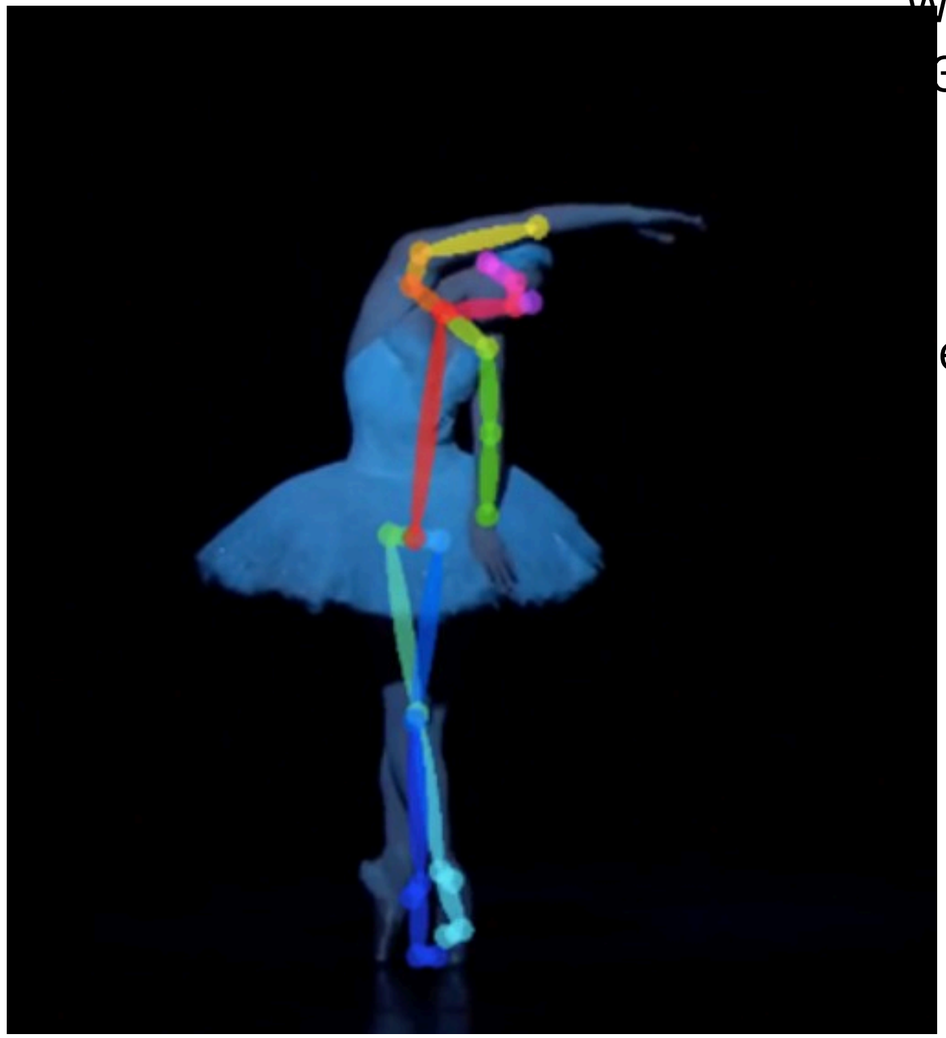
AI CHOREOGRAPHY PM REVIEW MEET

About our Project



Welcome to our groundbreaking project focused on music-based 3D dance generation using AI technology. In the world of video games, animations, movies, and music videos, creating captivating and synchronized dance sequences is an arduous and resource-intensive task. However, our project aims to evolutionize this process by developing an AI model capable of producing new and eye-catching dance moves that perfectly complement the music.

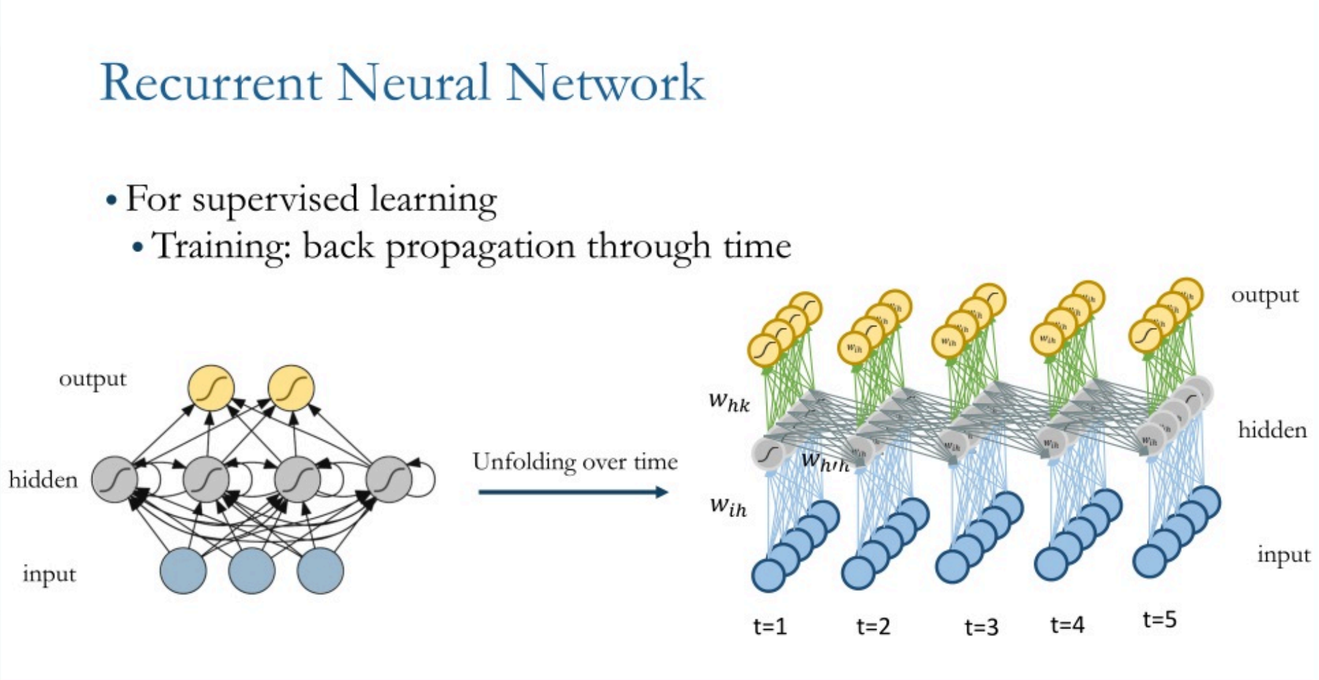
Our Goal

Develop a model that can take any music as input
and generate full dance moves
syncing with the beats of the music.

SECTION 1 : LEARNING PHASE

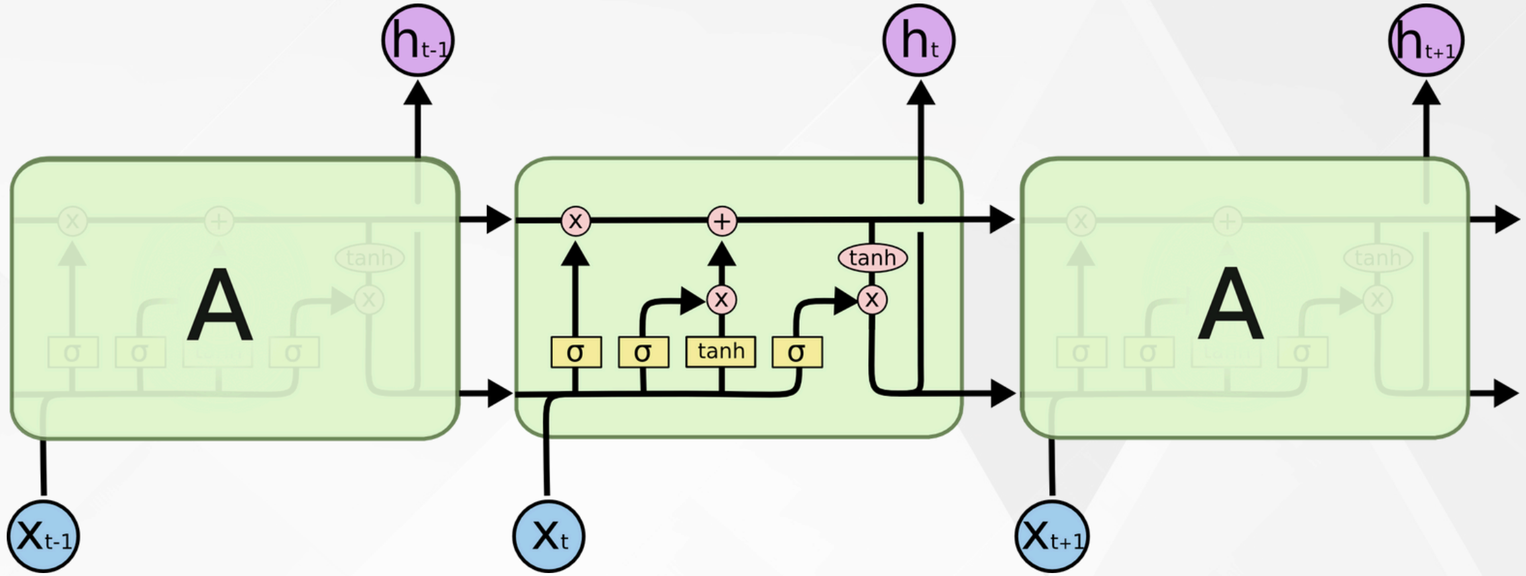
(SUMMARY OF ALL OF OUR THE LEARNING TASKS)

