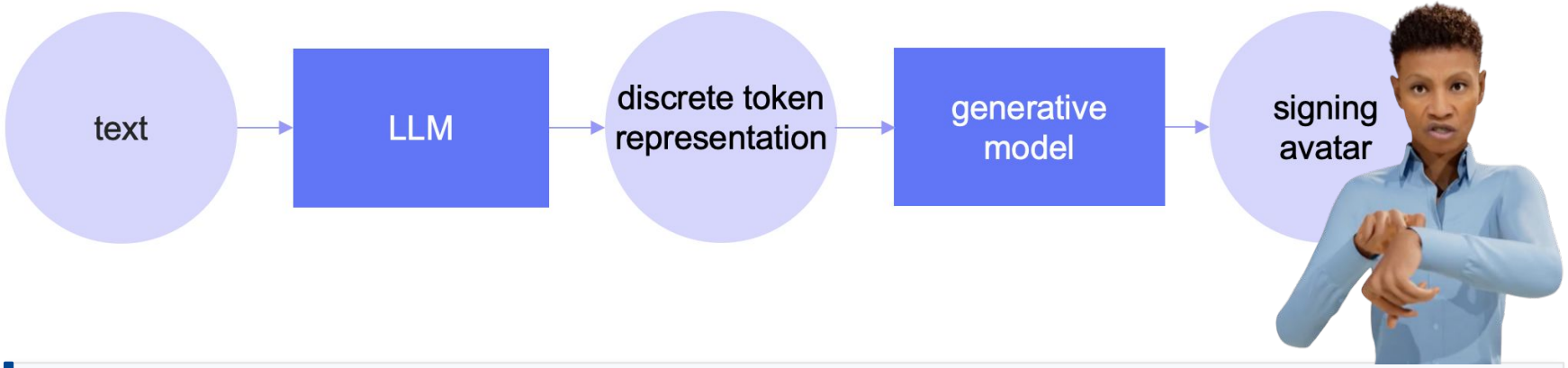
Comparison of Low Bitrate Quantizers for Encoding Swedish Sign Language

Anna Klezovich[1], Johanna Mesch[12], Gustav Eje Henter[1], Jonas Beskow[1]
[1]Speech, Music and Hearing, KTH Royal Institute of Technology, [2]Stockholm University

annkle@kth.se, beskow@kth.se

General Idea Behind Our Project



- Our goal

  → generative 3D signing avatar model for Swedish Sign Language (STS)

- The way to reach it

  → representation learning on lots of 2D and 3D data and motion generation on high quality 3D motion capture data

Why Representation Learning at All?

1 Better generative models

Learned discrete tokens capture domain-specific patterns (phonological units in SL, dance moves, speech phonemes).

2 Efficient training

We can compress the data at lower bitrate → smaller codebooks → faster training.

3 Enables LLM-based translation

Applying LLMs to sign language translation requires sign data to be at a text-comparable bitrate.

Our contribution

We benchmark 2 quantizers, ***k*-means** baseline and a residual vector-quantization model **(RQ-VAE)**, on Swedish Sign Language (STS) motion capture **at various bitrates** to find the most optimal compression-quality tradeoff.

What Data Do We Use? Our Own! 🤩

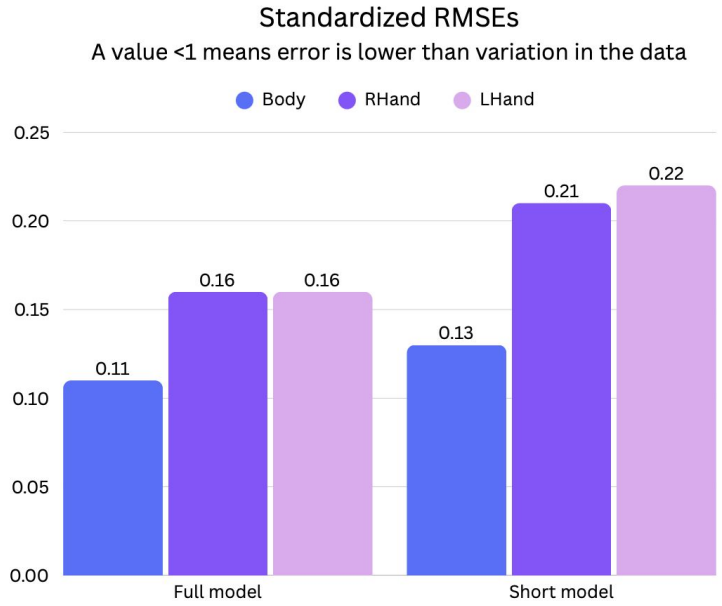
- **4 hours** of high-quality 3D motion capture
- **3.5 hrs** simplified news dataset in Swedish (8sidor) & **30 min** STS dictionary signs & sentences
- Sentence by sentence translation **from text** in Swedish
- **1 CODA signer** in consultation with a deaf L1 signer
- Recorded **face, body, and fingers** motion, but the **face data was not used** for this study
- Body + hands recorded at **120 fps**

