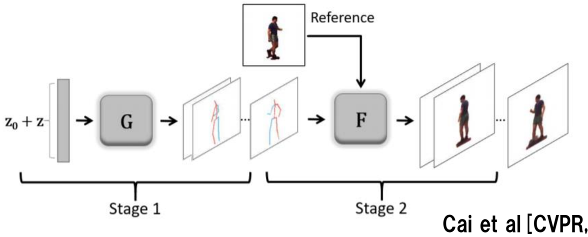
Pose Sequence Generation with a GCN and an initial Pose Generator

The University of Electro-Communications Kento Terauchi, Keiji Yanai

Background

Research Objective: natural pose sequence generation



Cai et al [CVPR, 2017]

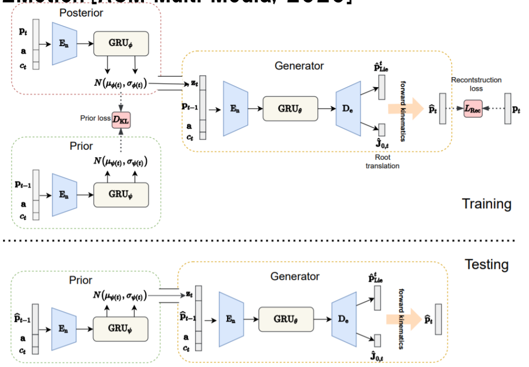
Some two-step video generation consisting of generation of pose sequences and video generation from pose sequences.

↓

By making pose generation natural, it is possible to generate human motion videos that perform natural motions.

Related Works

Action2Motion [ACM Multi Media, 2020]



They introduces Prior Loss which brings the distributions of the outputs of the encoder of the previous frame and the encoder of the next frame closer together.

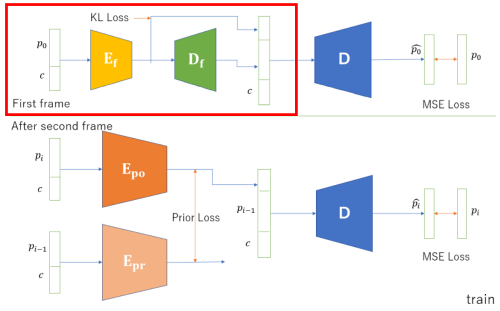
During testing, the encoder of the previous frame produces an output similar to the encoder of the next frame.

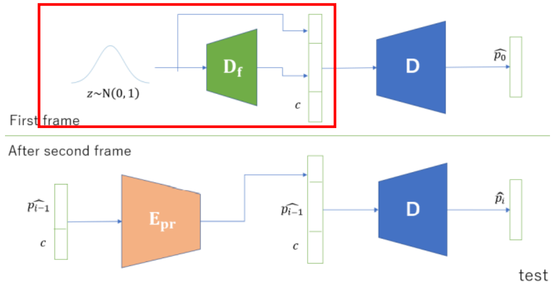
From the pose of the previous frame, the pose of the next frame can be generated sequentially.

Method

Make two changes to Action2Motion

Method: change (1) Add module for initial frame consideration





We encode initial frame encoder to encode initial frame and initial condition generator to the model of Action2Motion.

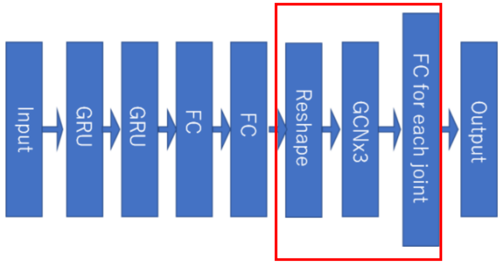
We encode the first pose with the initial frame encoder and decode it to condition of decoder with the initial condition decoder.

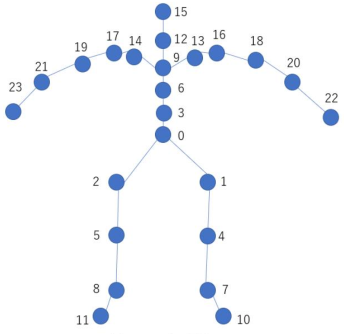
After the second frame, Learn to bring the outputs of the previous frame encoder and the next frame decoder closer to each other.

At the time of generation, the first frame is a condition generated from the noise of the normal distribution using the initial frame condition decoder.

After the second frame, the next frame is sequentially generated by encoding the previous frame with the previous frame encoder.