RNNs- Recurrent Neural Networks



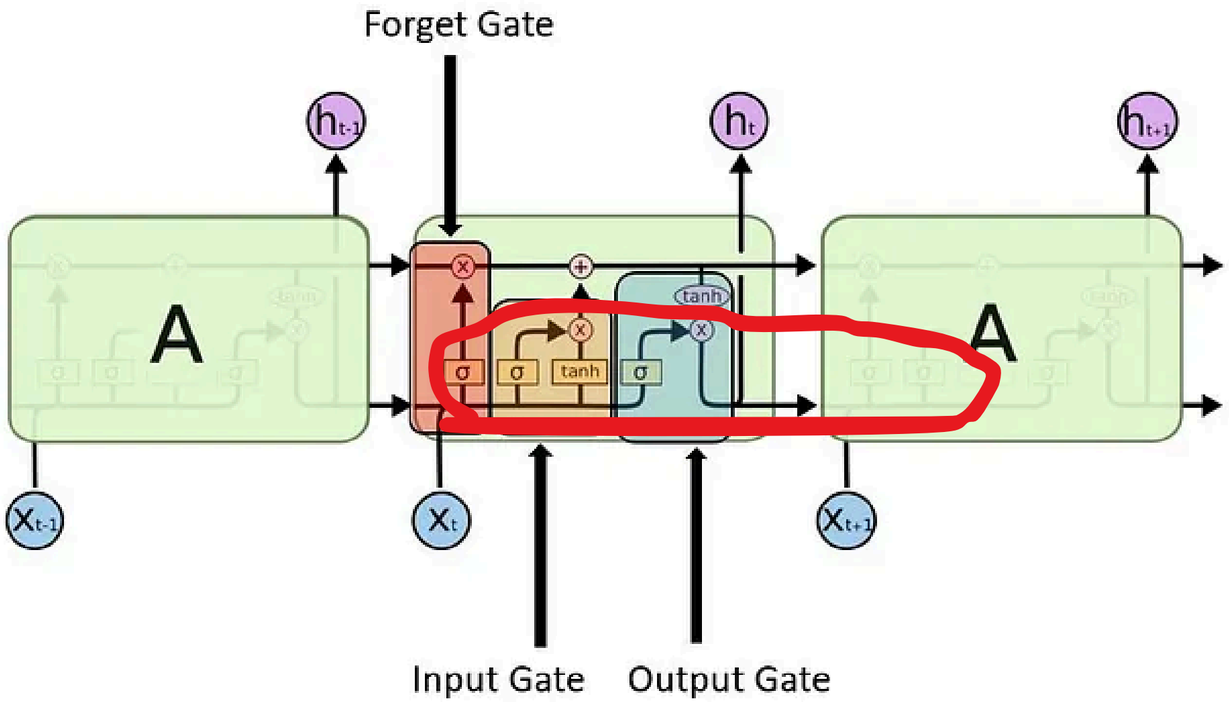
Recurrent Neural Network(RNN) is a type of **Neural Network** where the output from the previous step is fed as input to the current step. In traditional neural networks, all the inputs and outputs are independent of each other, but in cases when it is required to predict the next word of a sentence, the previous words are required and hence there is a need to remember the previous words . The main and most important feature of RNN is its Hidden state, which remembers some information about a sequence.

LONG SHORT TERM MEMORY (LSTM'S)



Long Short Term Memory networks – usually just called “LSTMs” – are a special kind of RNN, capable of learning long-term dependencies.
Remembering information for long periods of time is practically their default behavior, not something they struggle to learn!
All recurrent neural networks have the form of a chain of repeating modules of neural network. In standard RNNs, this repeating module will have a very simple structure, such as a single tanh layer.
LSTMs also have this chain like structure, but the repeating module has a different structure. Instead of having a single neural network layer, there are four, interacting in a very special way.

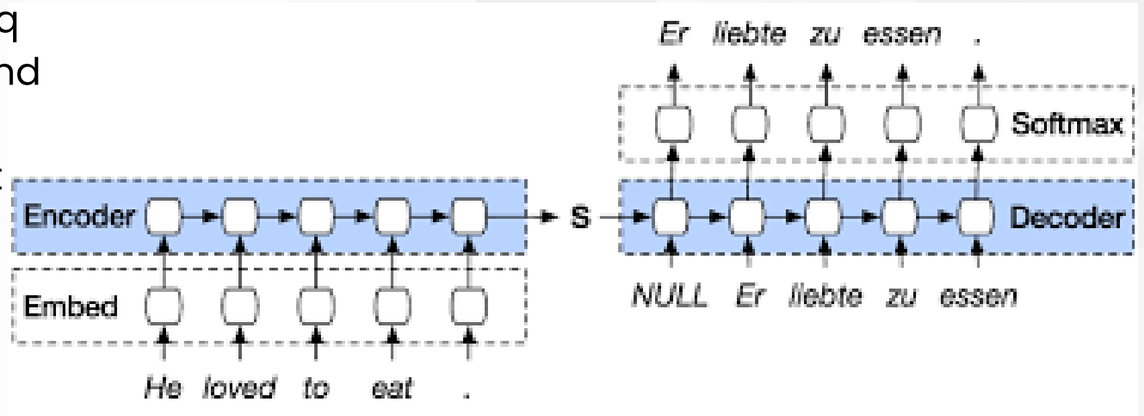
LSTM Architecture



The key to LSTM's, the cell state or the context vector, with each individual cell having gates that regulate flow of information.

SEQ2SEQ MODELS

We decided to first understand encoder-decoder type models starting with Seq2Seq models. Encoders reads the input vector and converts it into a context vector and decoder uses this context vector to predict the next possible states of the sequence

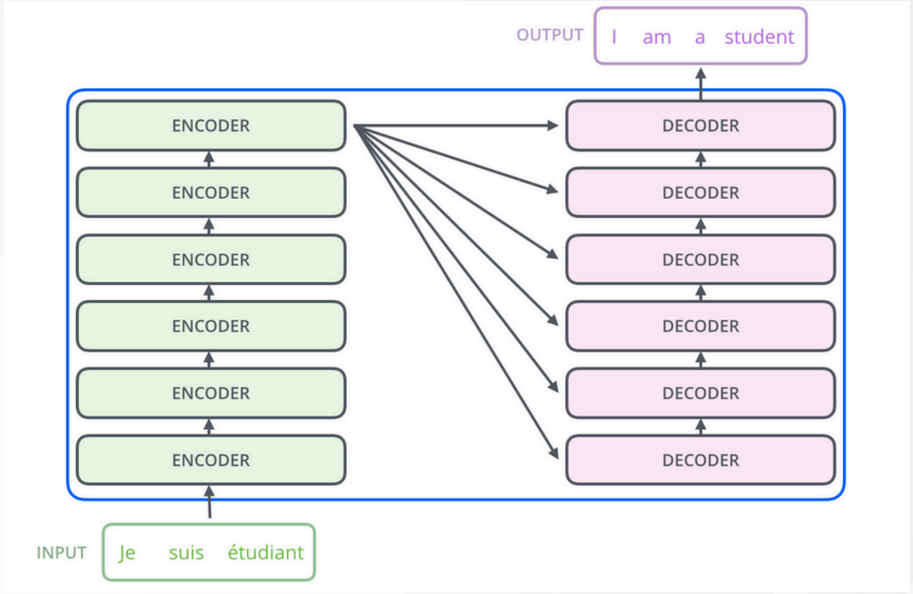


We implemented the Seq2Seq model on simple machine translation task .

THE TRANSFORMER

The Transformer – a model that uses **attention** to boost the speed with which these models can be trained.

The Transformer is basically our LSTM model but with the added benefit of self and multi-head attention.
The Transformer has the basic structure made of an encoder component and a decoder component. The encoder component comprises of a few encoders and same for the decoder.