STS Mocap v1

See also our LREC 2026 paper: *Klezovich, A., Mesch, J., Henter, G.E., and Beskow, J. (2026). How much Data is Enough Data? A New Motion Capture Corpus for Probabilistic Sign Language Generation.*

4 Hours Is Not Much, Why Do You Think Generative Modelling Will Work Here?

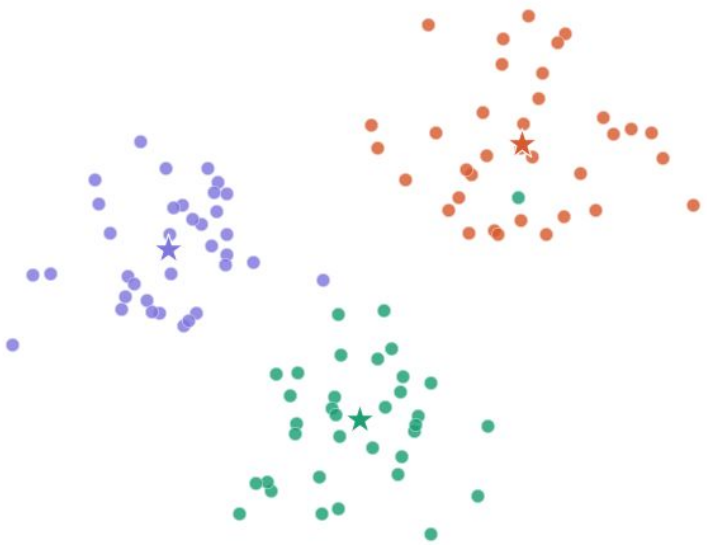
- In LREC main conference submission we have already experimented with 2D-to-3D task
- We compared training on 30 minutes of data with training on the whole dataset
- **We showed that 4 hours is enough data for a generative modeling task of generating 3D motion from 2D**



See also our LREC 2026 paper: *Klezovich, A., Mesch, J., Henter, G.E., and Beskow, J. (2026). How much Data is Enough Data? A New Motion Capture Corpus for Probabilistic Sign Language Generation.*

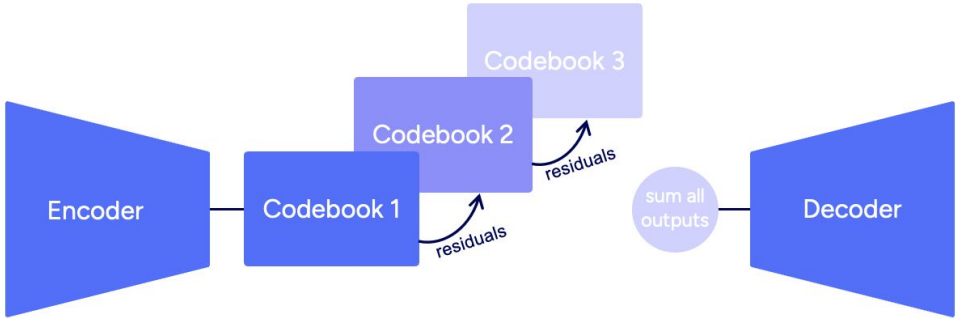
Quantization Methods We Compare

k-means



✗ Not a network, doesn't *learn* representations
✓ Simple, no training needed

RQ-VAE (Residual VQ)



✗ Hard to train
✓ Learned encoder & decoder provide more refined representations

How to Calculate Information?

$$\text{bitrate} = \log_2(k) \times \text{frames per second}$$

How many bits required to record k clusters/vectors?

Turn it into bits per second

bitrate-distortion tradeoff

Calculates the minimum amount of data required to encode a source to a specific level of fidelity.
In our case we measure it with a reconstruction error.

