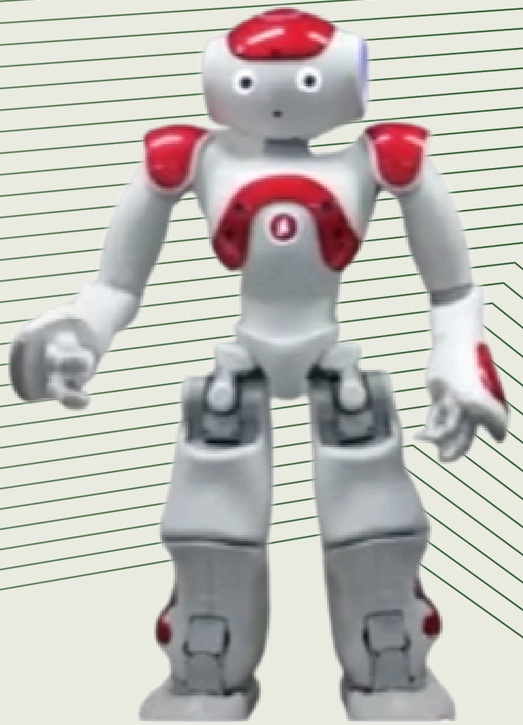
CO-SPEECH GESTURE GENERATION

Robots Learn Social Skills: End-to-End Learning of Co-Speech Gesture Generation



WEI, Yujia, Mateus
Supervisor: Decky
1 December, 2025

Research Background & Motivation

- Importance of Co-Speech Gestures
  - **Gestures**: core component of communication
  - **Function**: emphasize prosody,
    - convey intentions
    - illustrate concepts vividly
    - support language comprehension
- Current Limitations
  - Rule-based Methods
  - Early Learning-based Methods
- Our Research Goal
  - Propose an end-to-end generative model
  - Enable robots to acquire human-like non-verbal social skills

Figure 1. Speech of TED (Source: Wikipedia)

Baseline: End-to-End Neural Network Architecture

- Sequence-to-Sequence (Seq2Seq) Model
  - Adopts an Encoder-Decoder structure inspired by Neural Machine Translation.
  - Directly maps speech text words to a sequence of upper-body poses.
- Encoder: Speech Text Understanding
  - Input: Word sequences converted to vectors using pre-trained GloVe embeddings.
  - Structure: Bi-directional GRU (Gated Recurrent Unit) to capture full speech context.
- Decoder: Motion Generation
  - Structure: Recurrent Neural Network (GRU) with linear projection layers
  - Soft Attention Mechanism: Enables the model to focus on specific words while generating corresponding gestures
  - Continuity: Inputs n previous poses to generate smooth, continuous subsequent motions

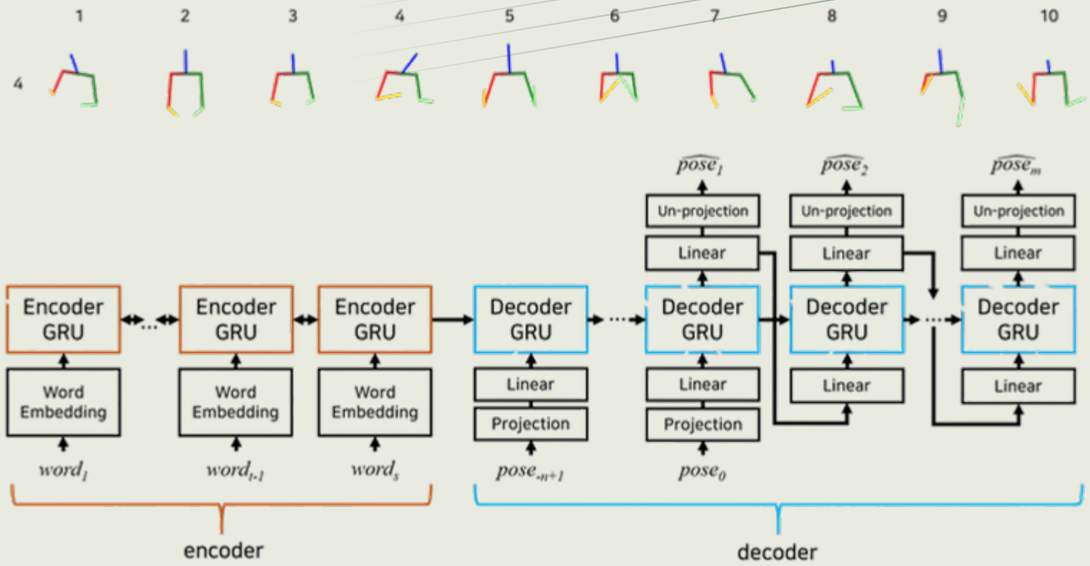
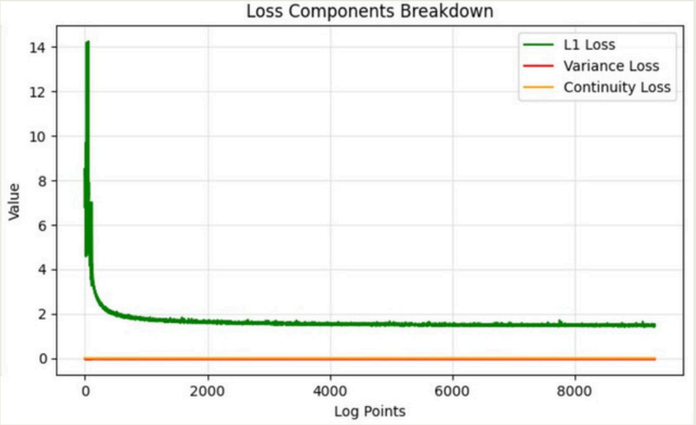


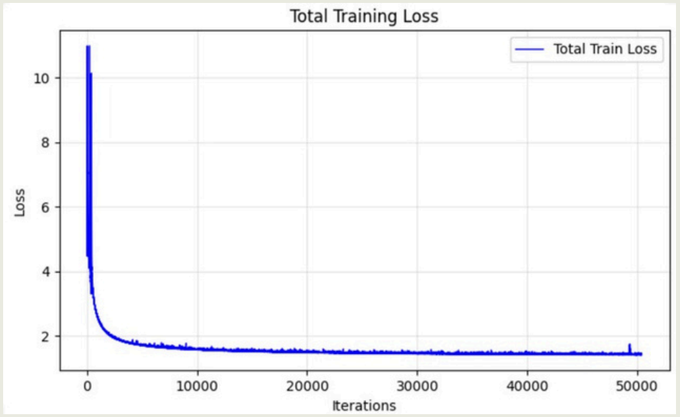
Figure 2. Network Architecture [1]

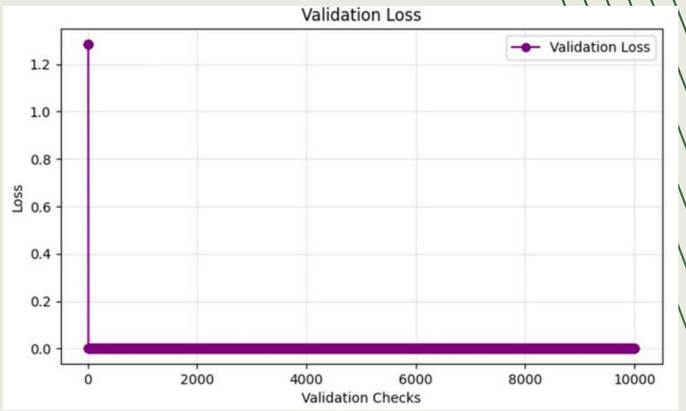
Analysis of Loss Components

The model optimizes a multi-objective loss function: $\mathcal{L} = \mathcal{L}_{rec} + \alpha\mathcal{L}_{cont} + \beta\mathcal{L}_{var}$

- L1 Reconstruction Loss (**Accuracy**)：Measures the distance between generated poses and ground truth.
- Continuity Loss (**Smoothness**)：ensure the robot's motion is fluid and natural
- Variance Loss (**Dynamism**)：Crucial for preventing the "Mean Pose" problem.





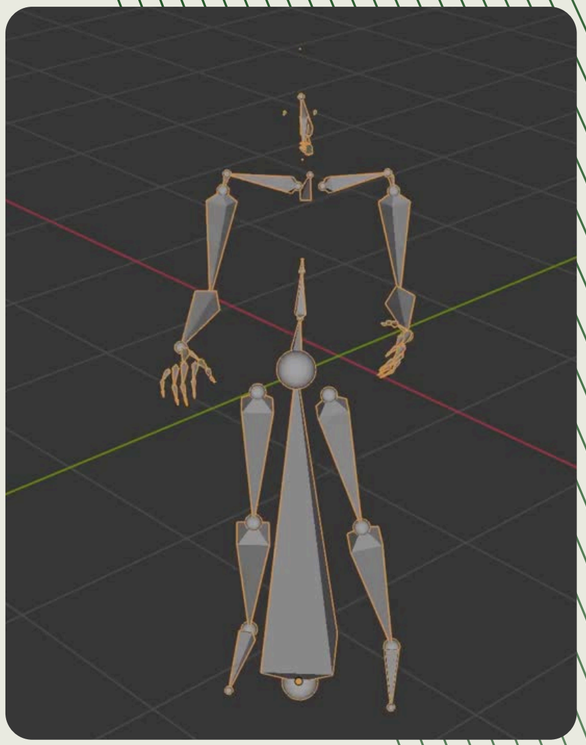


Dataset and evaluations metrics

Yeah. I agree.Yeah. I would be feeling horrible.

We use Genea [2] dataset that consists of :

- Statistics:
  - Training (371 samples):
    - Main Agent : sound and movement.
    - Interlocutor : sound and movement.
  - Validations (41 samples):
    - Main agent : sound and movement.
    - Interlocuter : sound and movement.
  - Test (70 samples):
    - Main agent : sound and no movement.
    - Interlocuter : sound and movement (used as proxy test).



Quantiative metrics:

MSE (Mean Squared Error): Quantifies the average squared difference between generated and ground truth motion per frame, measuring absolute reconstruction accuracy.

AVE (Average Variance Error): Measures the deviation in statistical variance between generated and real motion, assessing motion dynamics and liveliness.