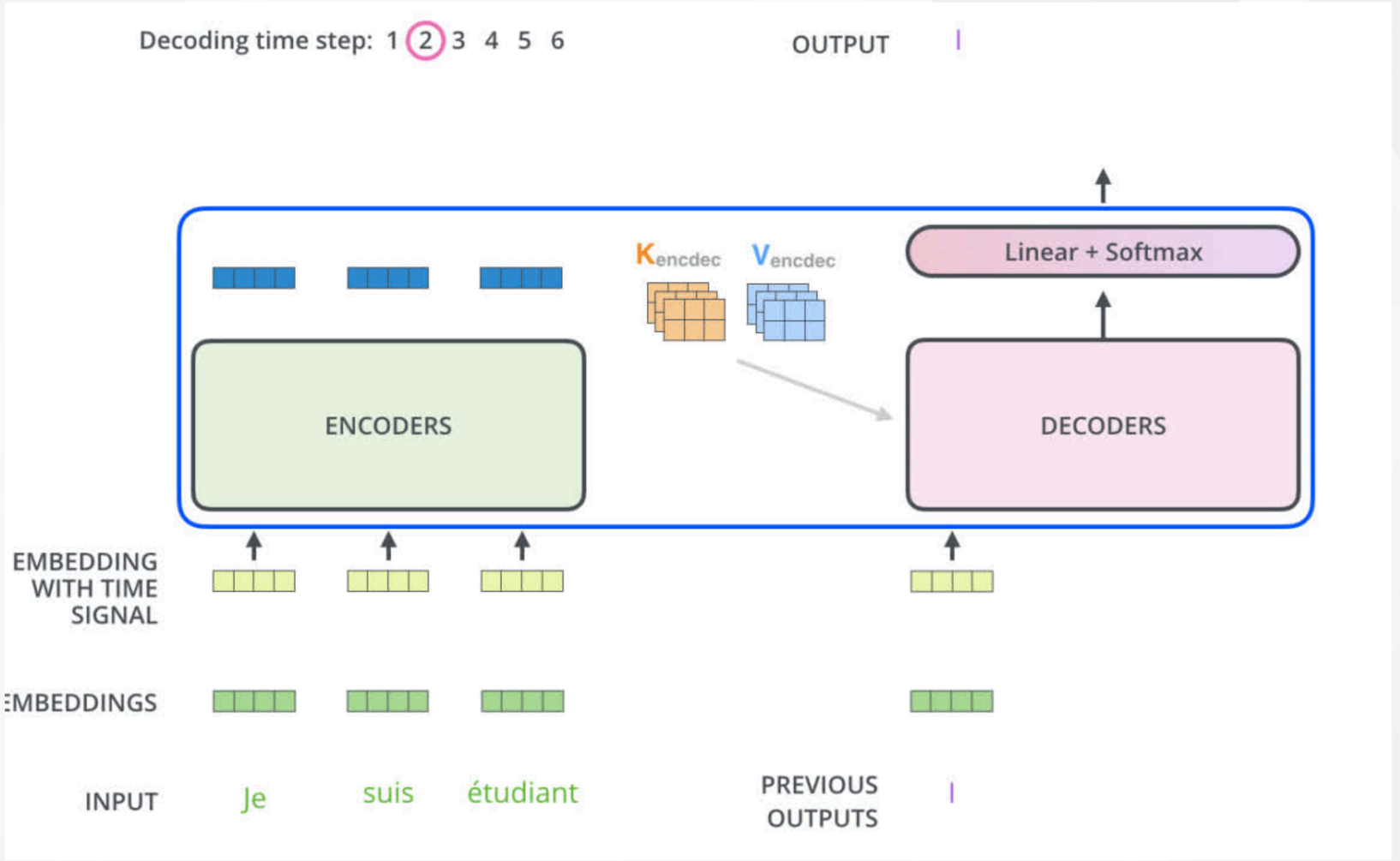
ATTENTION:

Self-attention is the method the Transformer uses to bake the understanding of other relevant components of the dataset into the one we're currently processing. For example: In translation, if you have the sentence
"The animal didn't cross the street because it was too tired"
then attention helps us to associate the word it with animal.

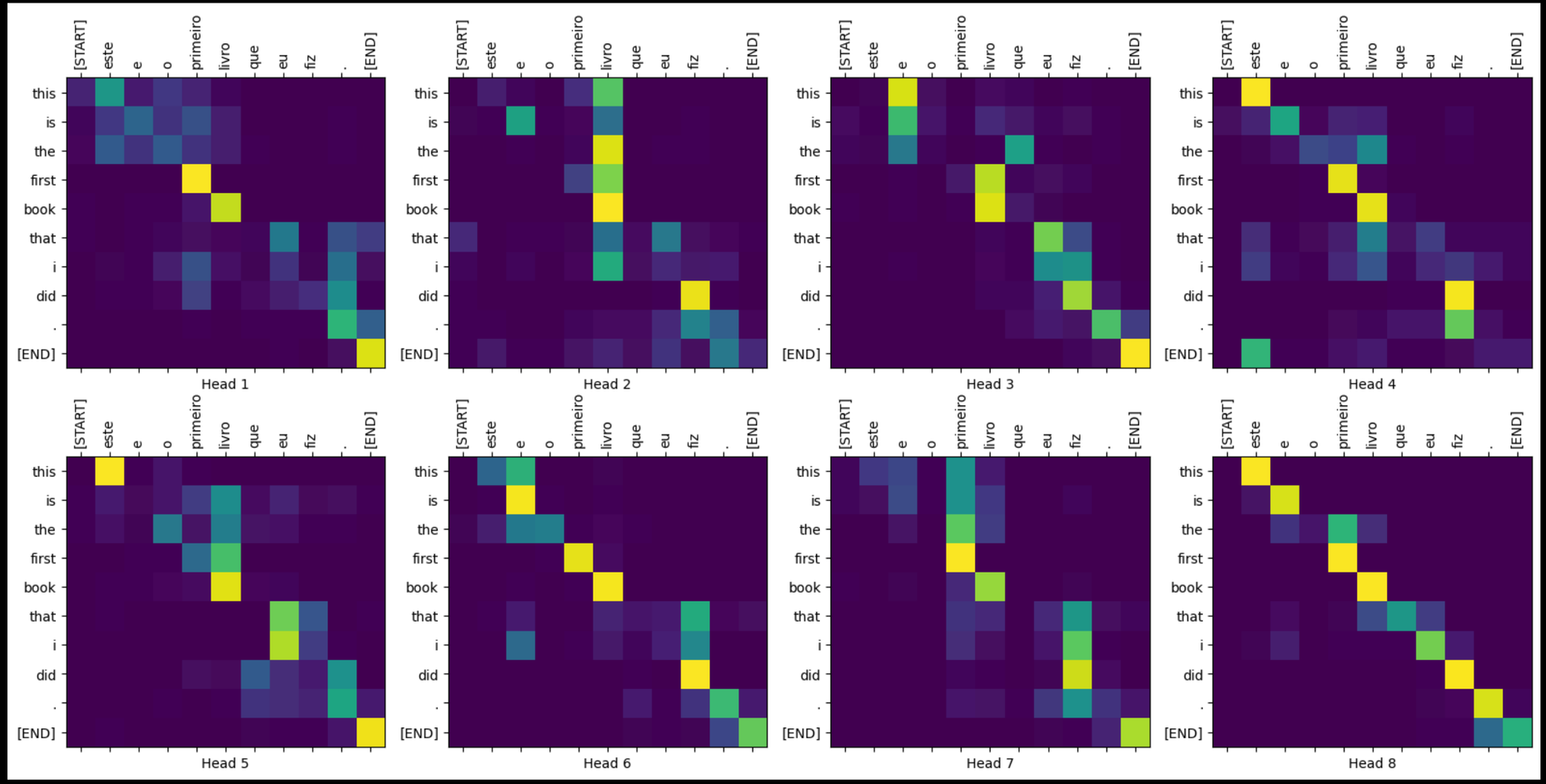
ATTENTION IS ALL YOU NEED!

RESEARCH PAPER

SCHEMATIC REPRESENTATION OF TRANSFORMERS



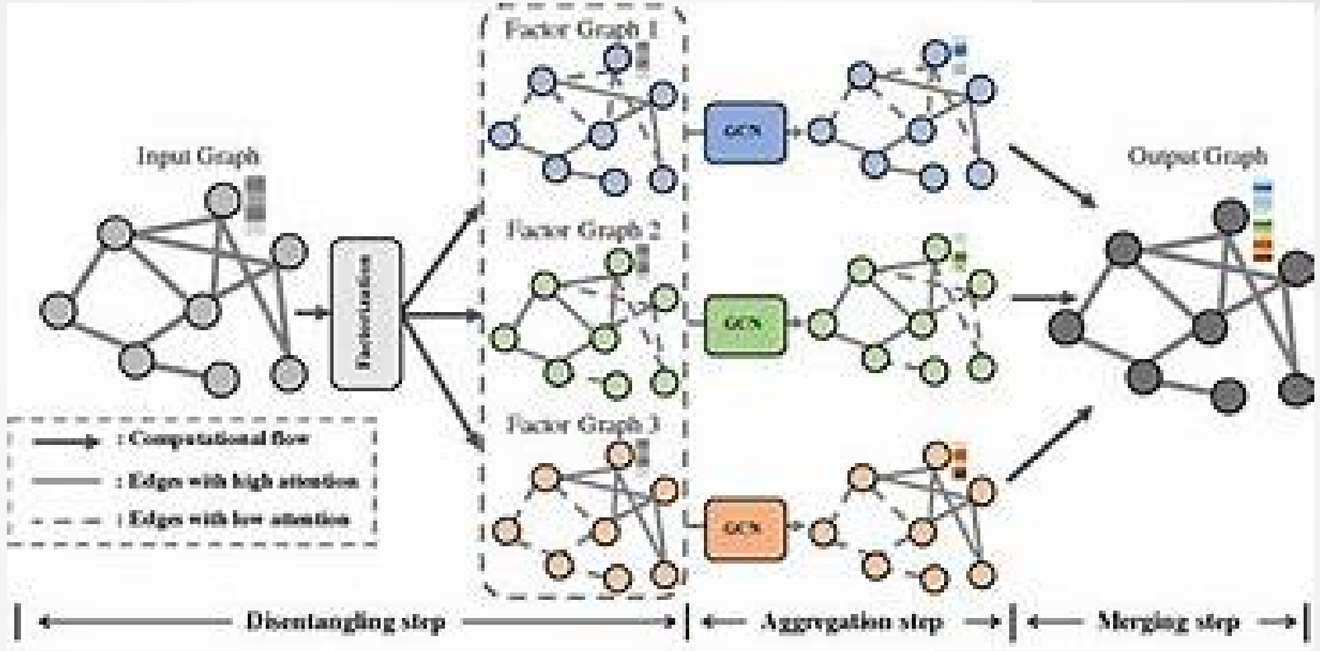
INPUT: ESTE É O PRIMEIRO LIVRO QUE EU FIZ.
PREDICTED TRANSLATION: THIS IS THE FIRST BOOK I MADE .