Method: Change (2) introduce GCN to the decoder





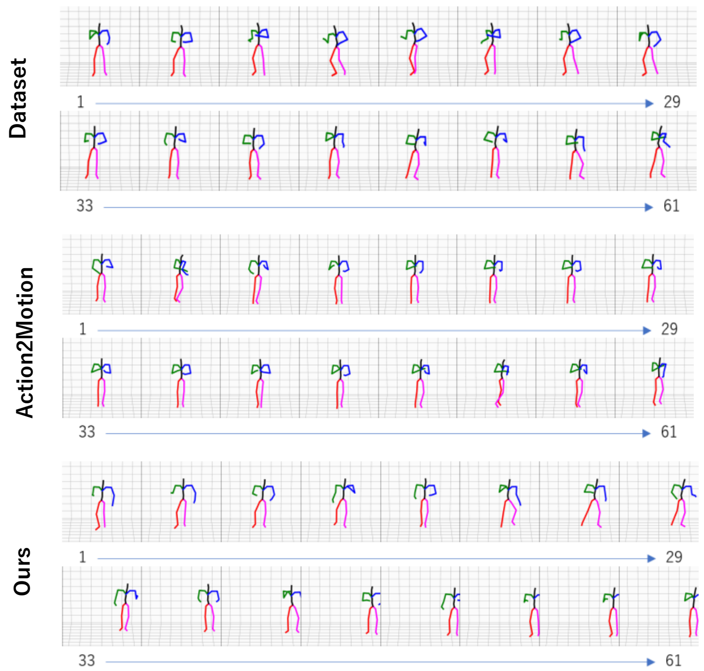
HumanAct12

It learn the structural representation by using three layers of GCN to get a complex representation.

The output is obtained by passing through a fully connected layer with different weights for each joint.

Experiments

Examples of generated results on Run action



Qualitative Evaluation

The proposed method, the results of which are represented as ˵Ours (FULL) " in the table, outperforms the existing methods in FID while maintaining the same accuracy.

Method	Accuracy↑	FID↓	Diversity→	Multimodality→
Groundtruth (GT)	0.997	0.092	6.857	2.449
Action2Motion	0.923	2.458	7.032	2.870
Ours(FULL)	**0.924**	2.252	**6.962**	**2.861**
Ours(IPG only)	0.864	**1.979**	6.924	3.388
Ours(GCN only)	0.542	13.599	5.933	3.309

Ours (IPG only) outperformed the baseline regarding FID.

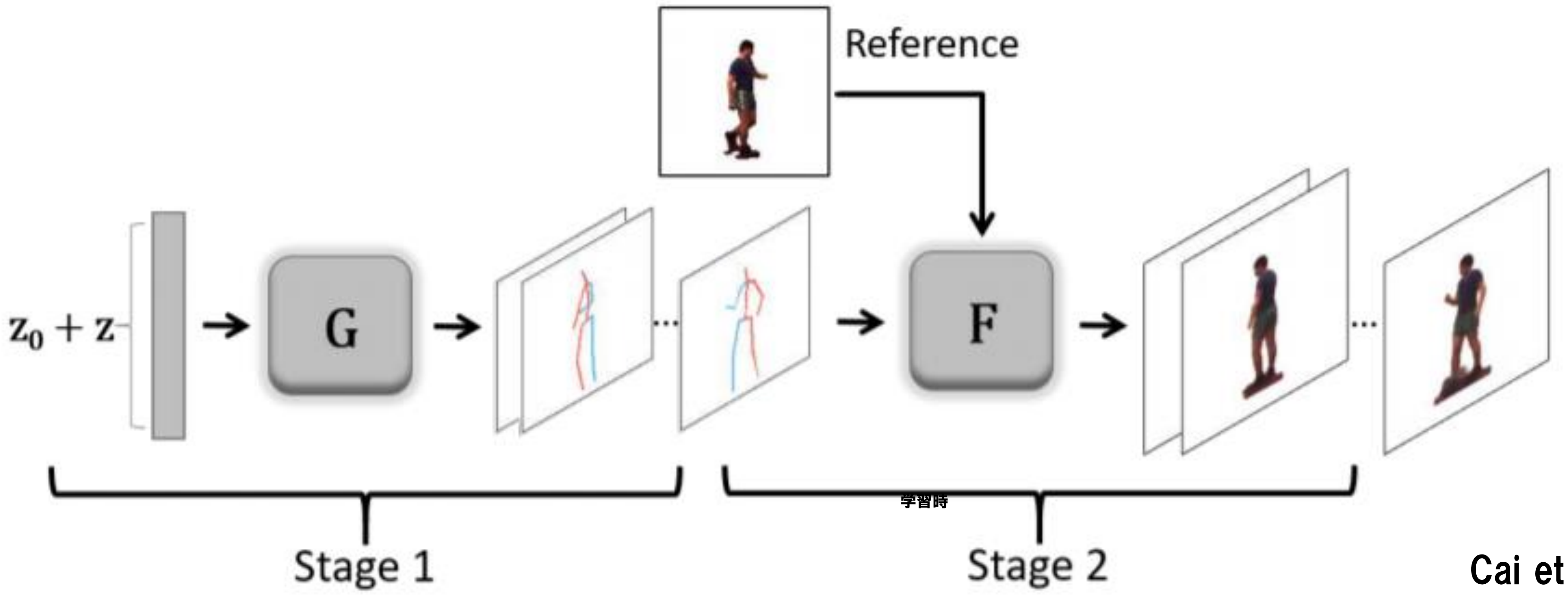
Its accuracy was inferior to Ours (FULL), and it could not generate a clear pose sequence.