Example of result

Rhythm Synchronization:

The model successfully produces rhythmic hand movements that coincide with speech emphasis.

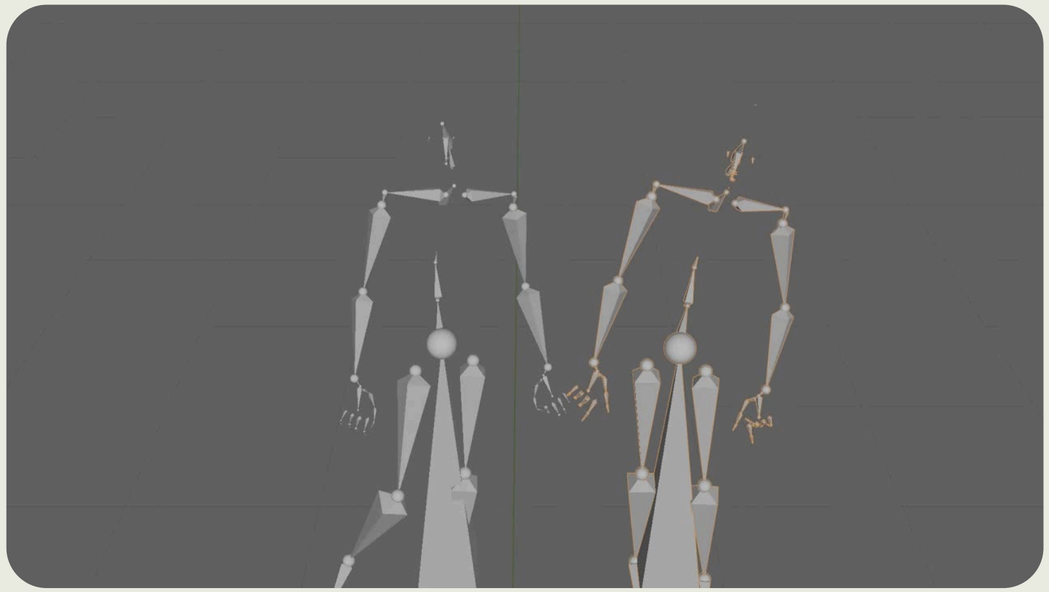
NUM	MODEL	Train		Val		Test(interloctr)	
		MSE	AVE	MSE	AVE	MSE	AVE
1	Baseline (Sequence - Sequence)	80.28	183.55	63.25	157.78	187.52	631.31

Semantic Consistency:

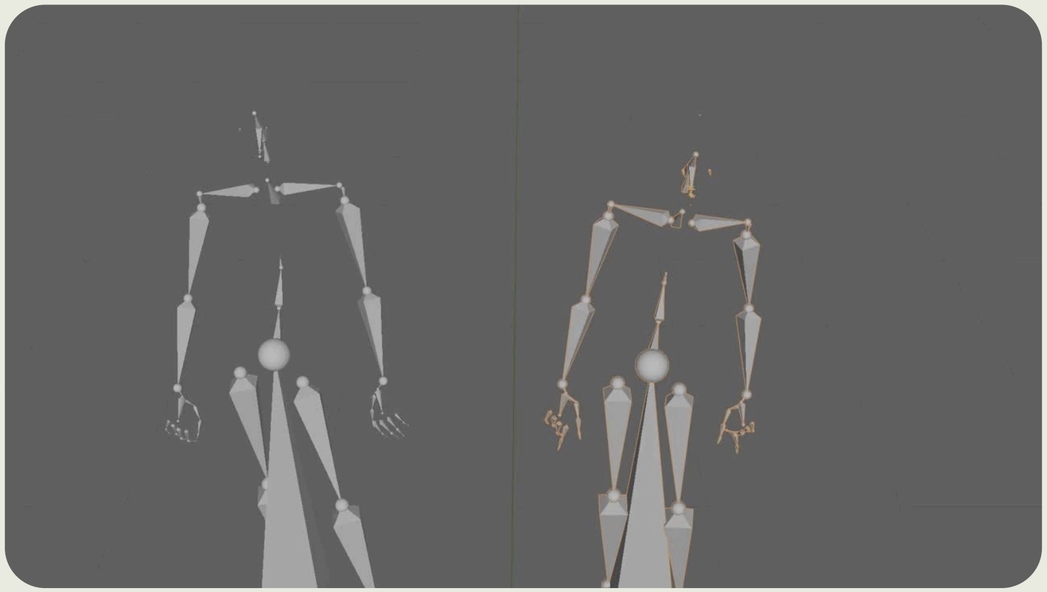
Observations show appropriate gesture types (e.g., opening arms) correlated with the text content.

The generated motions are on the **right side**

Okay. Okay. I feel bad now# ### you are...



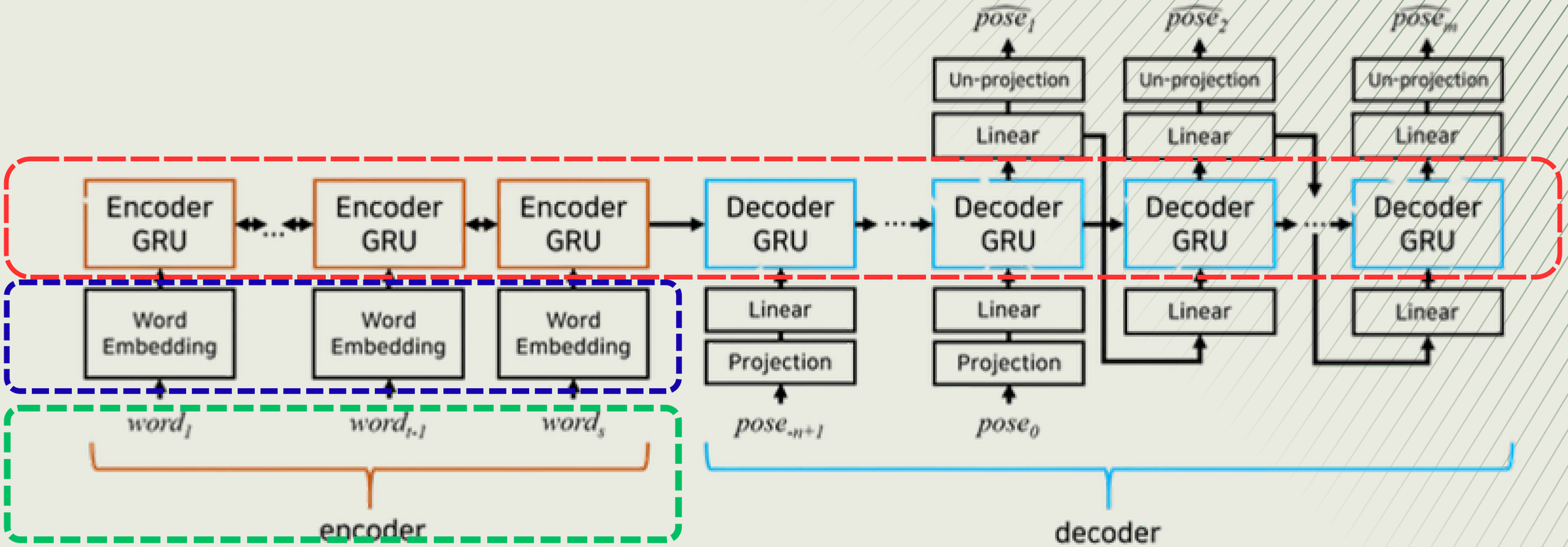
Um, that is kind of funny. Yeah. Um...



Proposed advancements

Three advancements:

- Change the **methods** → from seq-seq (current) to **Difusion** based model
- Change the **input** → word embedding to **BERT** based (instead of GLOVE) embedding
- Change the **input** → to use **sound** (in **spectogram**) in place of text input



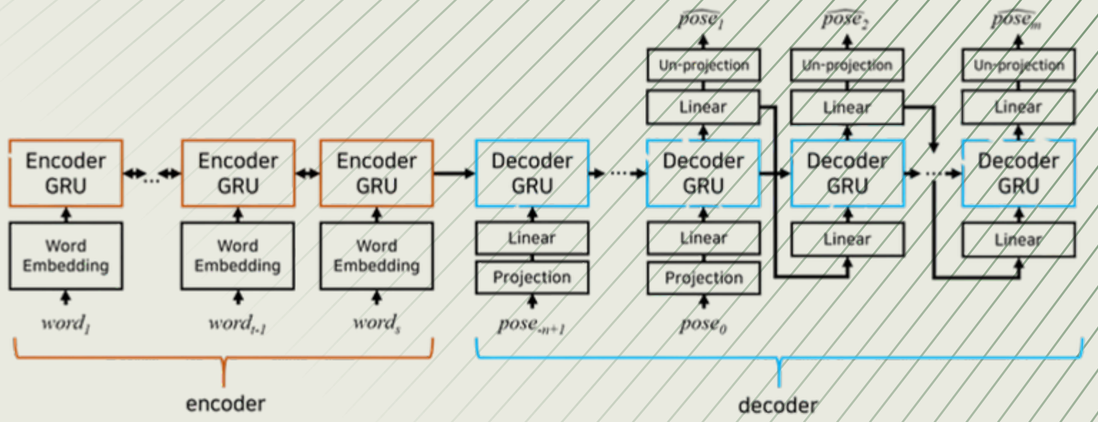
Advancement: integration of diffusion method

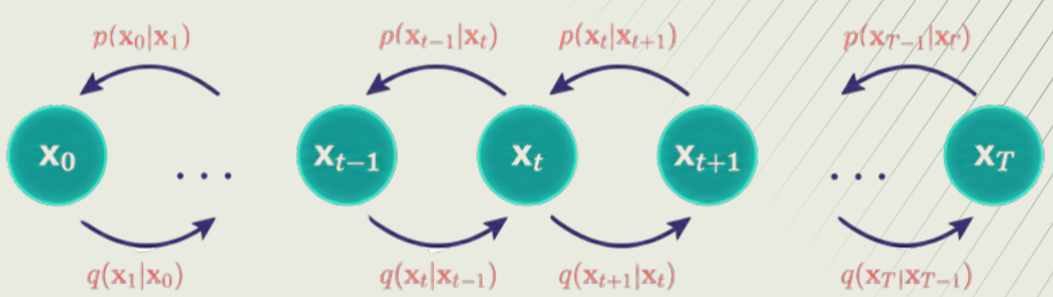
Advancement: Method to Diffusion

Motivation:
Beyond Deterministic Generation
The current GRU-based model is deterministic. For a fixed input text, it always generates the same gesture sequence.