REAL TRANSLATION: THIS IS THE FIRST BOOK I'VE EVER DONE.

GCNs- Graph Convolutional networks

GCN is a type of convolutional neural network that can work directly on graphs and take advantage of their structural information. it solves the problem of classifying nodes (such as documents) in a graph (such as a citation network), where labels are only available for a small subset of nodes (semi-supervised learning).

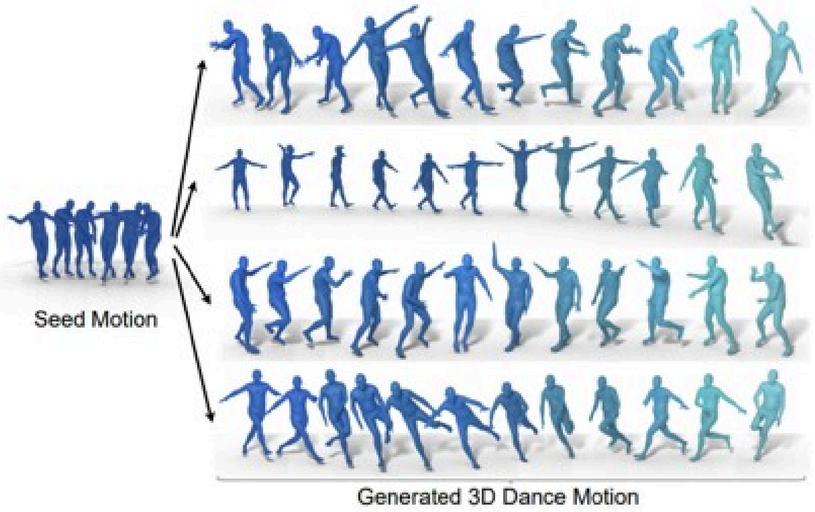


The architecture is similar to a traditional CNN but it takes graphs as input, also the convolution and pooling operations are different in principle. Here what we do is have the input as graph, which has nodes and edges. Each node contains some information about an object and the edges represent the relation between the 2 nodes. With each forward pass step, some information is shared between adjacent nodes. This is used to determine relations between 2 objects.

SECTION 2 : DATASET ANALYSIS

(BRIEF INTRODUCTION TO THE DATASET WE ARE USING)

Dataset we will be using...



Multi-modal dataset of 3D dance motion & music...

AIST++

The dataset will be taken from a research paper, AI Choreographer.

....the largest dataset of the kind we will be requiring.....

Contains ⟹ 5.2 hours of 3D dance motion in 1408 sequences, covering 10 dance genres (Old School (Break, Pop, Lock and Waack) and New School (Middle Hip-hop, LA-style Hip-hop, House, Krump, Street Jazz and Ballet Jazz) with multi-view videos with known camera poses.

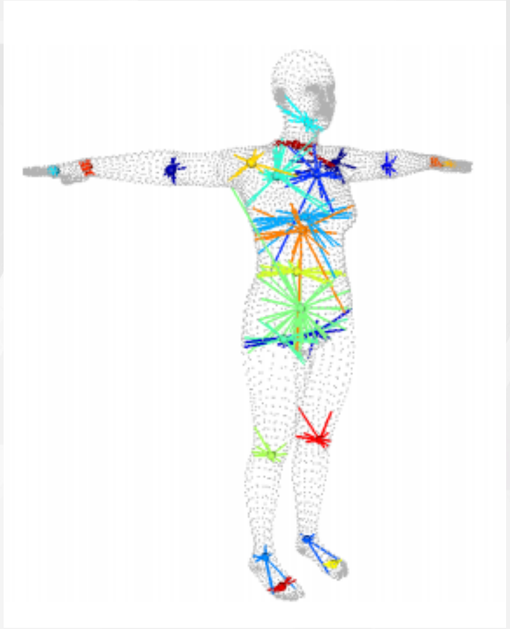
Each genre= 85% basic choreographies + 15% advanced choreographies.

Special sequence model FACT is applied to the above dataset...

FACT in short: - model involving a deep cross-modal transformer block with full-attention that is trained to predict N future motions instead of one, unlike normal transformer. This minimizes the accumulated error in each step of attention.

SKINNED MULTI PERSON LINEAR MODEL

3D OBJECTS ARE OFTEN REPRESENTED BY VERTICES AND TRIANGLES THAT ENCODES THEIR 3D SHAPE. HOWEVER, FOR OBJECTS LIKE HUMAN, THE 3D MESH REPRESENTATION COULD BE COMPRESSED DOWN TO A LOWER DIMENSIONAL SPACE WHOSE AXES ARE LIKE THEIR HEIGHT, FATNESS, BUST CIRCUMFERENCE, BELLY SIZE, POSE ETC. THIS REPRESENTATION IS OFTEN SMALLER AND MORE MEANINGFUL.



THE SMPL IS A STATISTICAL MODEL THAT ENCODES THE HUMAN SUBJECTS WITH TWO TYPES OF PARAMETERS:

- **SHAPE PARAMETER:** A SHAPE VECTOR OF 10 SCALAR VALUES, EACH OF WHICH COULD BE INTERPRETED AS AN AMOUNT OF EXPANSION/SHRINK OF A HUMAN SUBJECT ALONG SOME DIRECTION SUCH AS TALLER OR SHORTER.

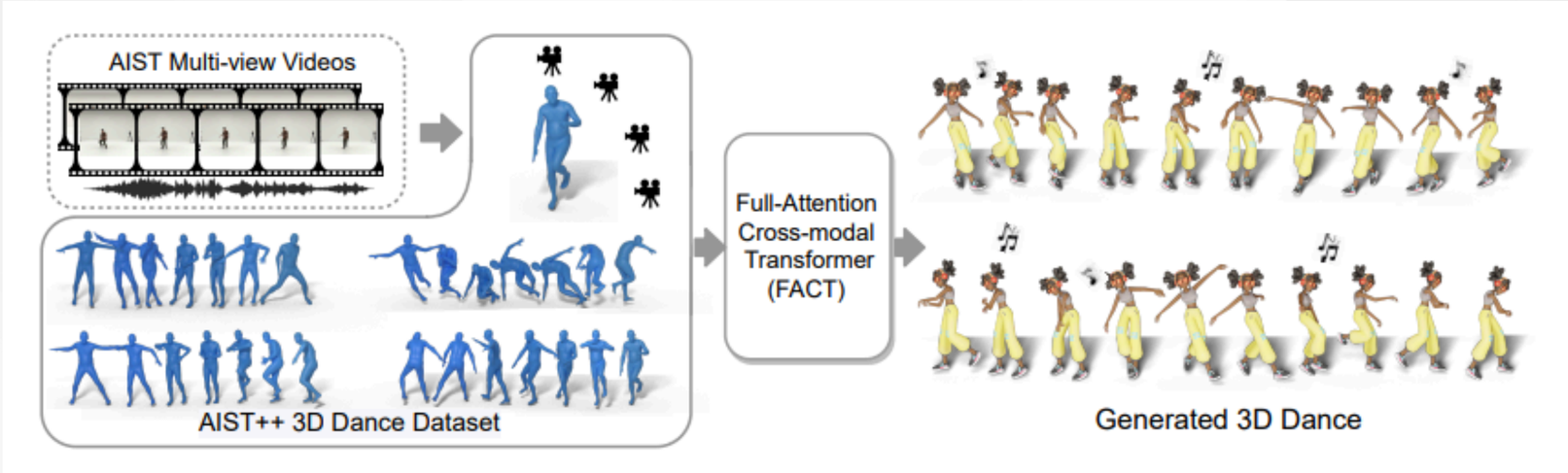
- **POSE PARAMETER:** A POSE VECTOR OF 24X3 SCALAR VALUES THAT KEEPS THE RELATIVE ROTATIONS OF JOINTS WITH RESPECTIVE TO THEIR PARAMETERS. EACH ROTATION IS ENCODED AS A ARBITRARY 3D VECTOR IN AXIS-ANGLE ROTATION REPRESENTATION.