Ours (GCN only) failed to generate a clear pose sequence because both Accuracy and FID are by far less than others.

Pose Sequence Generation with a GCN and an initial Pose Generator

Kento Terauchi, Keiji Yanai
The University of Electro-Communications

Research Objective: natural pose sequence generation



Cai et al [CVPR, 2017]

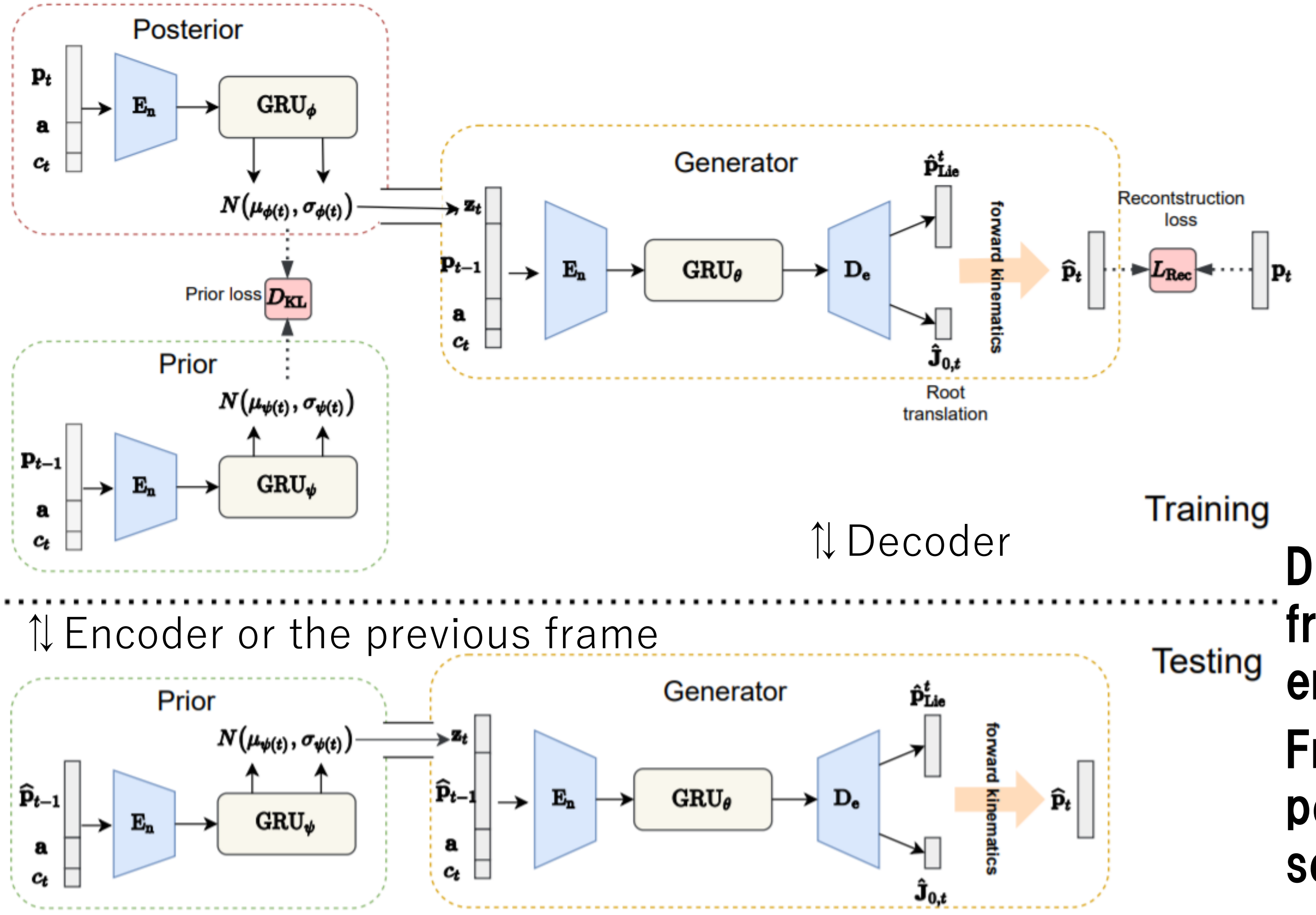
Some two-step video generation consisting of generation of pose sequences and video generation from pose sequences.