The Goal: Human gestures are stochastic. We aim to model this probabilistic nature to generate diverse and creative motions for the same speech.

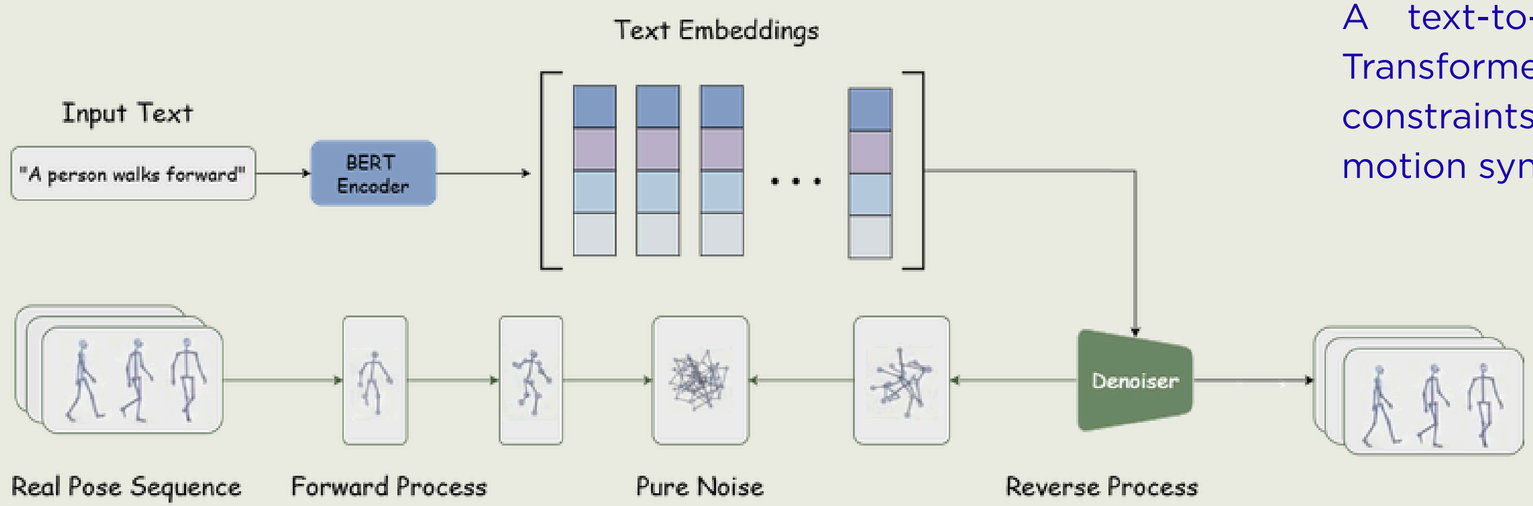
Preliminary task: Text generation using diffusion method.

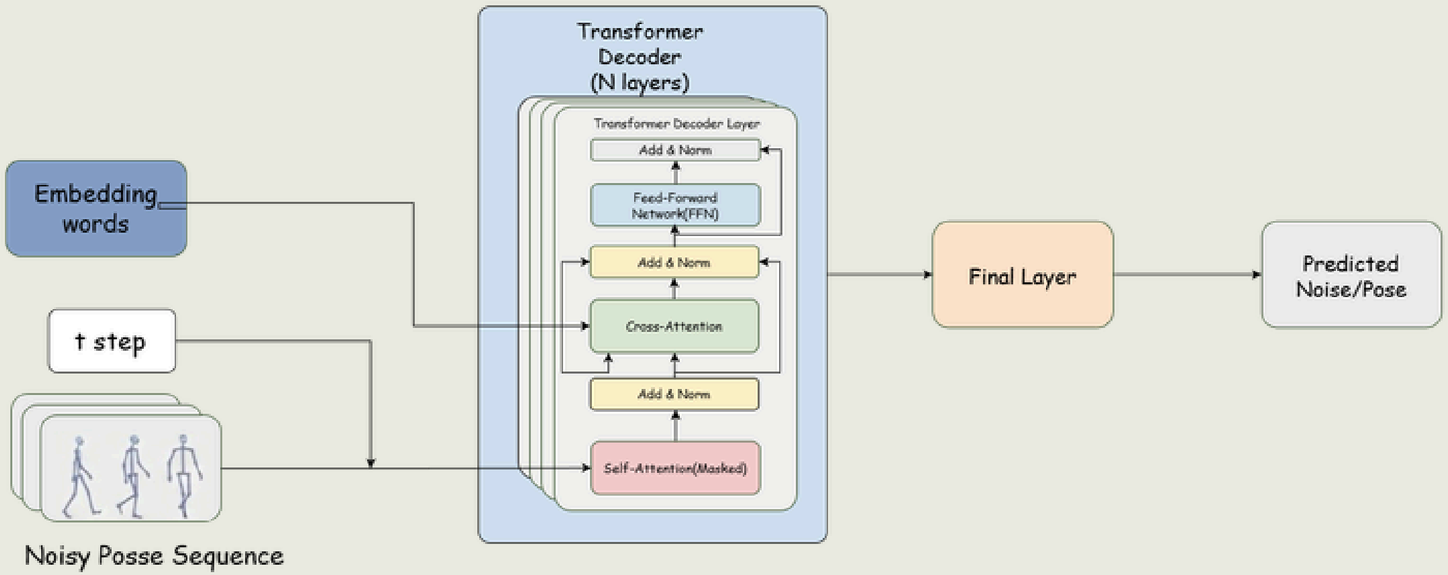




NEW METHOD: Diffusion + Transformer Decoder

A text-to-gesture diffusion framework using a Transformer Decoder, optimized with geometric constraints (SVD) and smoothing for robust 3D motion synthesis

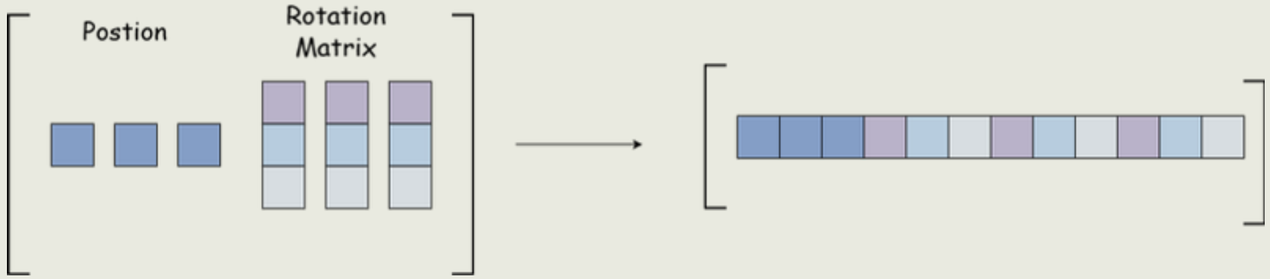




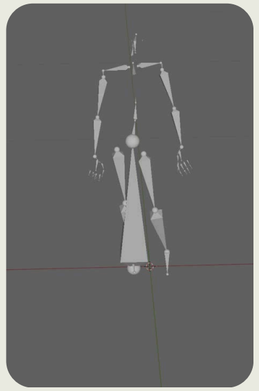
Data Preprocessing

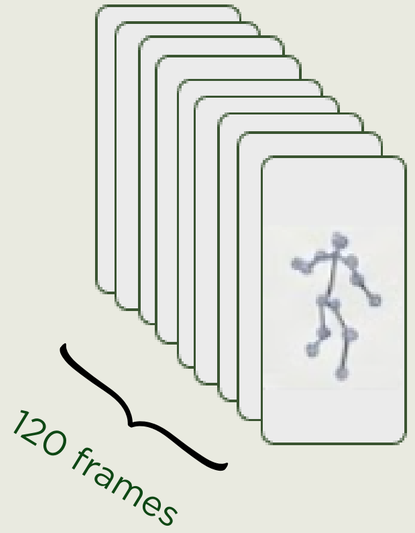
The preprocessing stage converts raw BVH data into normalized 12D feature vectors (incorporating rotation matrices) and slices them into fixed 120-frame sequences stored in LMDB for efficient training.

TSV file	Start time	End time	Word
	...	...	...



