# Paper Implementation:
# [CVPR 2023] MoFusion: A Framework for Denoising-Diffusion-based Motion Synthesis

(Group 32) 110550022 賴柏允, 110612117 張仲瑜

## -Goal

To implement a diffusion-based motion synthesis model that can generate a sequence of motion based on the given text prompt.

## -Related Works

Motion diffusion models from the same year.

[ICLR2023]

MDM: Human Motion Diffusion Model
*-A carefully adapted classifier-free diffusion-based generative model for the human motion domain.*

[ICCV 2023]

PhysDiff: Physics-Guided Human Motion Diffusion Model
*-A novel physics-guided motion diffusion model which incorporates physical constraints into the diffusion process.*

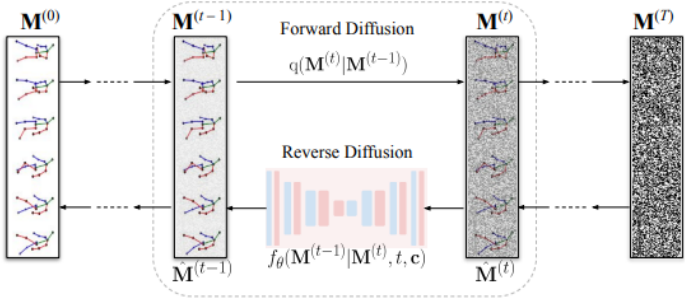
[SIGGRAPH 2023]

Listen, denoise, action! Audio-driven motion synthesis with diffusion models
*-Denoising diffusion probabilistic models for audio-driven motion synthesis, trained on motion-capture datasets of dance and gesticulation.*

## -Proposed Approach

The paper introduces a denoising-diffusion model (DDM) for human motion synthesis.
It uses a time-varying weight schedule to produce outputs that are temporally plausible and semantically accurate with the conditioning signal.

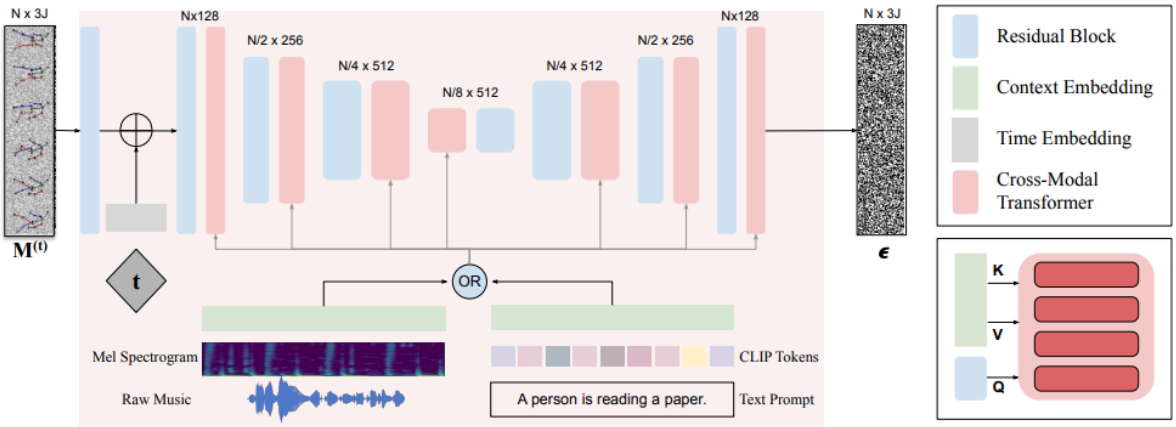
The authors formulate the motion generation task as a reverse diffusion process of sampling a random noise vector from a noise distribution to produce a meaningful motion sequence.




During training, Gaussian noise is successively introduced to a motion sequence with the Markov diffusion kernel, transforming the data to a noise matrix.

The overall loss for the training process is a weighted sum of two broad loss types, where the weights are determined by the timestep the generated motion sequence is on to avoid training on extremely noisy data. The first loss is the commonly used L2 distance between the noise used for forward diffusion and the estimated noise function by the model. While this loss alone is sufficient to approximate the underlying data distribution, the sequence may not be physically and anatomically plausible, so the second loss is introduced, which is a compound-loss comprising 3 loss terms: A skeleton-consistency loss that ensures consistent skeleton bone length across time, an anatomical constraint that penalize asymmetry of bone lengths, and finally a ground-truth supervision on motion synthesis. And since the denoising network is trained to estimate the noise instead of the motion sequence itself, the workaround is to apply the losses to the final reverse-diffused motion using the re-parameterisation trick.

To generate a motion sequence from the noise matrix, the model iteratively reverse-diffuses the noisy sequence with a model trained to predict the original noise. The architecture uses a 1-D UNet to approximate the noise function, allowing for training on motions of varying lengths.



The network consists of three downsampling blocks that first successively reduce the feature length, and each 1D residual block is followed by a cross-modal transformer that incorporates the conditioning context into the network. The time-embedding is generated by passing the sinusoidal time embedding through a two-layer MLP. The context is incorporated by treating intermediate residual motion features to get the query vector while using the conditional signal to compute the key and value vector. As in standard cross-attention, the relevance scores are first computed with the softmax, and then used to weigh the value.

## -Expected Result

Things that-

Are guaranteed (We'll try our best!) to work:

- Code of building the Unconditional generation
- Code of text to motion synthesis
- Plan B: We will make a blender animation if things above don't work out the way we hope.

May (might not) work :

- Audio to motion synthesis
- Same performance as the thesis (due to poor hardware, RTX 3060 Ti vs RTX A40)

## -Tasks of Each Member

- 110550022 賴柏允-- Code, Paper proposal, Report
- 110612117 張仲瑜-- Code, PPT proposal, Report