Research Question

So when we compare RQ-VAEs with the simpler *k*-means clustering at various bitrates, we are basically asking:

What affects the reconstruction error of sign language data more

- **amount of information we encode, aka. bitrate**

  or

- **complexity/quality of the quantizer?**

Experiments

bits/frame	*k*-means	RQ-VAE
6	k = 64	D = 8, Q = 2
9	k = 512	D = 8, Q = 3
10	k = 1024	D = 32, Q = 2
12	k = 4096	D = 8, Q = 4

That's a very low bitrate!

Formula for *k*-means:
$bitrate = log_2(k) \times fps$

Formula for RQ-VAE:
$bitrate = log_2(D) \times Q \times fps$

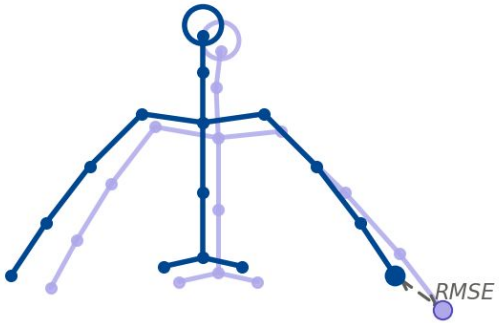
All experiments trained/fitted on joint angles (exp. maps) at 30 fps, body & hands modelled separately.

Evaluation Metrics

Full-pose RMSE

Global

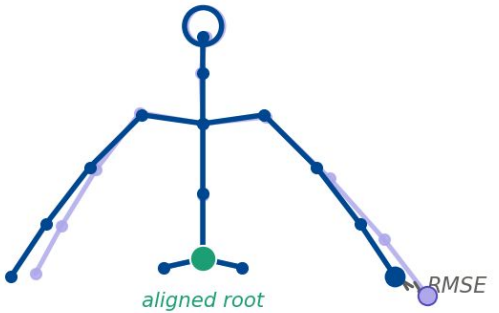
RMSE on raw 3D joint positions, averaged distance (cm) between original and reconstructed joints across the whole body.



Kabsch-aligned RMSE

Detailed

First, we align rotation and translation of root, e.g. hips joint, and then calculate RMSE. Isolates shape error from global translation/rotation drift. Separately for body and hands.

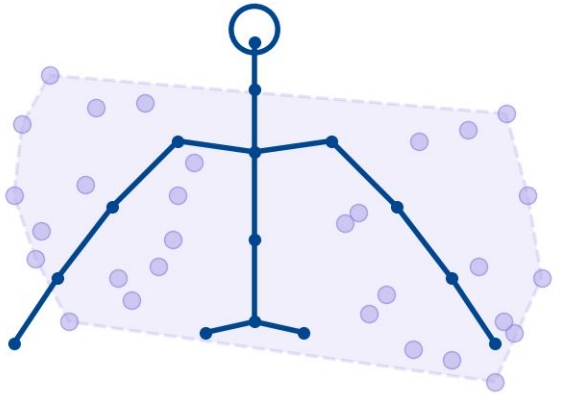


Evaluation Metrics

Convex hull volume

Expressivity

Wrist 3D point-cloud convex hull as % of original. Larger = more expressive signing space. Computed over full sequence with scipy.