BVH file	Define struct	Postion	Rotation
	Root ...	XYZ	ZXY



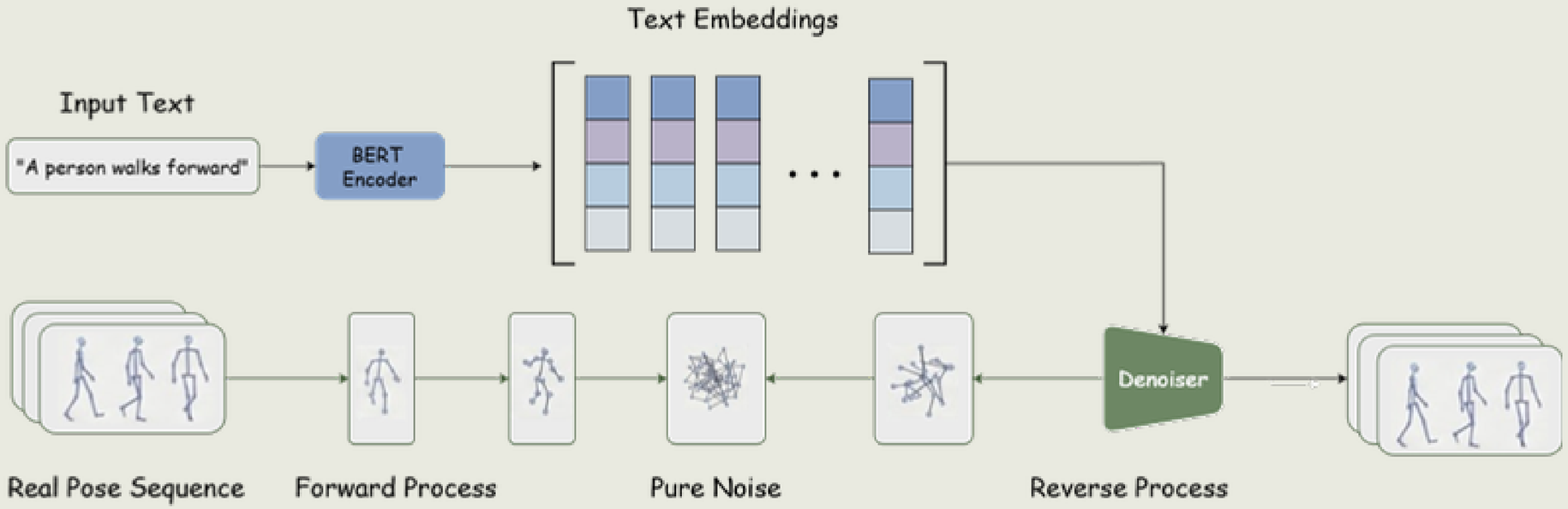


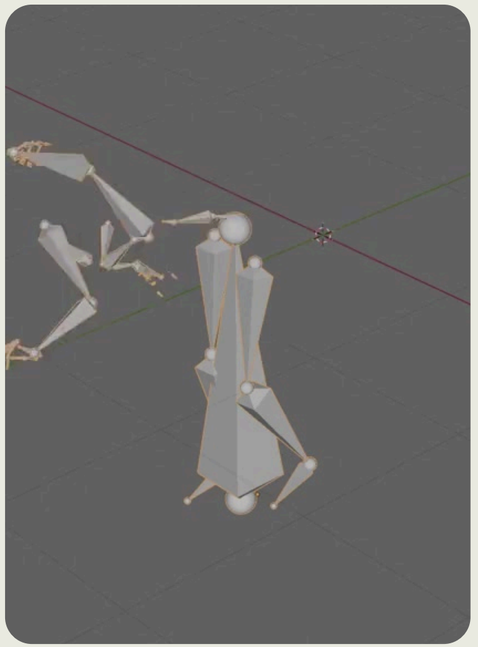
Normalization

[Batch_Size, 120, 216]

NEW METHOD: Diffusion

The diffusion module executes a text-conditioned reverse denoising process, utilizing the Transformer backbone to iteratively reconstruct coherent motion sequences from Gaussian noise over 1000 timesteps.



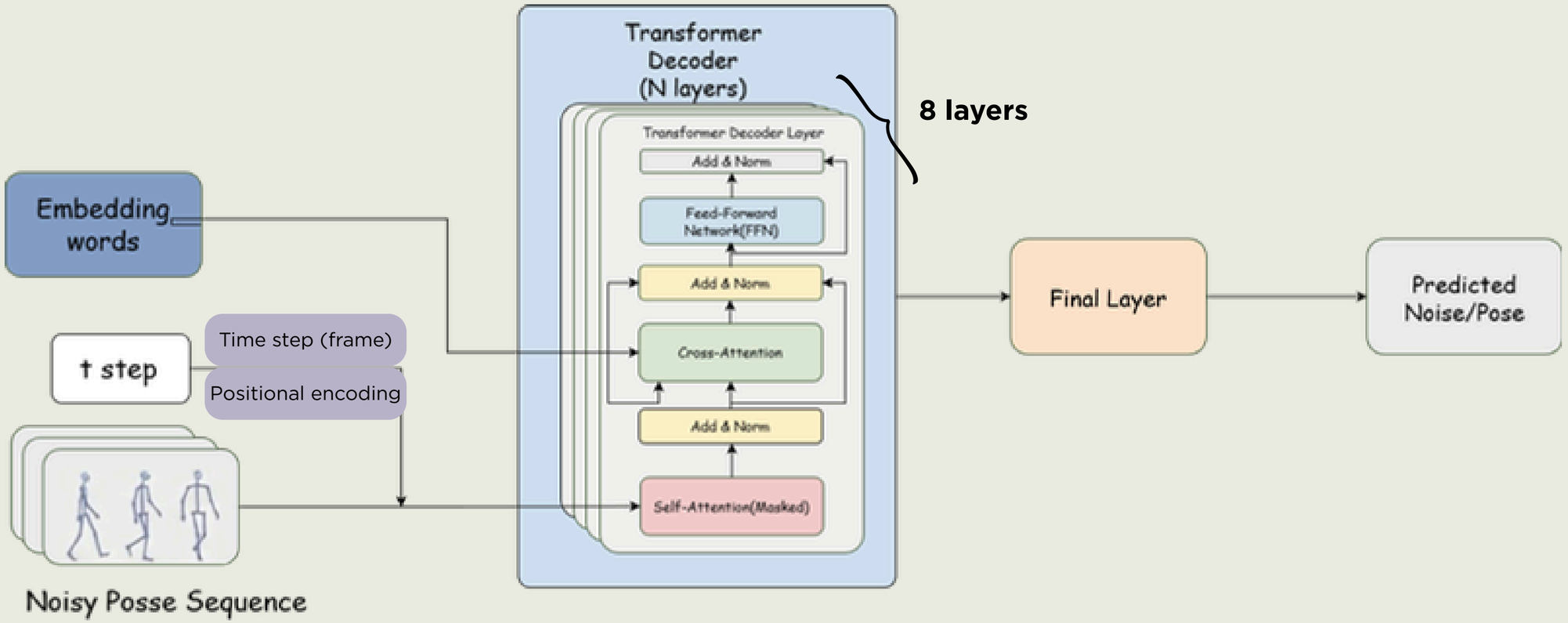


NEW METHOD: Transformer Decoder

Self-Attention

Cross-Attention

Feed-Forward Network

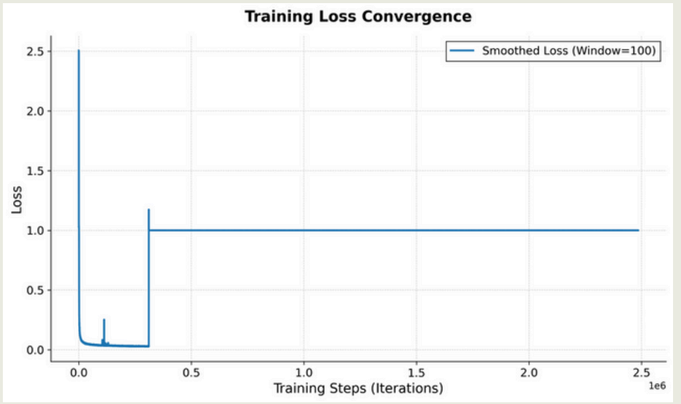


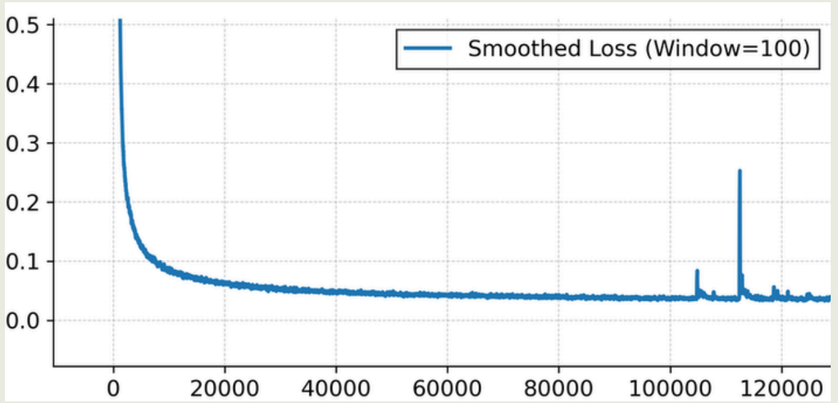
Outcome:

Training converged initially but collapsed later due to instability; evaluation is based on the best checkpoint before collapse

High MSE indicates the model generates diverse motions rather than overfitting to the mean pose.

Low AVE confirms the model captures the true dynamics and liveliness of human motion.

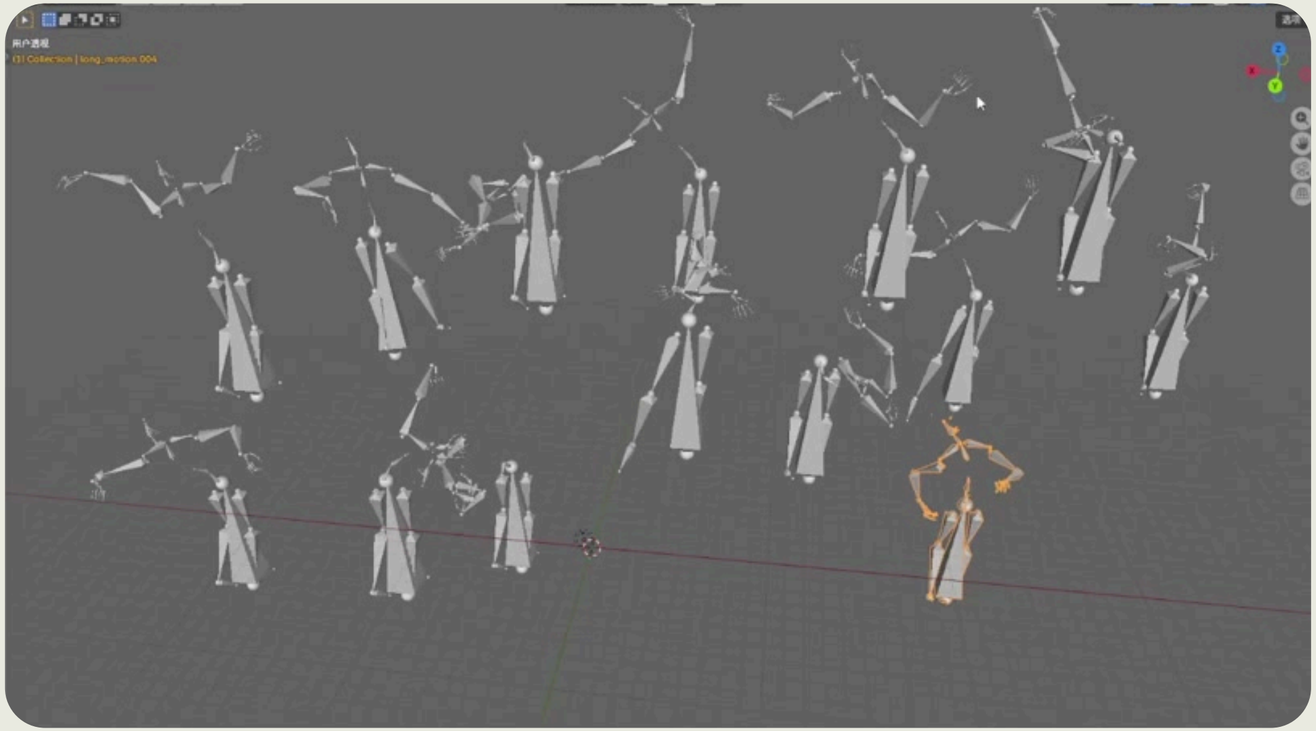




NUM	MODEL	Train		Val		Test (interloctr)	
		MSE	AVE	MSE	AVE	MSE	AVE
1	Baseline (Sequence - Sequence)	80.28	183.55	63.25	157.78	187.52	631.31
2	Baseline with Diffusion Backbone	304.12	83.14	293.20	93.25	279.10	101.87