↓

By making pose generation natural, it is possible to generate human motion videos that perform natural motions.

Related Works: Action2Motion [ACM Multi Media, 2020]

↓ Encoder or the next frame



⇅ Decoder

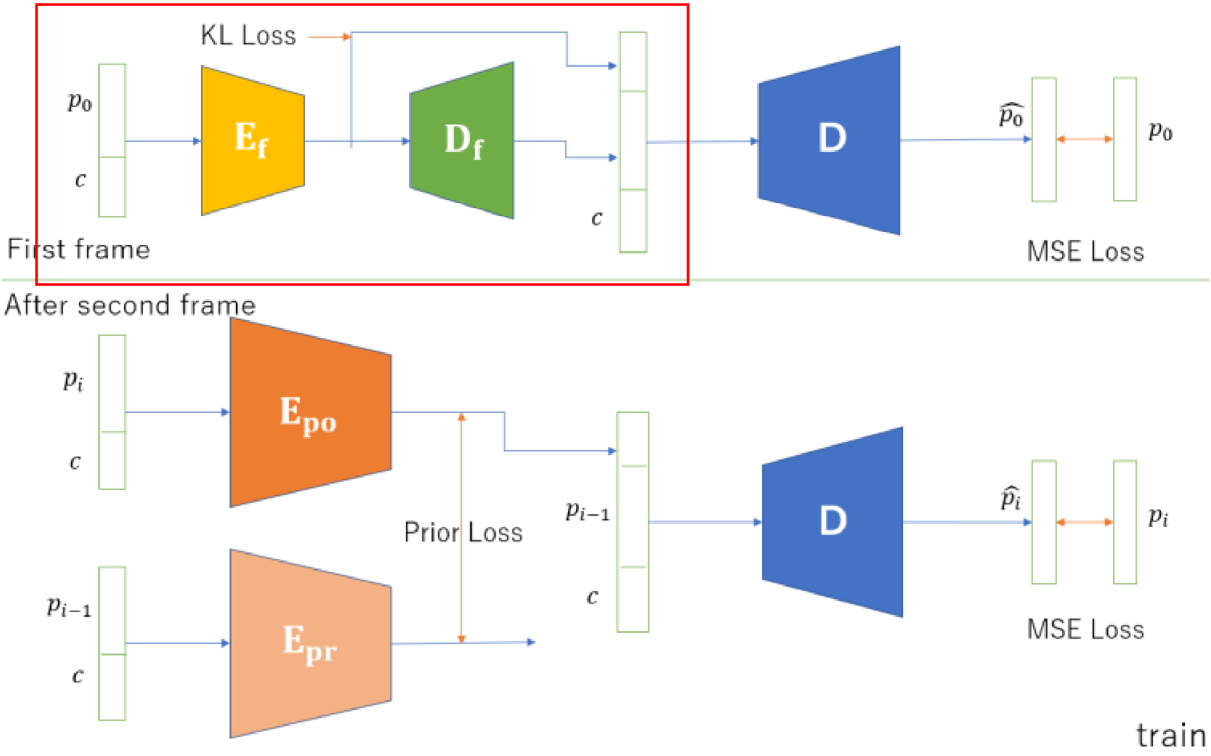
⇅ Encoder or the previous frame

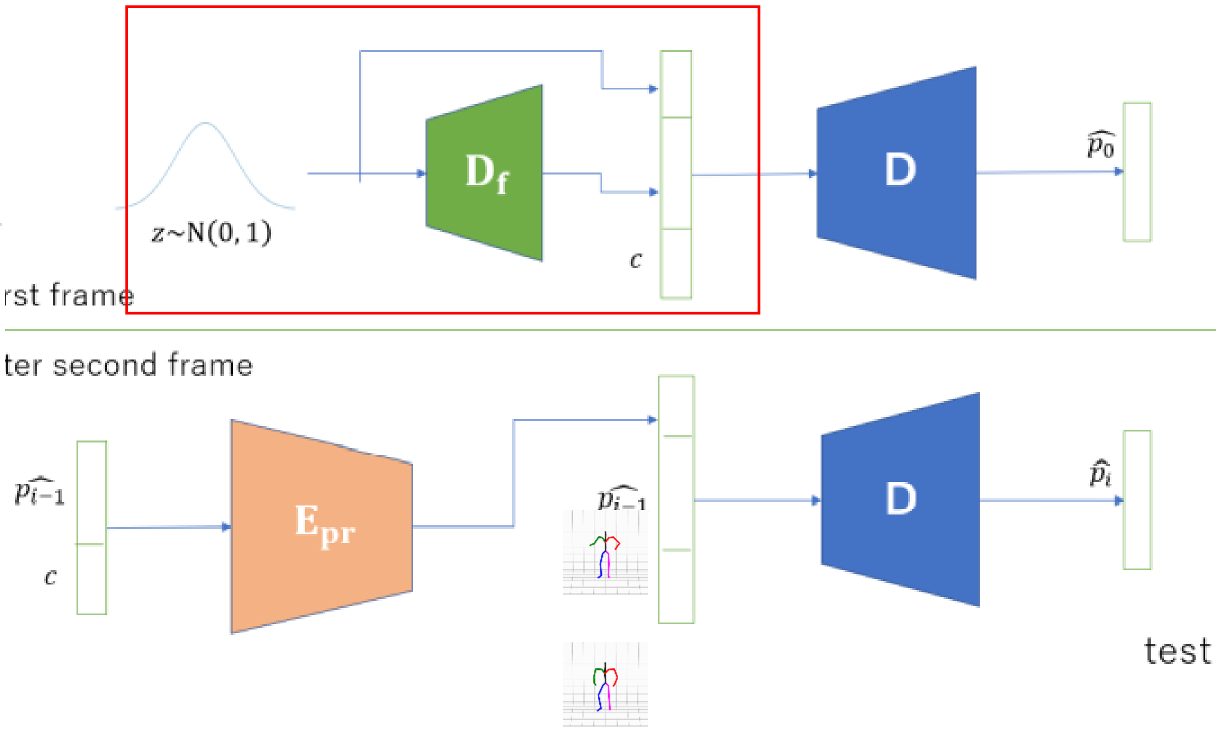
They introduces Prior Loss which brings the distributions of the outputs of the encoder of the previous frame and the encoder of the next frame closer together.