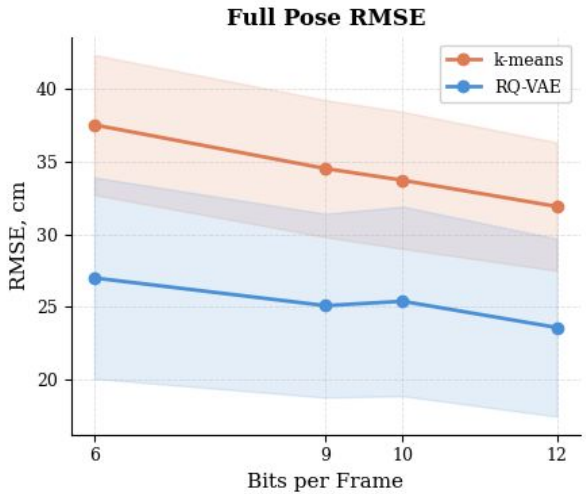


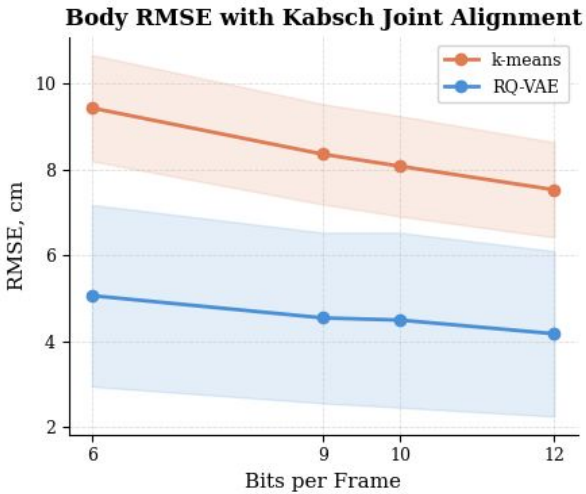
Motion smoothness

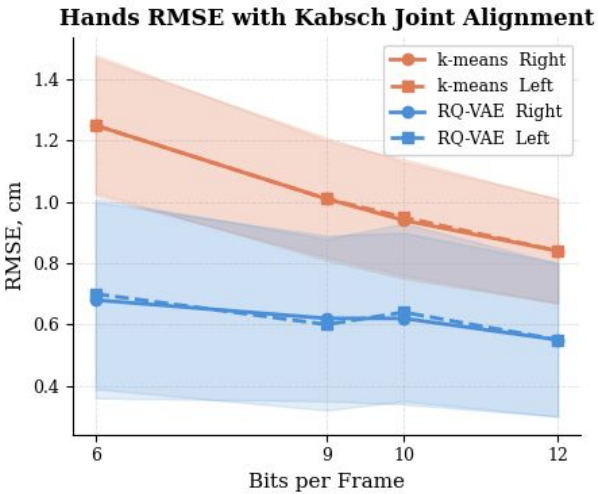
Smoothness

Measures if a downstream generative model learned realistic motion dynamics from quantized tokens. Calculated as average motion jerk.

Results: RMSE vs. Bitrate





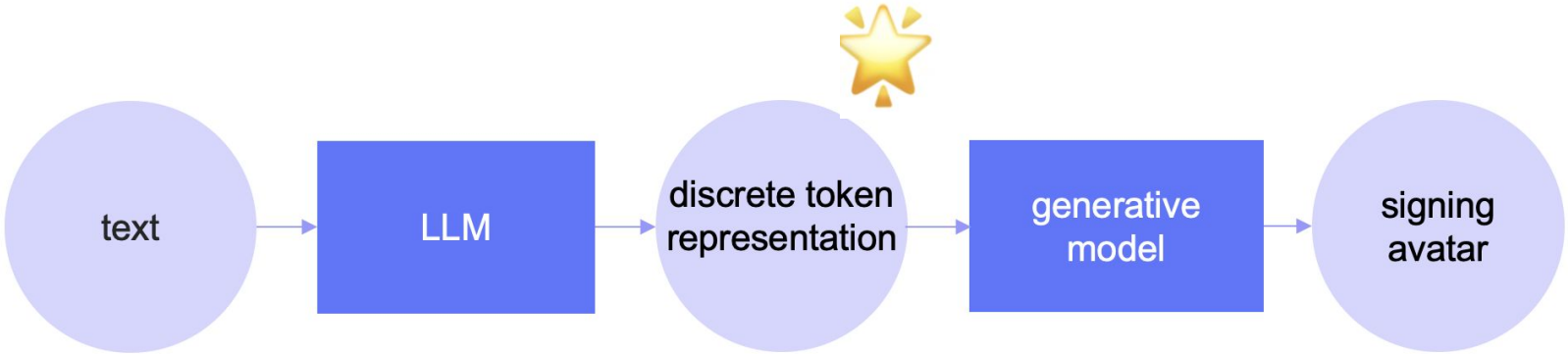


RQ-VAE slightly better at all bitrates for full pose. Advantage grows at low bitrates.

RQ-VAE body RMSE **approximately 1.8x lower** than *k*-means at all bitrates

Results overlapped more at higher bitrates, while RQ-VAE was still on average better than *k*-means.

Intermediate Discrete Representations



Now we built models that produce discrete token representations.
Let's look at them then!