Outcome：

Despite high-frequency jitter and occasional artifacts, the generated motion exhibits **realistic amplitude** and **continuous dynamics**. unlike the baseline which suffers from mean **collapse** (freezing), our diffusion model successfully synthesizes expressive gestures throughout the entire sequence.



Conclusion:

Core Architecture: Implemented a **Transformer-based Gaussian Diffusion** model utilizing 12D motion features to achieve robust **text-to-gesture synthesis**.

Engineering Optimization: Integrated **SVD orthogonalization** and **Gaussian smoothing** to guarantee geometric validity and eliminate high-frequency jitter.

Evaluation Results: Achieved **low AVE**, confirming the model captures **authentic motion dynamics** and successfully **avoids mean collapse**.

Pros & Cons

Avoids Mean Collapse

Geometric Robustness

Strong Semantic Control

High-Frequency Jitter

Slow Inference Speed

Lack of Physical Constraints

Long-Sequence Consistency

Future Improvements

Training Objectives

Architecture Upgrade

Multi-modal Input

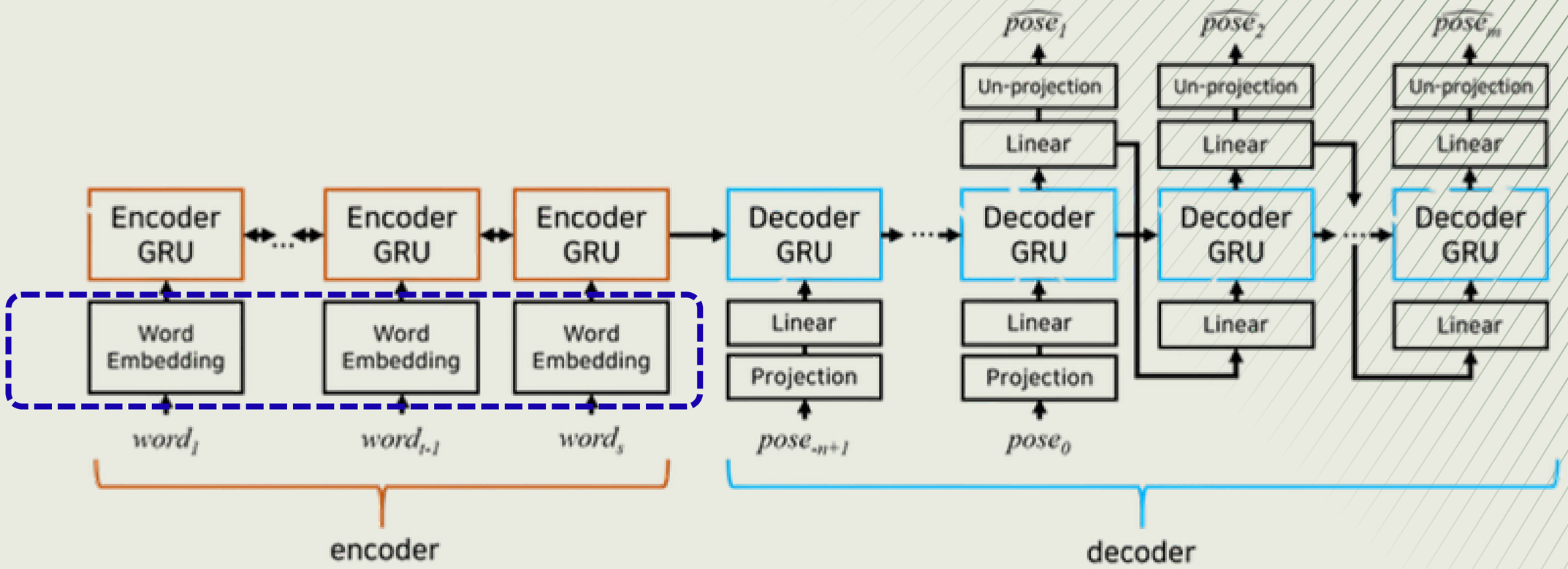
Long-Sequence Generation

Advancement: using BERT embedding

Proposed advancements

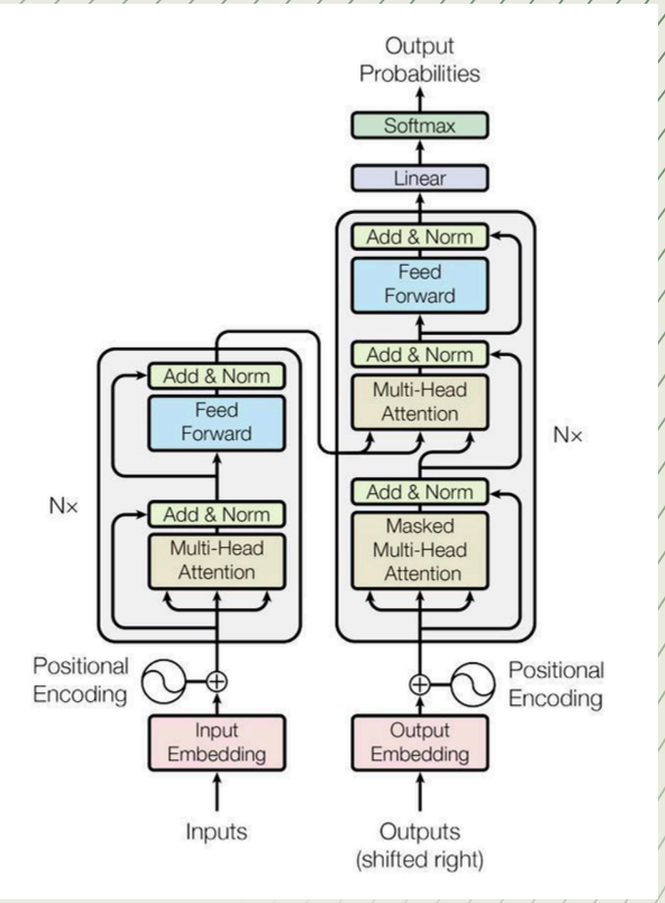
AAdvancements:

- Change the **input** → word embedding to **BERT** based (instead of GLOVE) embedding



Extraction of BERT features

- **Input:** Transcript files (.tsv) containing words and timestamps.
- **Core Model:** bert-base-multilingual-cased (Frozen).
- **Logic:** Process text sentence-by-sentence.
- **Output:** Numpy archives containing semantic vectors and timing data.



Extraction of BERT features

- **Semantic & Temporal Segmentation Logic:**
  - **Iterative Reading:** The script reads raw words from the subtitle wrapper.
  - **Sentence Detection:** Accumulates words until a delimiter is found (., ?, !).
  - **Timestamp Tracking:**
    - **Start:** Time of the first word in the phrase.
    - **End:** Time of the last word in the phrase.
  - **Result:** A clean list of phrases mapped to specific time intervals.

Extraction of BERT features

- **Tokenization Process:**
  - **Special Markers:** Adds **[CLS]** (Start) and **[SEP]** (End) to every phrase.
  - **Sub-word Splitting:** Handles complex words by breaking them down (e.g., embeddings: em, ##bed, ##ding, ##s).
  - **Tensor Conversion:** Converts strings into numerical IDs.

"Hello World" → ['[CLS]', 'Hello', 'World', '[SEP]'] → [101, 7592, 2088, 102]

Extraction of BERT features

- **Hidden State Extraction:**
  - **Model Forward Pass:** model(input, output_hidden_states=True).
  - **Layer Selection:** We discard the final output and grab the **last 4 hidden layers**.
  - **Concatenation:** torch.cat(last_four, dim=2).
  - **Dimensionality Math:** 768 (BERT dim) x 4 (Layers) = 3072 dimensions