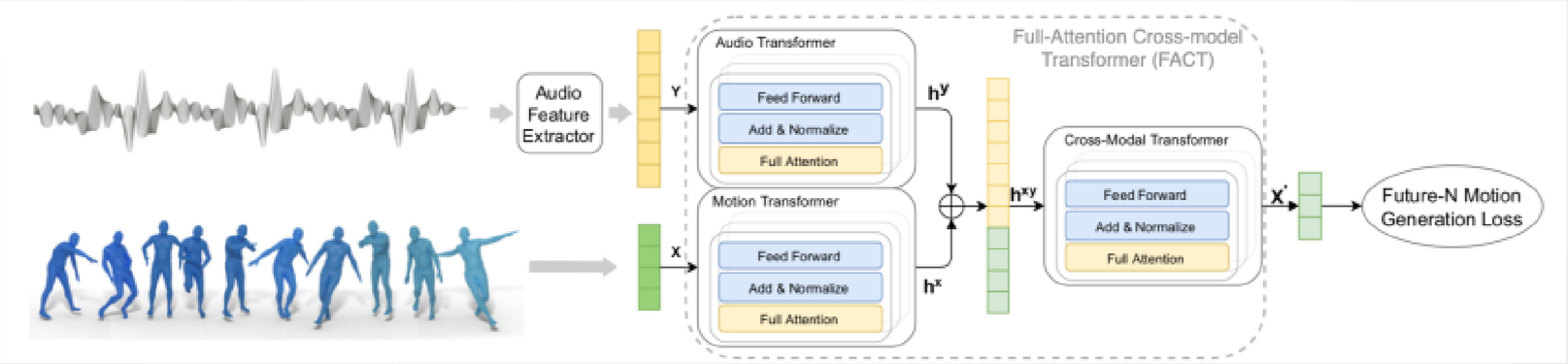
SECTION 3 : LITERATURE REVIEW

(SUMMARY OF ALL THE MAIN PAPERS WE WILL BE USING FOR THIS PROJECT)

PAPER 1 : FACT : FULL ATTENTION CROSS MODEL TRANSFORMER

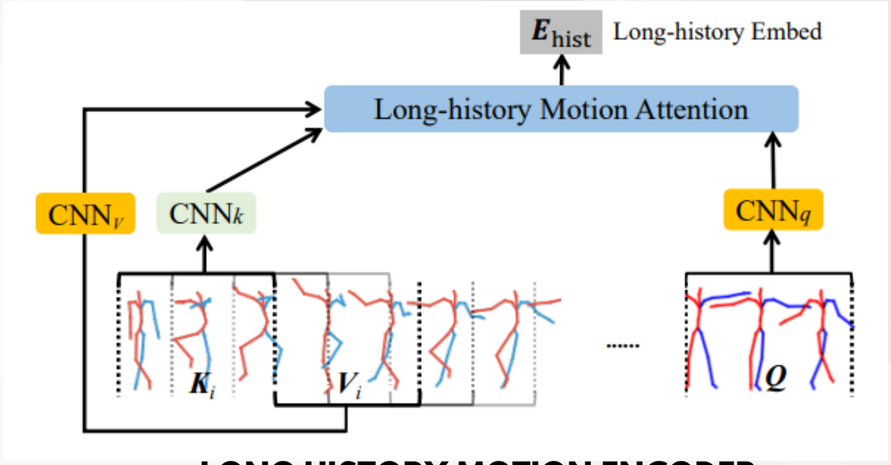
FACT MODEL INVOLVES A DEEP CROSS MODAL TRANSFORMER BLOCK WITH FULL ATTENTION THAT IS TRAINED TO PREDICT N FUTURE MOTIONS



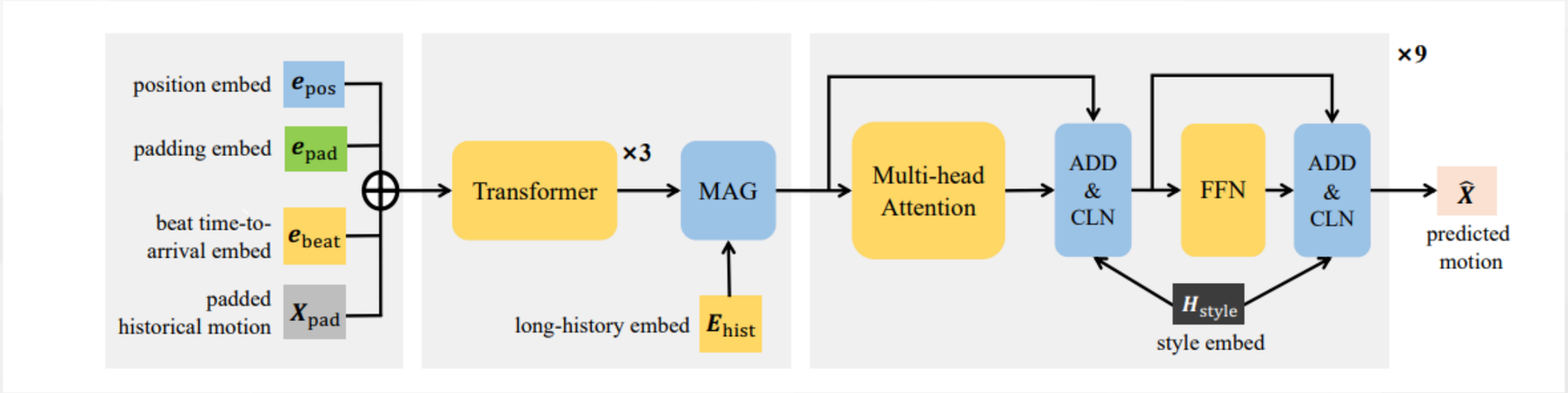


PAPER 2 : ROBUST DANCER: LONG-TERM 3D DANCE SYNTHESIS USING UNPAIRED DATA

- THIS RESEARCH PAPER AIMS TO DEVELOP A NOVEL 3D DANCE SYNTHESIS SYSTEM THAT CAN ROBUSTLY GENERATE IMPRESSIVE DANCES ACCOMPANYING WITH A PIECE OF LONG MUSIC. TO OUR BEST KNOWLEDGE, IT IS SUPPOSEDLY FIRST SYSTEM THAT SUCCESSFULLY USES ONLY UNPAIRED DATA TO ACHIEVE MUSIC-DRIVEN DANCE GENERATION.
- THEY PROPOSE AN EFFICIENT UNPAIRED DATA TRAINING SCHEME TO ALLEVIATE THE PROBLEM OF LACKING DATA.
- THEY HAVE DEVISED AN INNOVATIVE LONG-HISTORY ATTENTION STRATEGY, WHICH EXPLICITLY MAINTAINS THE TEMPORAL COHERENCE BETWEEN MUSIC AND DANCE IN THE LONG-TERM TIME.



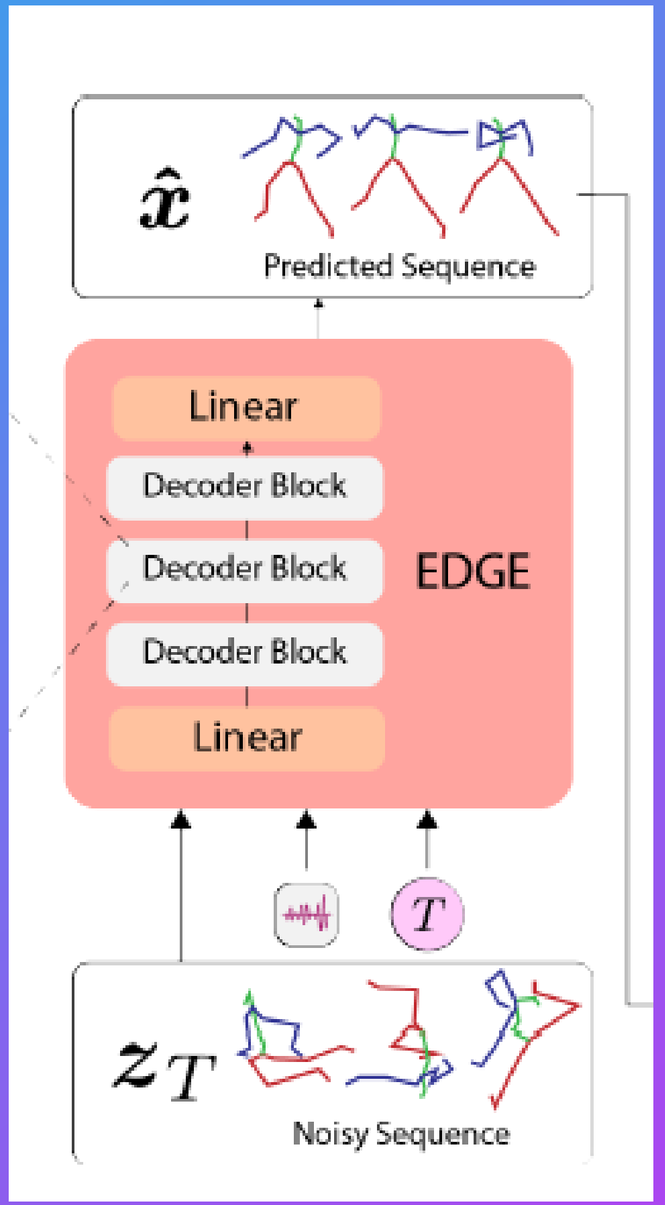
LONG HISTORY MOTION ENCODER



MOTION GENERATOR

Paper 4 : EDGE : Editable Dance Generation From Music

Editable **D**ance **GE**neration (EDGE), is a state-of-the-art method for editable dance generation that is capable of creating realistic, physically-plausible dances while remaining faithful to the input music. EDGE uses a transformer-based diffusion model paired with Jukebox, a strong music feature extractor, and confers powerful editing capabilities well-suited to dance, including joint-wise conditioning, and in-betweening.