Intermediate Discrete Representations

Original mocap	*k*-means, 10 bits/frame	RQ-VAE, 10 bits/frame

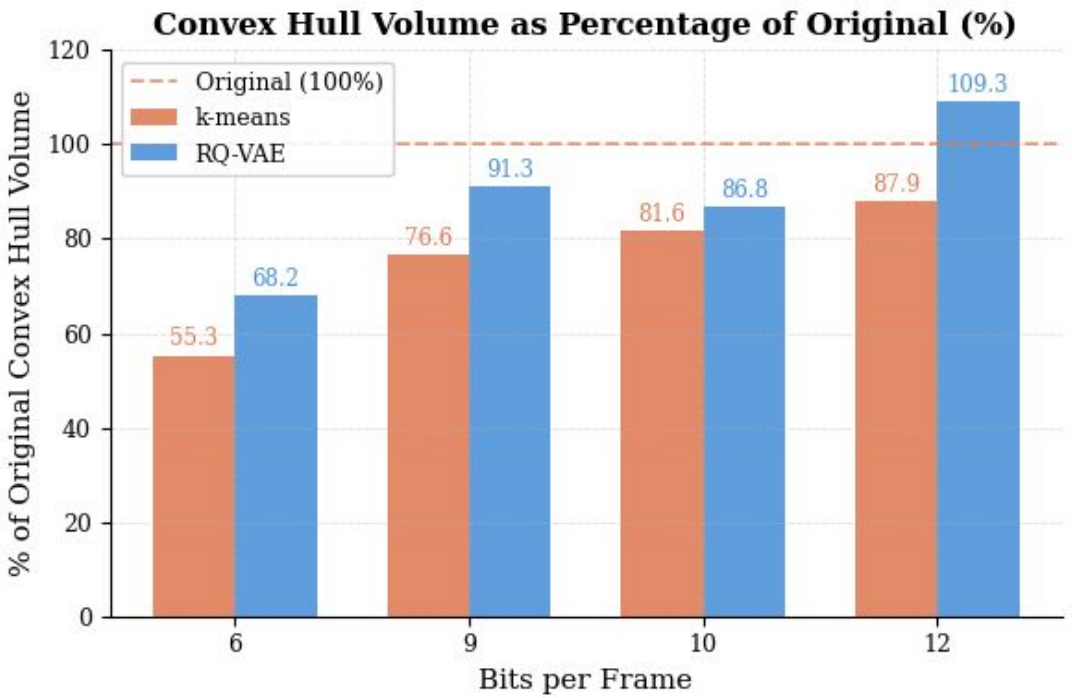
k-means: k = 1024, RQ-VAE: D = 32, Q = 2; Unreal Engine 5.4.4 & Metahuman avatars.

Intermediate Discrete Representations

Original mocap	*k*-means, 10 bits/frame	RQ-VAE, 10 bits/frame

k-means: k = 1024, RQ-VAE: D = 32, Q = 2; Unreal Engine 5.4.4 & Metahuman avatars.

Results: Signing Space Expressivity



** Ground truth average: 0.81 m³*

RQ-VAE uses more signing space at all bitrates.

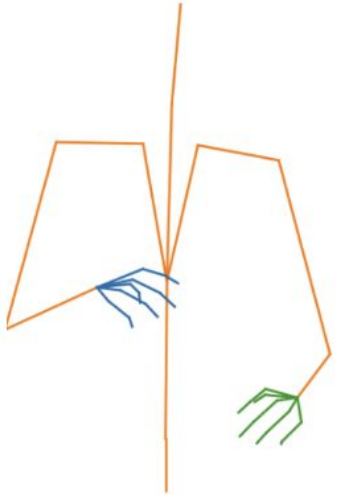
k-means growth is monotonic in bitrate.

RQ-VAE growth is non-monotonic, probably sensitive to hyperparameters.

At 12 bits per frame RQ-VAE reaches 109%, over-expansion.

Downstream Generative Model – Proof of Concept

Condition on cluster centroid poses from *k*-means

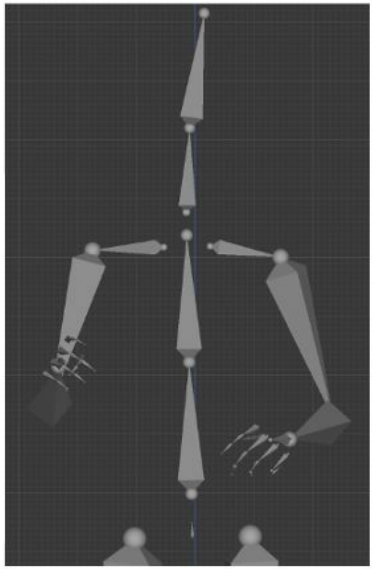


→

Match-TTSG flow-matching model (Mehta et al. 2024)

→

Output: reconstructed 3D motion



Generative Model Output

Downstream model produces smooth motion even on *k*-means clusters at low bitrate.