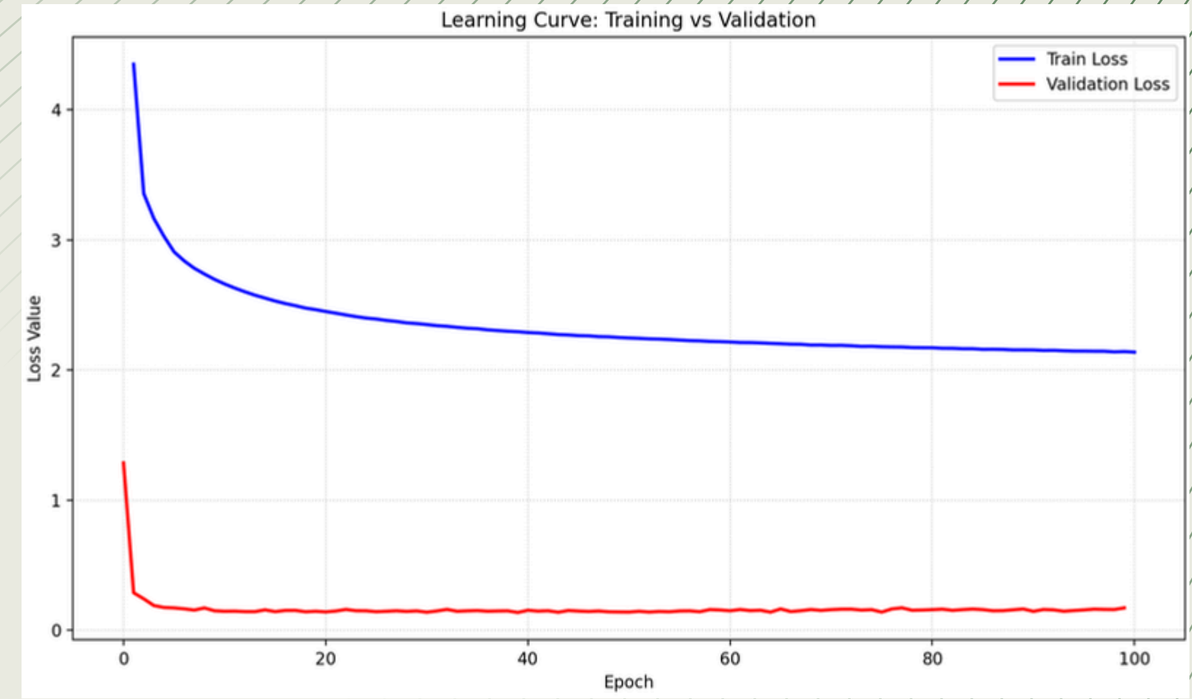
Extraction of BERT features

- **File Structure:**
  - Saved as **Numpy Zipped Archives** (.npz).
  - **Keys:**
    - 'phrase_0', 'phrase_1', etc.: The 3072-dim embedding matrices.
    - 'intervals': A generic array containing [start_time, end_time] for every phrase.
  - **Why?** Ensures perfect synchronization between the semantic meaning (Embedding) and the animation timing (Intervals).

Quantitative Results

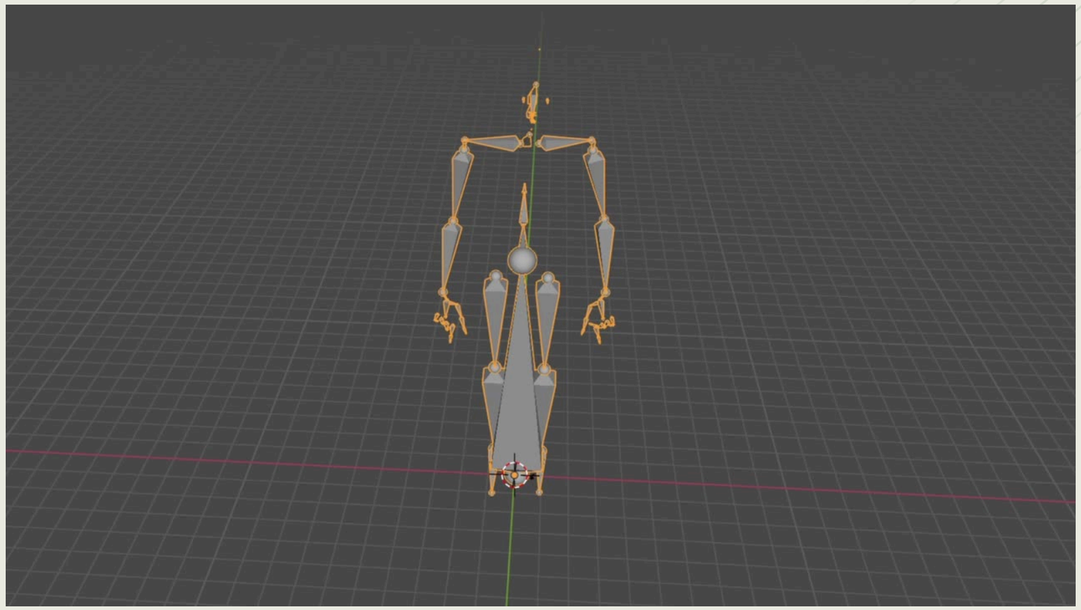
NUM	MODEL	Train		Val		Test(interloctr)	
		MSE	AVE	MSE	AVE	MSE	AVE
1	baseline	80.28	183.55	63.25	157.78	187.52	631.31
2	Baseline + BERT	79.45	181.2	59.6	159.1	182.4	628.15

Error for Baseline + BERT

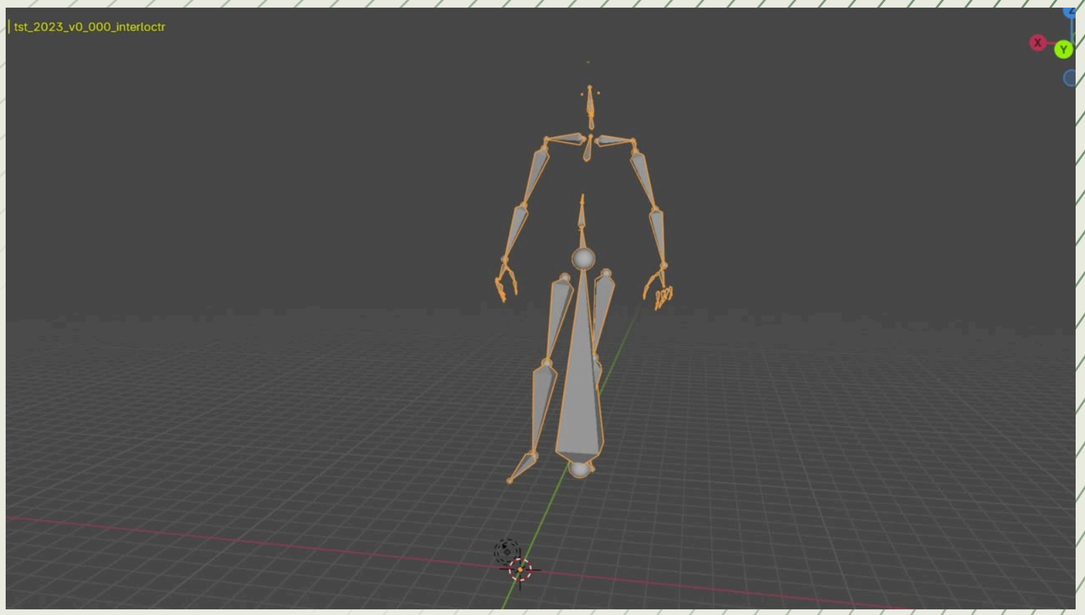


Learning Curve for Baseline + BERT

Qualitative Results



Interlocutor animation generated for Baseline + BERT

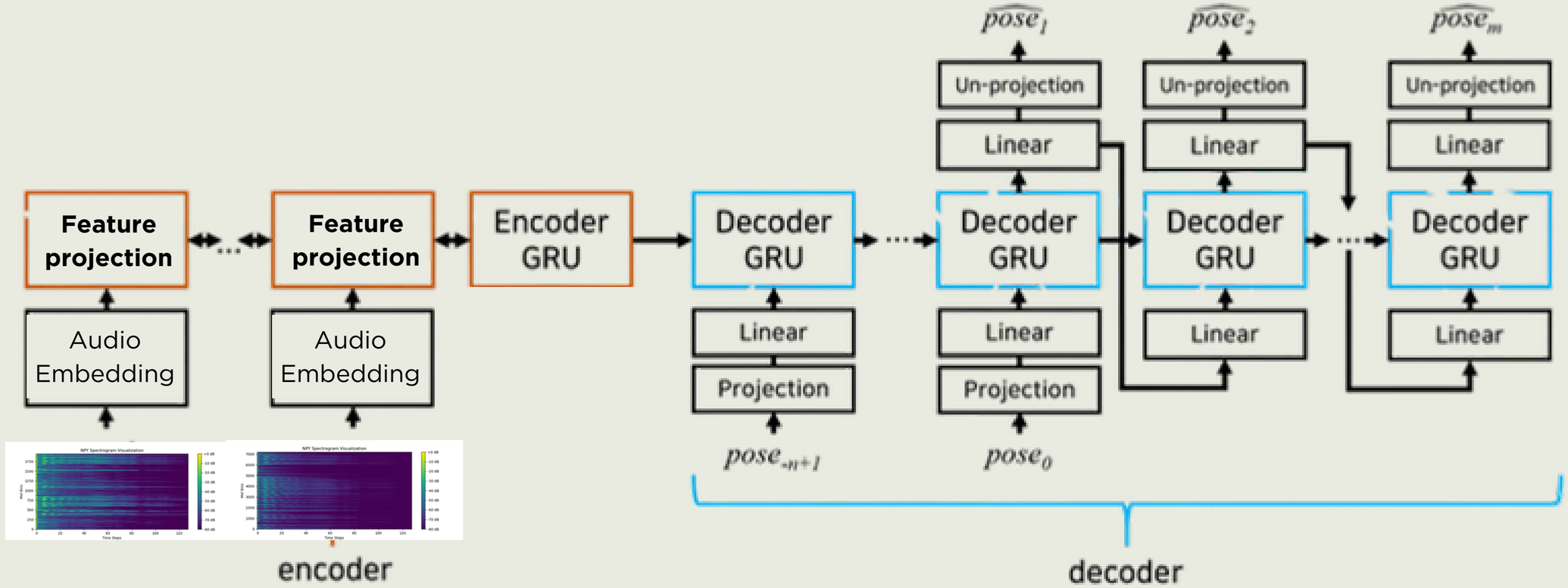


Interlocutor animation from dataset

The model generates coherent upper-body gestures, proving that semantic context drives expressiveness.