Paper 3: A Spatio-temporal Learning for Music Conditioned Dance Generation

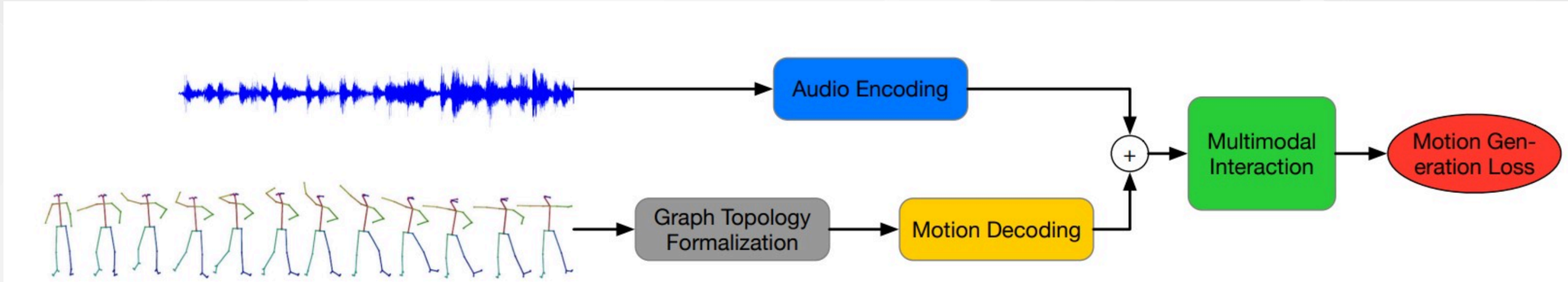


Figure 1: Cross-Modal Dance Generation Overview. We perform dance generation in a two-path spatio-temporal learning mode synchronized with audio features. The audio path consists of blocks of audio encoding, and the motion path consists of spatial-temporal graph topology formalization and layers of motion decoding. The output is a sequence of music-conditioned motions generated from multimodal interaction.

In this paper, what is proposed is:

- A music conditioned dance generation is constructed as a long term sequence to sequence multi modal task and introduce a novel spatio temporal learning framework for skeleton based dance generation.
- Then a positional GCN based block to decode the spatial and temporal features of motion is used.
- Then a regional attention based encoding mechanism for self feature learning and mutual feature fusion is introduced.

CONCLUSION

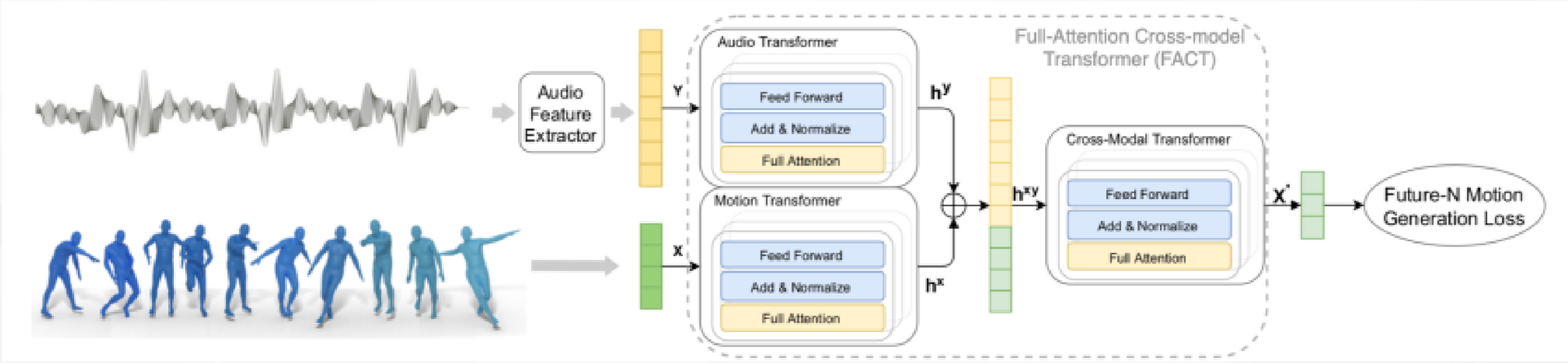
- We have gathered key ideas from all the previously mentioned research papers and combined them together to develop our own model.
- Our finally proposed model uses the spatial learning capabilities of a GCN and temporal learning ability of a transformer and combines them with music embeddings using a cross model attention layer.
- We have already started the implementation of baseline model which is expected to be ready by the end of this month.

Current approach :

We have combined the 2 models, the FACT transformer model and the GCN approach.

1st Component - Full Attention Cross Modal Transformer

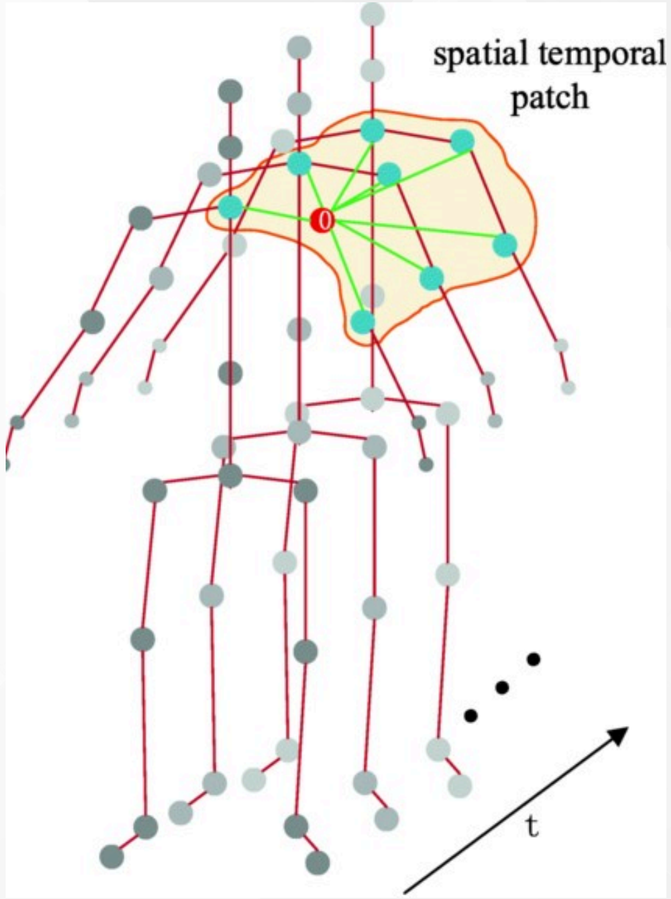
Fact model involves a deep cross modal transformer block with full attention that is trained to predict n future motions



2nd Component : Spatio-Temporal Graph Convolutional networks.

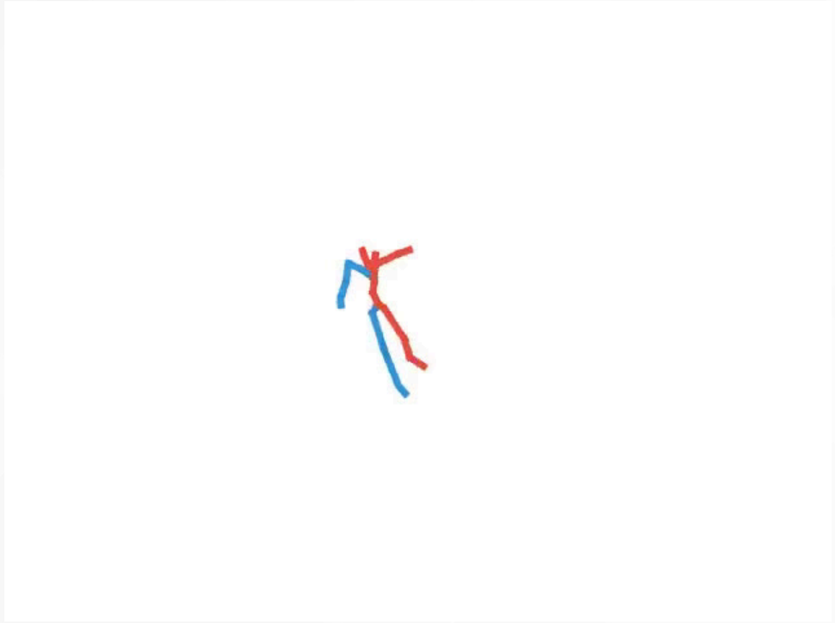
Spatial Features :
The graph convolution is employed directly on graphstructured data to extract highly meaningful patterns and features in the space domain. To reduce the computuation cost, The Chebyshev polynomials approxiamtion is deployed.