During testing, the encoder of the previous frame produces an output similar to the encoder of the next frame.

From the pose of the previous frame, the pose of the next frame can be generated sequentially.

Method: change（1） Add module for initial frame consideration





We encode initial frame encoder to encode initial frame and initial condition generator to the model of Action2Motion.

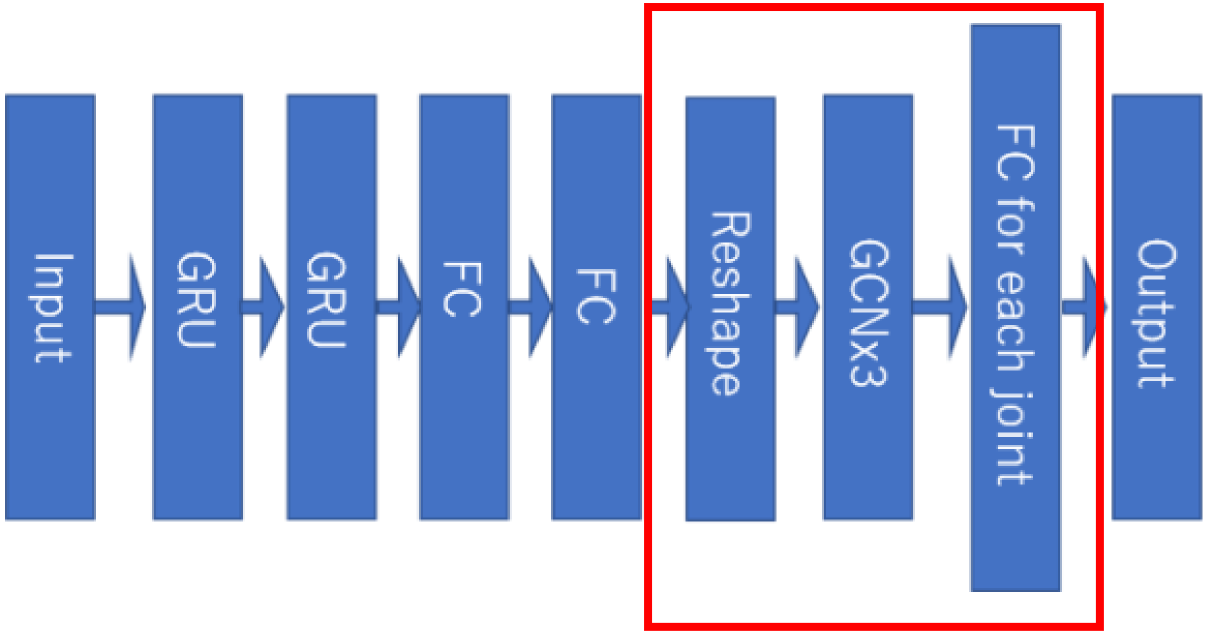
We encode the first pose with the initial frame encoder and decode it to condition of decoder with the initial condition decoder.