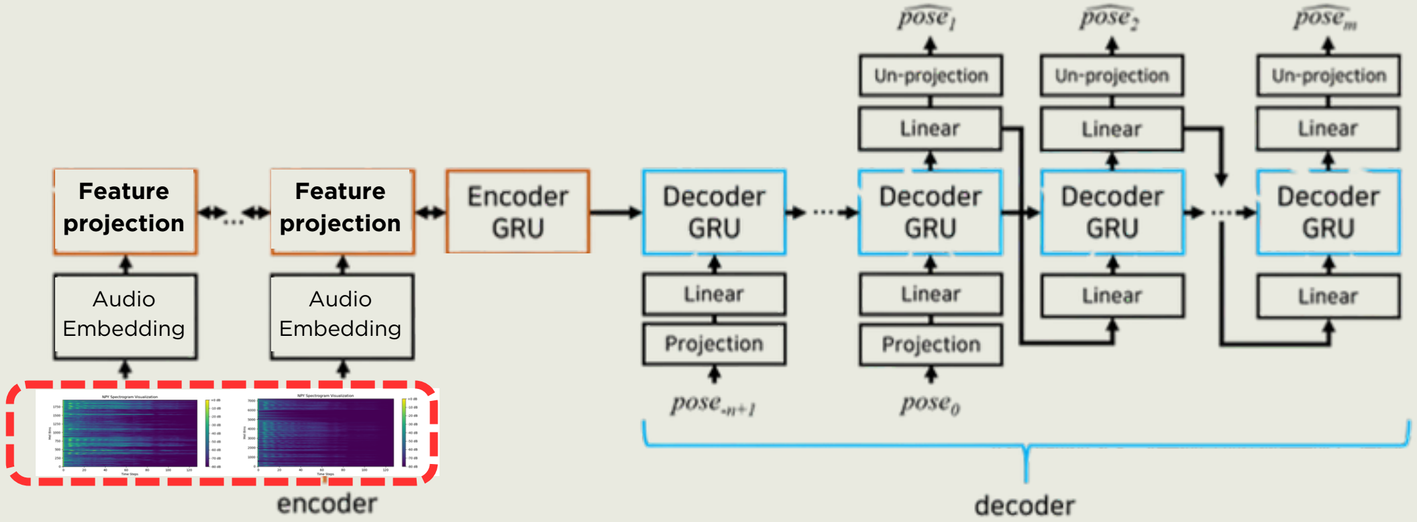
Advancement: audio processing

Pipeline



Input the audio as embedding,after the feature projection, put it into the encoder before

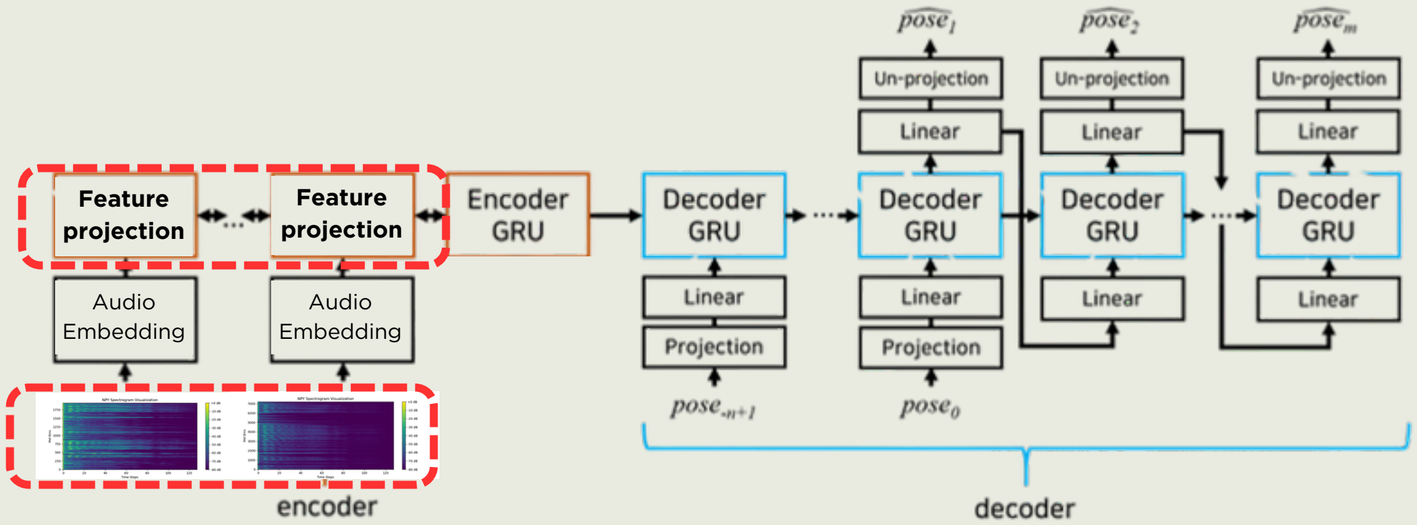
Pipeline



Data Preparation Phase (DataPreprocessor)

- **Input: Raw audio spectrogram features in .npy format.**
- Uses a **Sliding Window** to convert variable-length long videos into uniform sequences of length n_poses.
- sample_audio = clip_audio_features[start_idx : fin_idx] The audio features were pre-extracted to match the frame rate of the poses
- When the sliding window reaches the very end of a clip, there might not be enough frames left to fill the n_poses requirement. Instead of padding with zeros, use **Edge Padding**.
- **Output Shape: (Batch, Time, 128)**

Pipeline



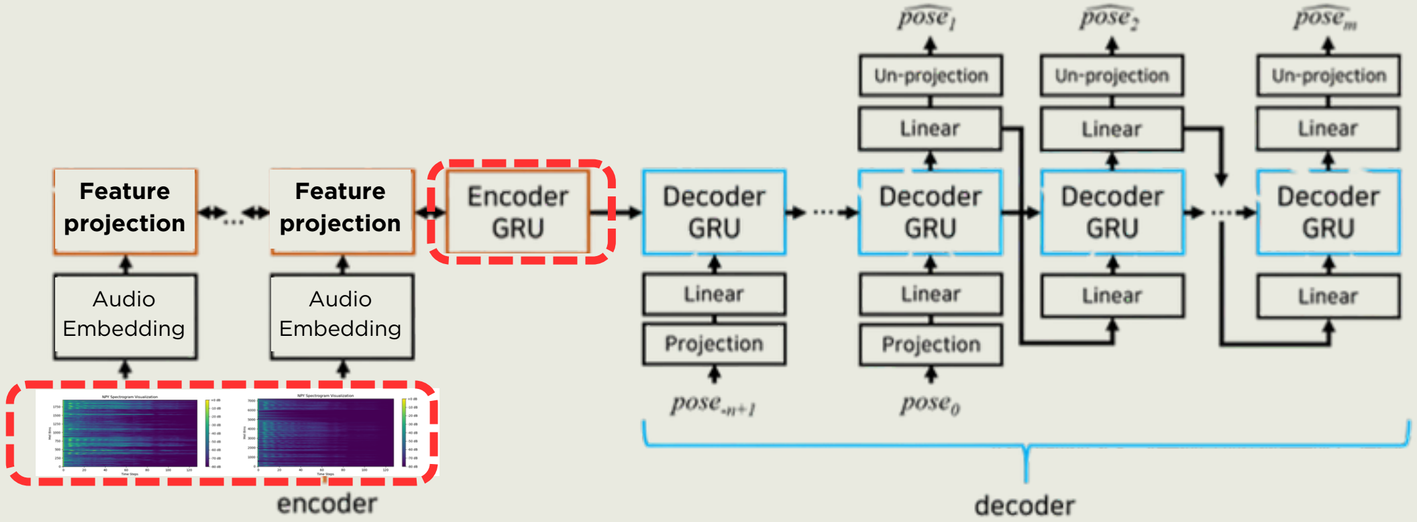
Feature Projection Layer (AudiofeatureProjector)

- This stage maps raw features into the model's hidden embedding space.

- Transforms (B, T, 128) into (B * T, 128)
- Linear(128 → 512) → LeakyReLU → Dropout
- Linear(512 → 512) → BatchNorm1d → LeakyReLU
- Linear(512 → Hidden_Size)
- Reshapes back to (B, T, Hidden_Size)

- **Transposes to (T, B, Hidden_Size)**

Pipeline



Semantic Encoding Layer (EncoderRNN)

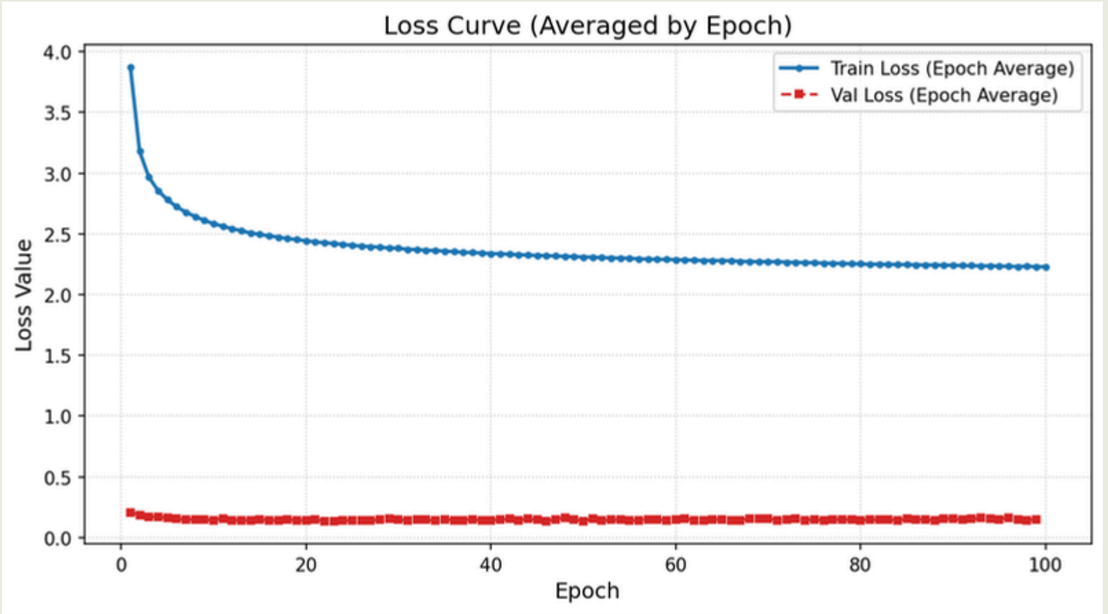
- Utilizes a **Bidirectional GRU** to capture temporal audio context.
- **Input: (T, B, Hidden_Size)**
- **Optimization:** Implements pack_padded_sequence to handle variable-length sequences and improve computational efficiency.
- **Feature Fusion (Bidirectional Sum):** outputs = forward_out + backward_out
- Summing bidirectional outputs preserves global context while maintaining dimension consistency with the Decoder.