Temporal Features :
As CNNs offer fast training and no dependency on the previous time steps, gated CNNs are used to extract the temporal features.The temporal convolutional layer employs 1D casual convolutions with gated linear units, which enable parallel and controlled training.

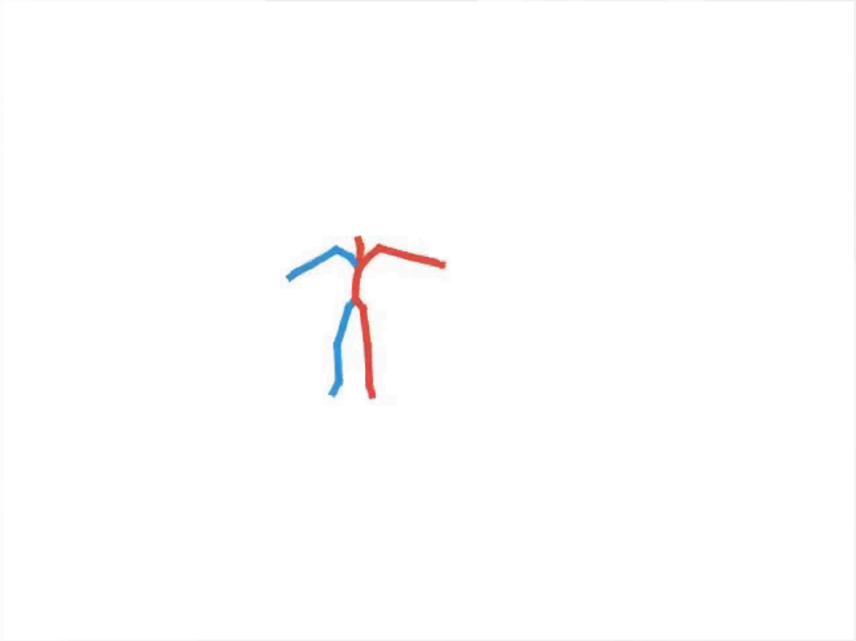


Results of current approach :

Provided seed motion



Model predicted motion



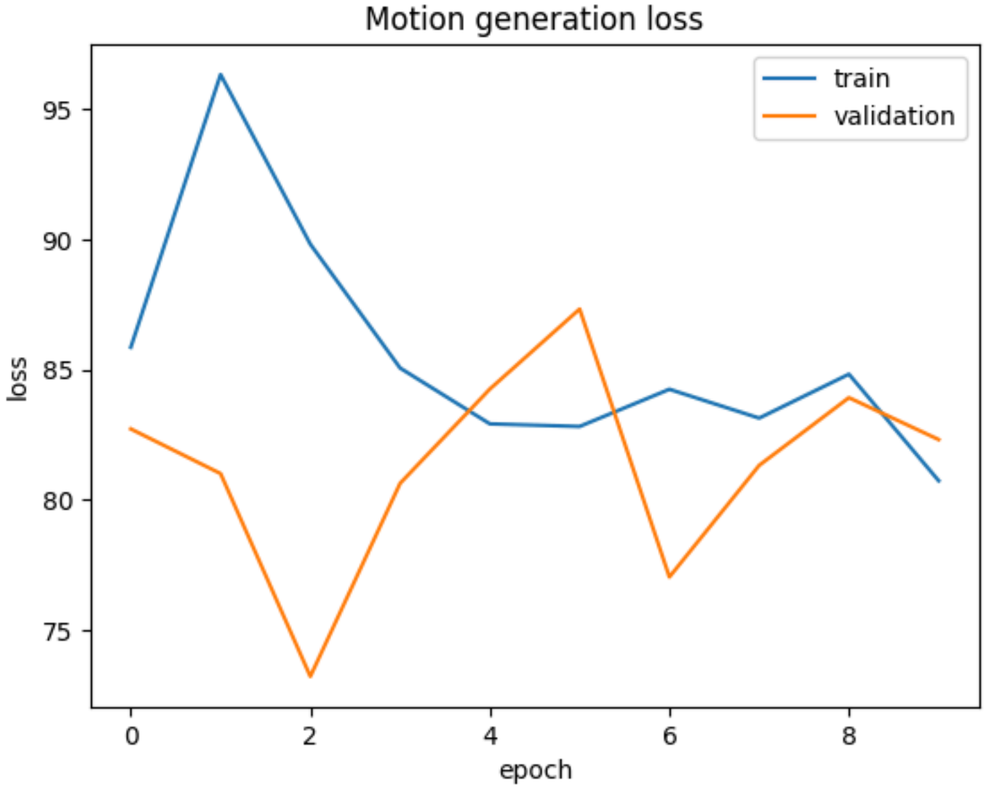
From the seed motion of 8 sec, our model managed to generate the dance sequence of the next 16 sec. Thus long form robust dance generation is successfully achieved by our current model.

Problems with current approach :

- Because of the use of 2 full attention transformers the training takes significant time. One epoch runs for almost 2 hours , so the approach is still computationally heavy.

- The model uses 240 frames of seed motion to make predictions of next 60 frames. So the model performance still depends on the seed motion that needs to be provided by the user.

- In our next approach we plan to reduce the number of transformers used and also eliminate the need for seed motion.



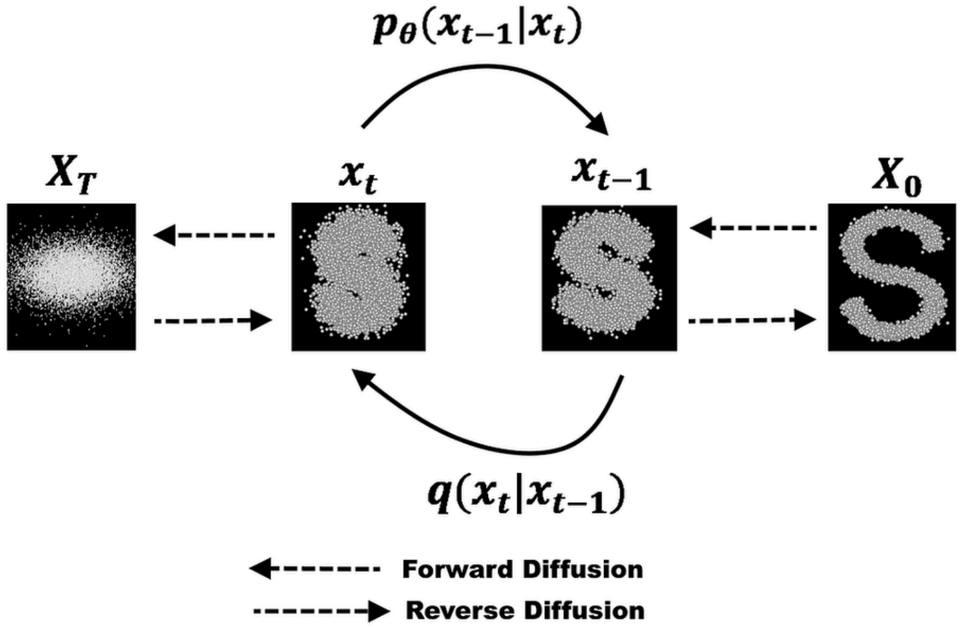
DIFFUSION MODEL APPROACH

- To overcome the problems mentioned before we plan to use to motion diffusion models to replace the two full attention transformers used in the current model.

- Diffusion Models are generative models which have been gaining significant popularity in the past several years, and for good reason. A handful of seminal papers released in the 2020s *alone* have shown the world what Diffusion models are capable of, such as beating GAN on image synthesis.

- Diffusion models work by destroying the training data through the successive addition of gaussian noise , and then learning to recover the data by reversing this noising process.

- Then we use this trained model to generate data by simply passing noise through learned denoising process.



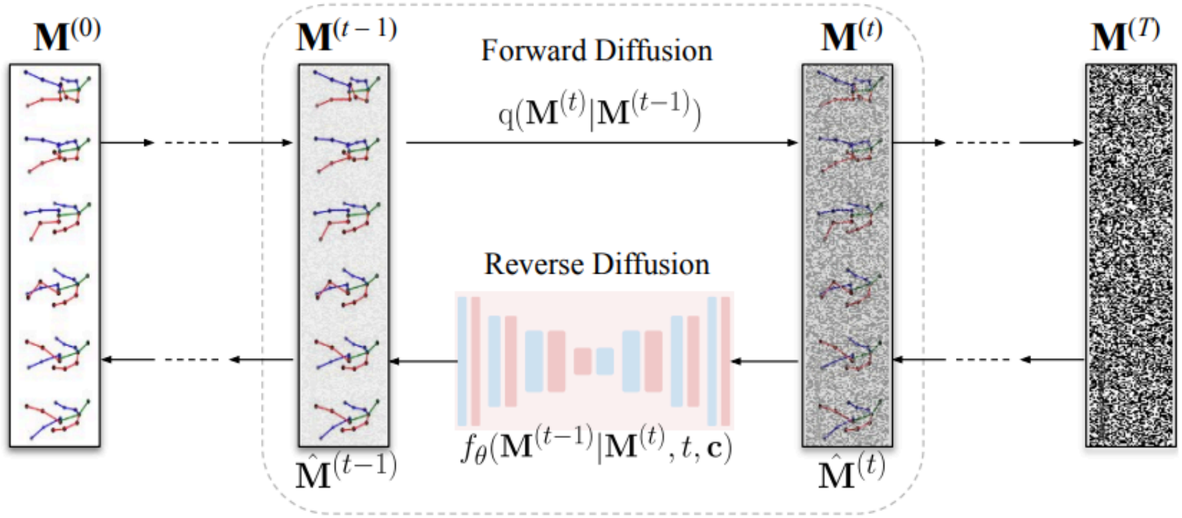
DIFFUSION MODEL APPROACH

- We plan on implementing a Diffusion Model for the 3D pose generation. We take our inspiration from the **“MoFusion: A Framework for Denoising-Diffusion-based Motion Synthesis ”** research paper.