After the second frame, Learn to bring the outputs of the previous frame encoder and the next frame decoder closer to each other.

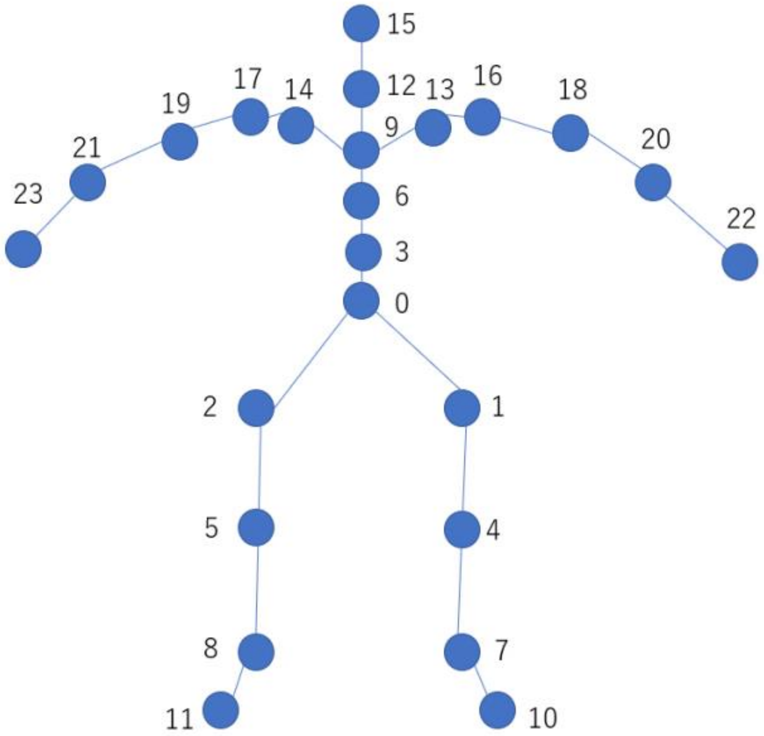
At the time of generation, the first frame is a condition generated from the noise of the normal distribution using the initial frame condition decoder.

After the second frame, the next frame is sequentially generated by encoding the previous frame with the previous frame encoder.