Original mocap

Synthesized with the model train on 10 bits/frame input

k = 1024, Unreal Engine 5.4.4 & Metahuman avatars

Generative Model Output

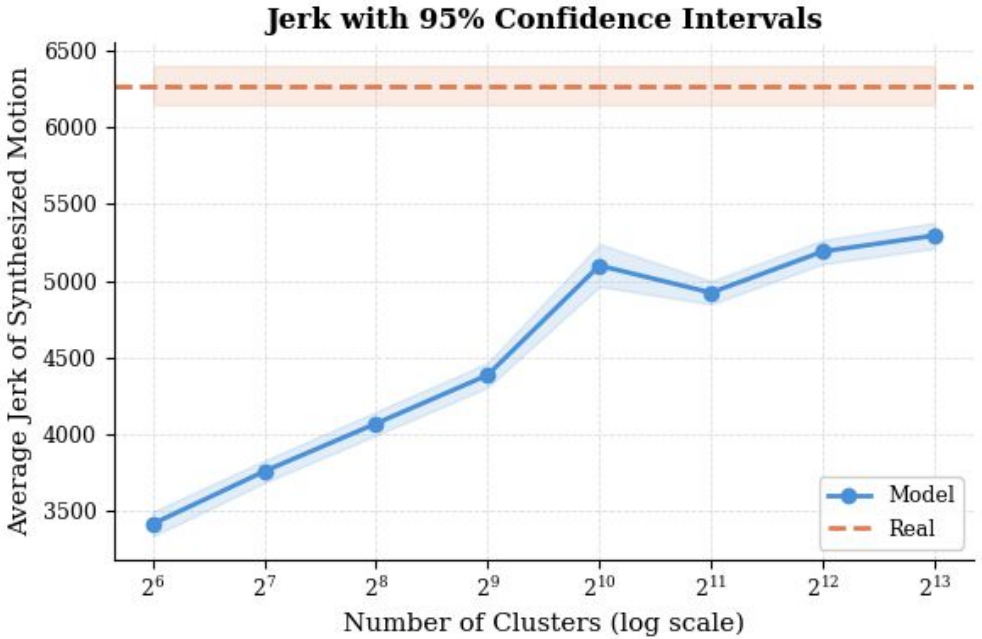
Downstream model produces smooth motion even on *k*-means clusters at low bitrate.

Original mocap

Synthesized with the model train on 10 bits/frame input

k = 1024, Unreal Engine 5.4.4 & Metahuman avatars

Generation Quality: Jerk vs. *k* of Cluster



Oversmoothness of *k*-means

Synthesis motion smoothness gets more “realistic” the closer the number of *k* clusters is to the size of the original dataset.

Findings

01

RQ-VAE outperforms *k*-means in RMSE

1.8× lower body error at all tested bitrates (6-12 bits per frame) with no overlap of standard errors.

02

Better signing space preservation in RQ-VAE

RQ-VAE uses a larger % of original signing space at all bitrates than *k*-means.

03

Discrete tokens can drive generation

A flow-matching model conditioned on *k*-means centroids at low bitrate produced smooth and plausible, but imperfect sign sequences.

04

6-12 bits per frame are likely insufficient for phonological accuracy

Upon visual inspection of synthesized motion sequences, one can notice a lot of inaccuracies due to the low bitrate, which is expected.

Future work: finetune an LLM on discrete sign tokens for text → 3D sign language translation

Even at a very veeery low bitrate of discrete tokens a generative model can reconstruct a plausible 3D sign language sequence!

Access our dataset on Huggingface 🤗

Thank you for your attention!

annkle@kth.se, beskow@kth.se

Selected references

- Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. 2023. Listen, Denoise, Action! Audio-Driven Motion Synthesis with Diffusion Models. ACM Trans. Graph., 42(4).
- Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2023. Flow matching for generative modeling.
- Shivam Mehta, Ruibo Tu, Simon Alexanderson, Jonas Beskow, Eva Szekely, and Gustav Henter. 2024. Unified speech and gesture synthesis using flow matching. In 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2024), pages 8220–8224.