- The motion generation task is formulated as a reverse diffusion process that requires sampling a random noise vector from a noise distribution to generate a meaningful motion sequence.

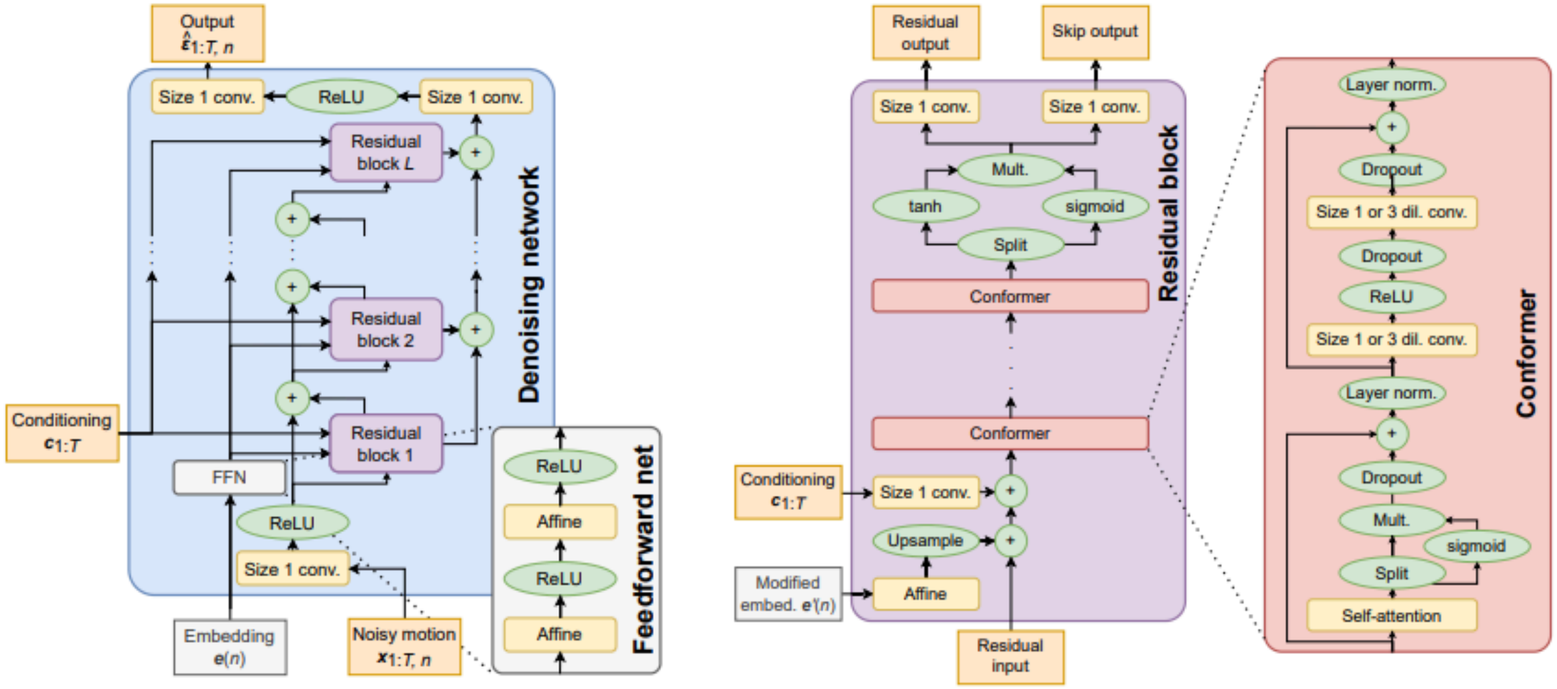
- Then the forward diffusion process requires successively corrupting motion sequence M(0) by adding Gaussian noise to a motion sequence for T timesteps in a Markovian fashion.

$$q\left(\mathbf{M}^{(1:T)}|\mathbf{M}^{(0)}\right) = \prod_{t=1}^{t=T} q\left(\mathbf{M}^{(t)}|\mathbf{M}^{(t-1)}\right)$$



- We plan to sample the poses in a similar way to the DiffWave implementation, which uses the residual blocks consisting of Bidirectional Dilated Convolution. Each residual block consists of many convolutions and fully connected layers with an input of the conditioned signal which gets correlated with the input poses to be trained upon.

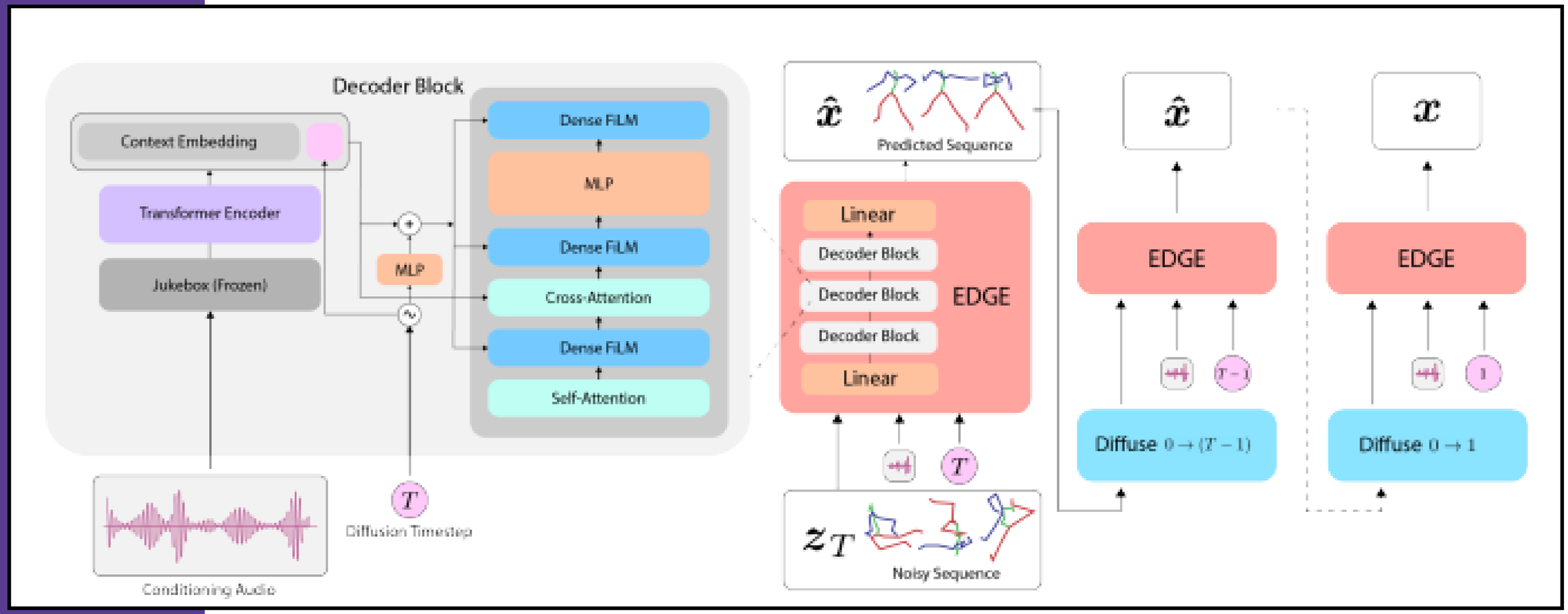
- Also instead of the simple Transformer, we are considering trying with the Conformer, which essentially is like a blend of the features of CNN and Transformers, composed of Self Attentions and Convolutions.



(a) Denoising network

(b) Residual block and Conformer

Editable Dance Generation from Music(EDGE)



EDGE uses a transformer-based diffusion model paired with Jukebox, a strong music feature extractor, and confers powerful editing capabilities well-suited to dance, including joint-wise conditioning, and in-betweening.

FUTURE WORK AND GOALS

- Implementing the diffusion model strategy described before and compare it with the performance of the currently deployed transformer based approach.

- Deploying the best performing model on a web server or API and launching this as easy to use product .One way is to use FASTAPI for building an interactive API.