Method: Change (2) introduce GCN to the decoder



Input → GRU → GRU → FC → FC → Reshape → GCNx3 → FC for each joint → Output

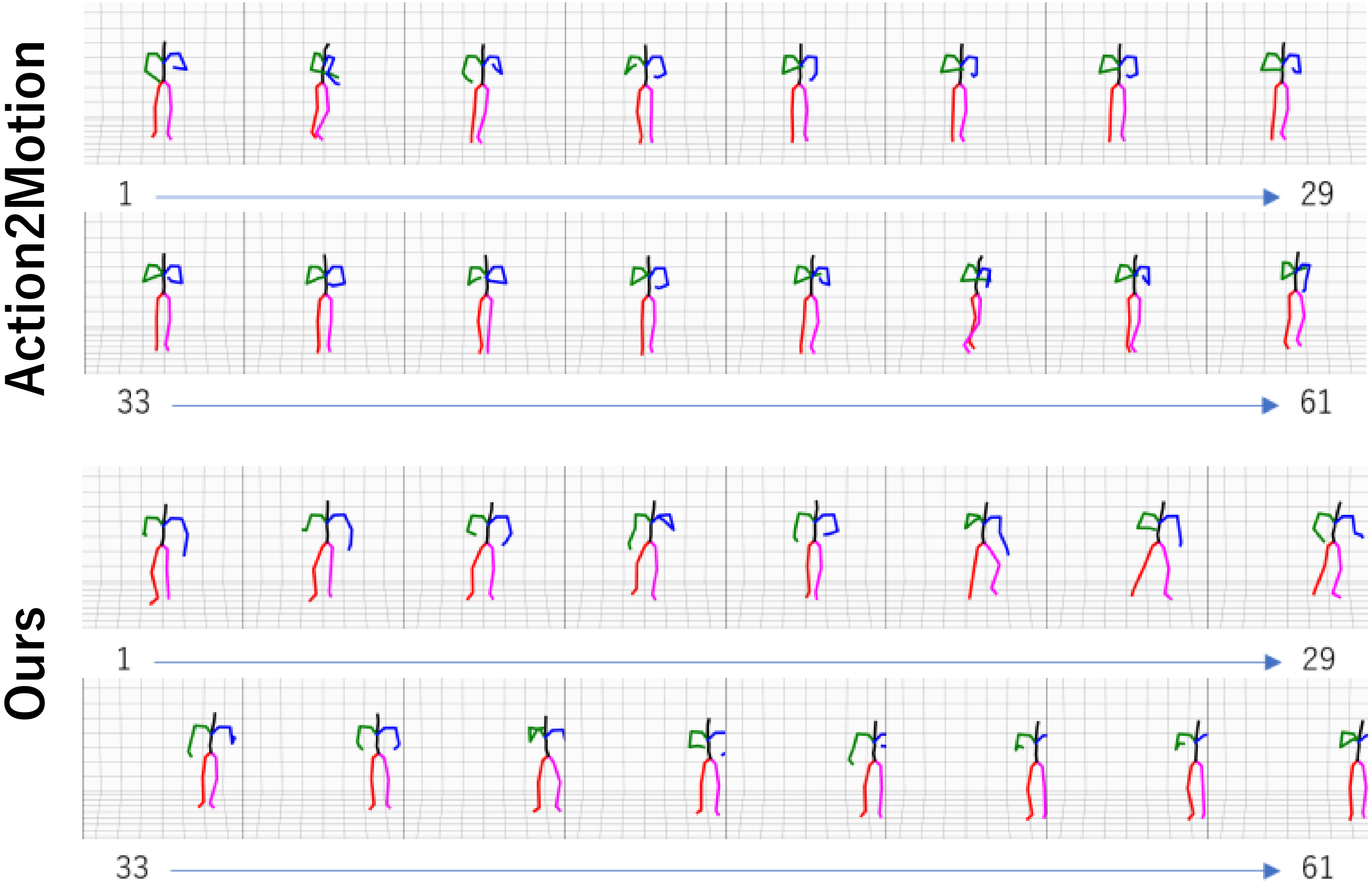


HumanAct12

It learn the structural representation by using three layers of GCN to get a complex representation.

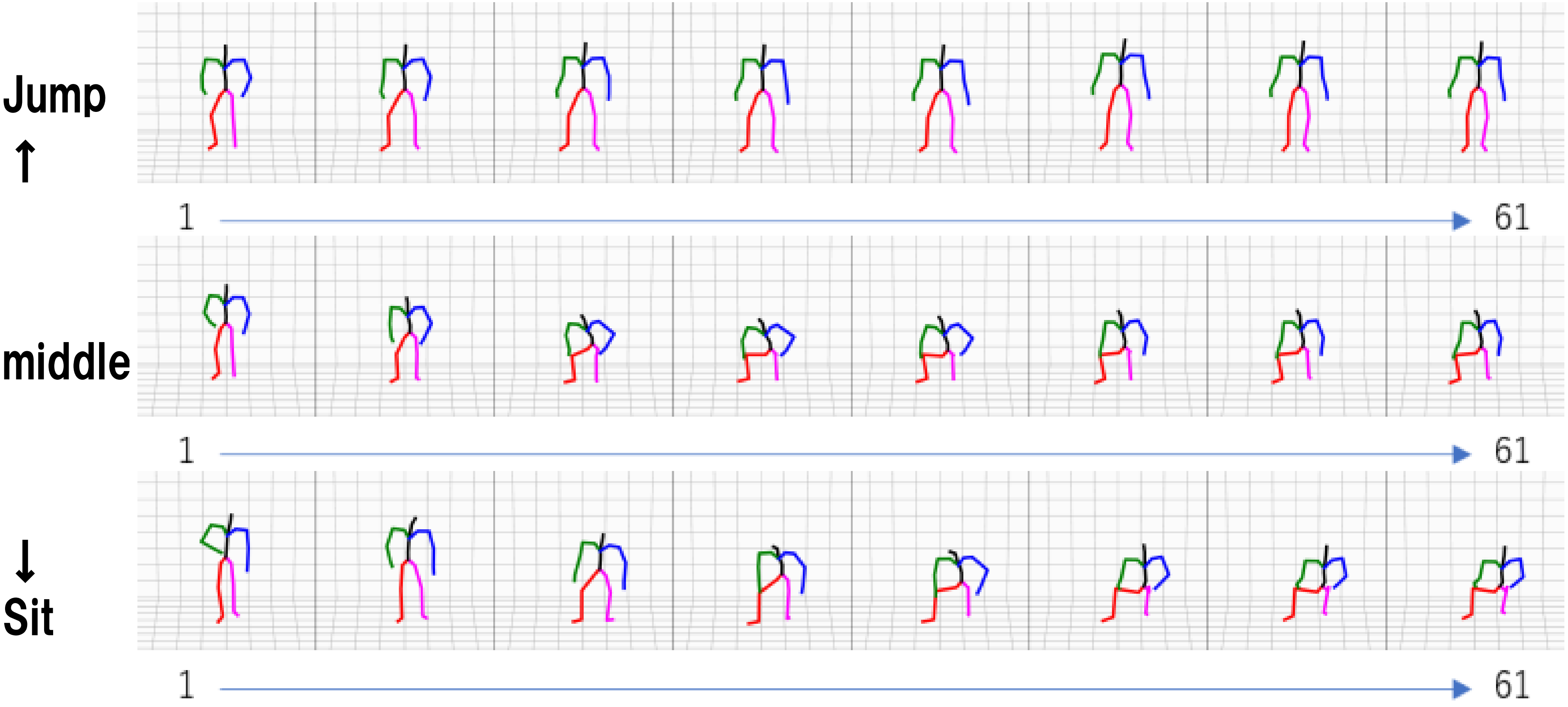
The output is obtained by passing through a fully connected layer with different weights for each joint.