- **Encoder Outputs: (T, B, Hidden_Size)**

Outcome:

The Train Loss (Blue) starts near 4.0 and steadily decreases to approximately 2.2

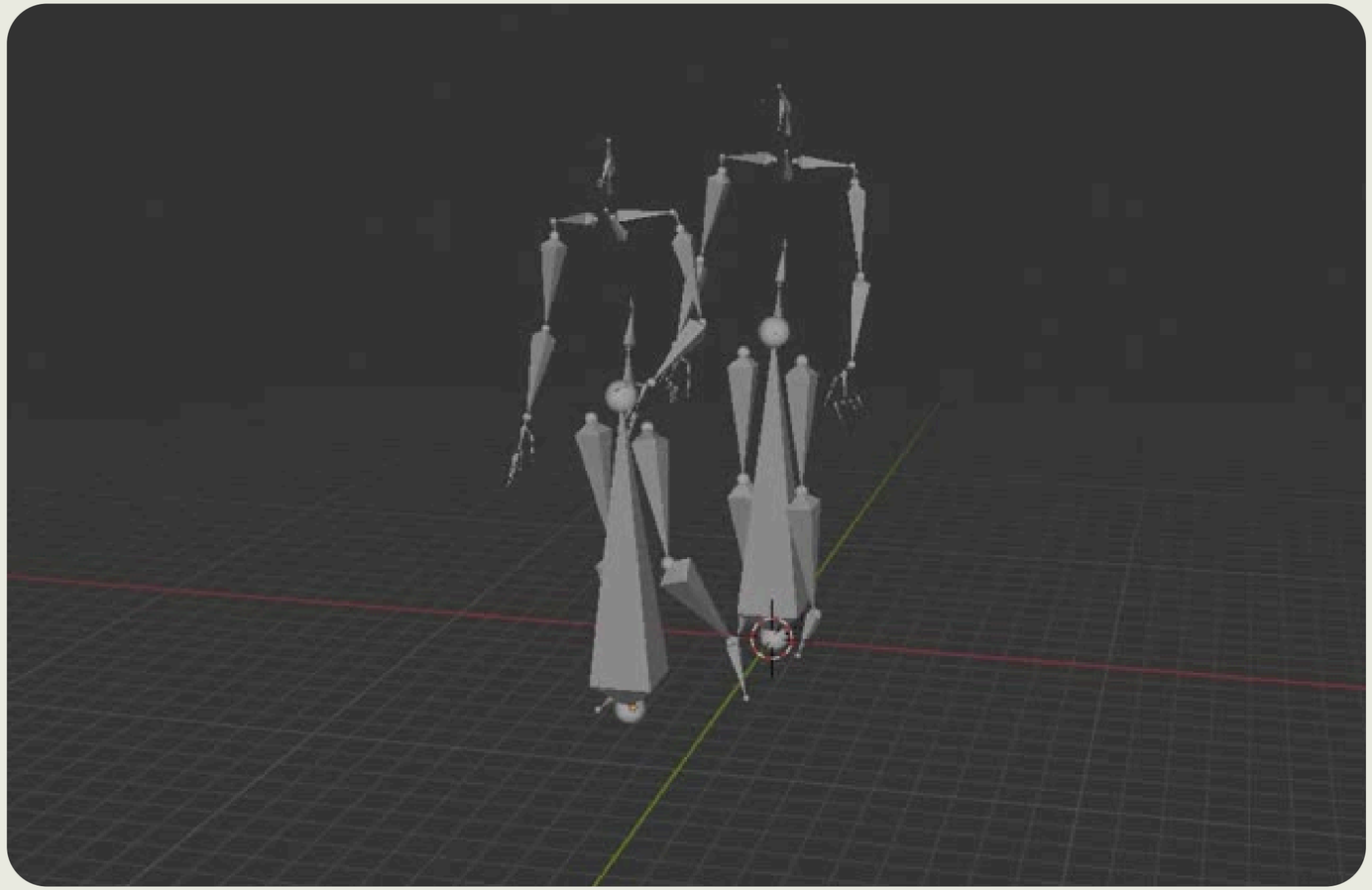
The Mse and the Ave is close to the baseline model



NUM	MODEL	Train		Val		Test (interloctr)	
		MSE	AVE	MSE	AVE	MSE	AVE
1	Baseline (Sequence - Sequence)	80.28	183.55	63.25	157.78	187.52	631.31
2	Baseline with audio as input	80.02	196.84	60.13	164.64	185.83	635.09

Outcomes

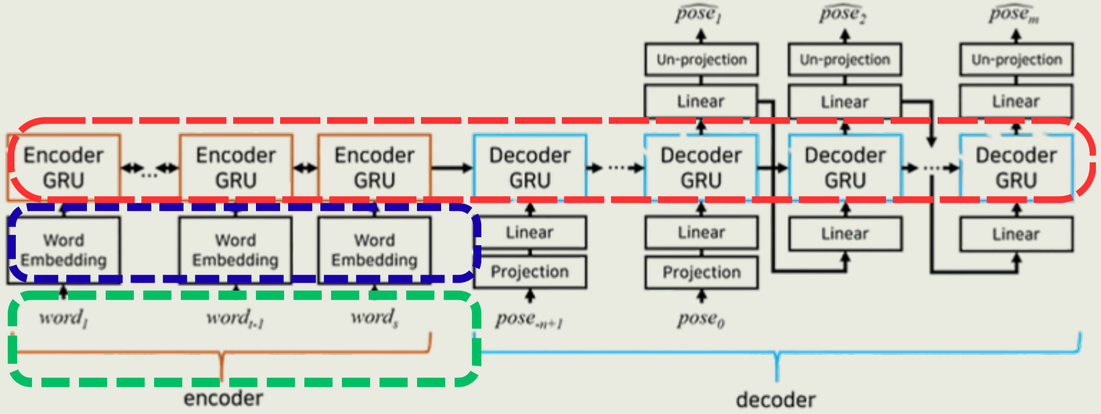
The left (front) is ground truth. We can see that there is movement generated by our method, even though the dynamic of movement happened mostly in the beginning part of the entire movements.

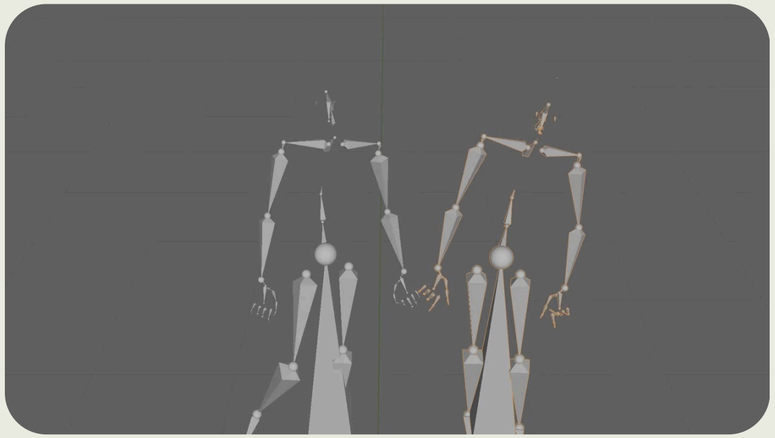


Conclusions

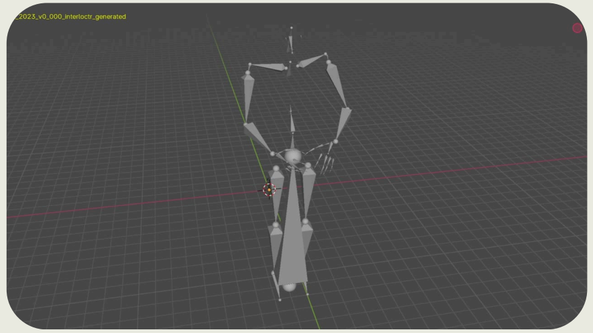
General conclusion

- The basis sequence-to-sequence model is capable to generate basic human movements given the text input of speaker
- The results from introduction of Diffusion based model are still inferior to baseline, but it shows some promising early generated movement.
- Changing the input to the audio improves the quantiative results (from the baseline), but some generated results are less animated than the baselines.
- Results from injection of new text embedding (BERT) results in general comparable results.





NUM	MODEL	Train		Val		Test (interloctr)	
		MSE	AVE	MSE	AVE	MSE	AVE
1	Baseline (Sequence - Sequence)	80.28	183.55	63.25	157.78	187.52	631.31
2	Baseline with Diffusion Backbone	304.12	83.14	293.20	93.25	279.10	101.87
3	Baseline with audio as input	**80.02**	**196.84**	60.13	**164.64**	185.83	**635.09**
4	Baseline + BERT	79.45	181.20	**59.60**	159.10	**182.40**	628.15



features (idem aboves)

What we learn

- From modelling:
  - Basic of generative model
  - Changing the input, we need to pay attention to the dimensionality of the input and output vector.
  - Diffusion
  - … …

- From computation perspective:
  - Server use
  - Library
  - LMDB
  - … …

Reference

[1] Proposed Method (Base Paper)
Y. Yoon, W. Ko, M. Jang, J. Lee, J. Kim, and G. Lee, "Robots Learn Social Skills: End-to-End Learning of Co-Speech Gesture Generation for Humanoid Robots," IEEE International Conference on Robotics and Automation (ICRA), 2019.

[2] Dataset Source
The GENEA Challenge 2023 Dataset (based on Talking With Hands 16.2M).
H. Joo et al., "Panoptic Studio: A Massively Multiview System for Social Motion Capture," ICCV, 2015.
G. Lee et al., "Talking With Hands 16.2M: A Large-Scale Dataset for Learning Gestures," ICCV, 2019.

[3] Future Work (Diffusion Model Implementation)
Minimal Text Diffusion: A minimal implementation of diffusion models for text generation.
Source Code: https://github.com/madaan/minimal-text-diffusion
Author: A. Madaan et al.

Thank you.

WEI, Yujia, Mateus
Supervisor: Decky
1 December, 2025