Qualitative Evaluation: Examples of generated results on Run Action



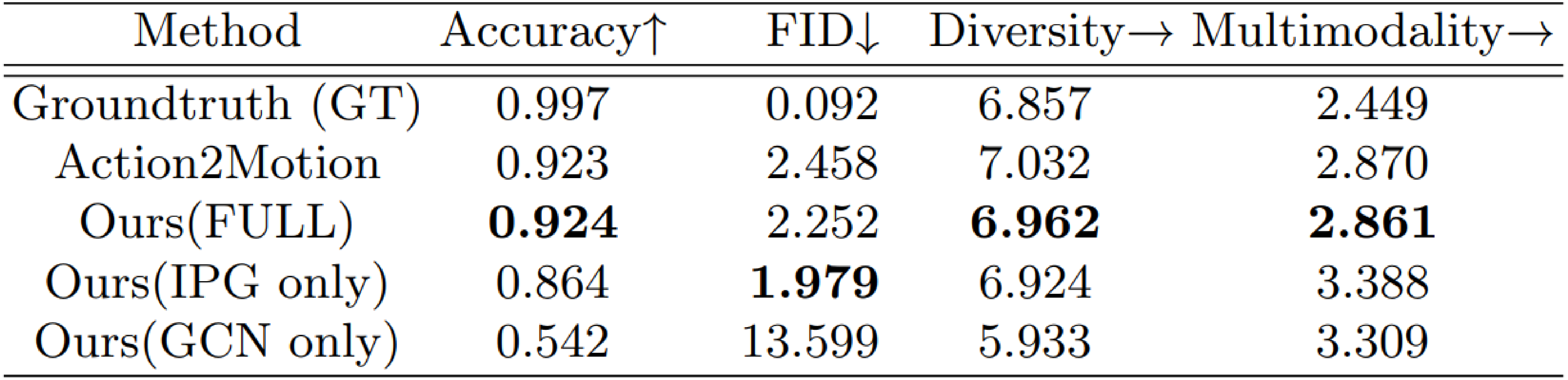
Interpolation of latent space

Interclass interpolation between jump and sit action



Qualitative Evaluation

The proposed method, the results of which are represented as ``Ours (FULL) " in the table, outperforms the existing methods in FID while maintaining the same accuracy.



Method	Accuracy↑	FID↓	Diversity→	Multimodality→
Groundtruth (GT)	0.997	0.092	6.857	2.449
Action2Motion	0.923	2.458	7.032	2.870
Ours(FULL)	**0.924**	2.252	**6.962**	**2.861**
Ours(IPG only)	0.864	**1.979**	6.924	3.388
Ours(GCN only)	0.542	13.599	5.933	3.309

Ours （IPG only） outperformed the baseline regarding FID.

Its accuracy was inferior to Ours (FULL) , and it could not generate a clear pose sequence.

Ours （GCN only） failed to generate a clear pose sequence because both Accuracy and FID are by far less than others.

Conclusion

- In this study, we proposed a model that explicitly captures the structural information by considering the initial frame with the intial pose generator and incorporating GCN.

- The qualitative results showed that the proposed method was capable of generating complex and diverse motions.

- Quantitatively, the proposed method outperformed the existing methods and showed more natural generation.