Let's Play Music: Audio-driven Performance Video Generation

Hao Zhu, Yi Li, Feixia Zhu, Aihua Zheng, and Ran He

School of Computer Science and Technology, Anhui University

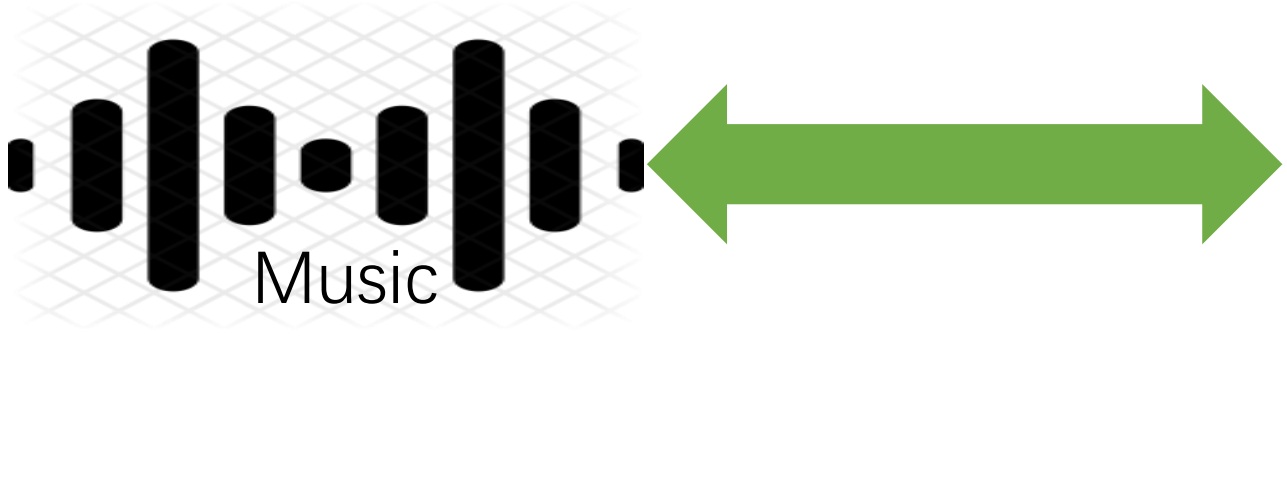
NLPR & CEBSIT & CRIPAC, Institute of Automation, CAS

School of Artificial Intelligence, University of Chinese Academy of Sciences

ICPR 2020

Task Definition

- There is a clear relationship between a piece of music and the movements performed.

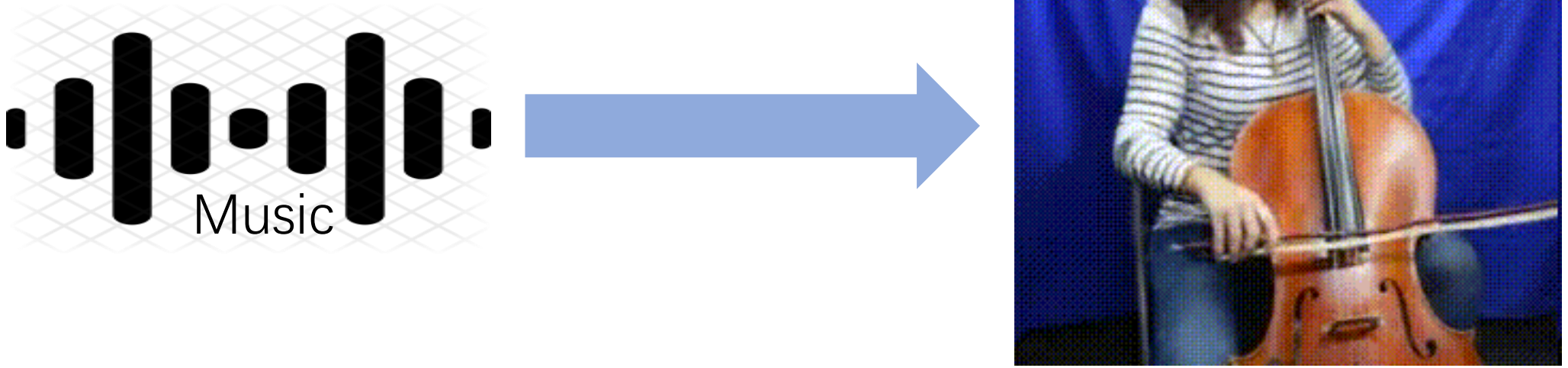


Task Definition

- How to construct the relationship between music and corresponding performance movements?

Task Definition

- How to construct the relationship between music and corresponding performance movements?
- **New Task: Audio-driven Performance Video Generation (APVG)**



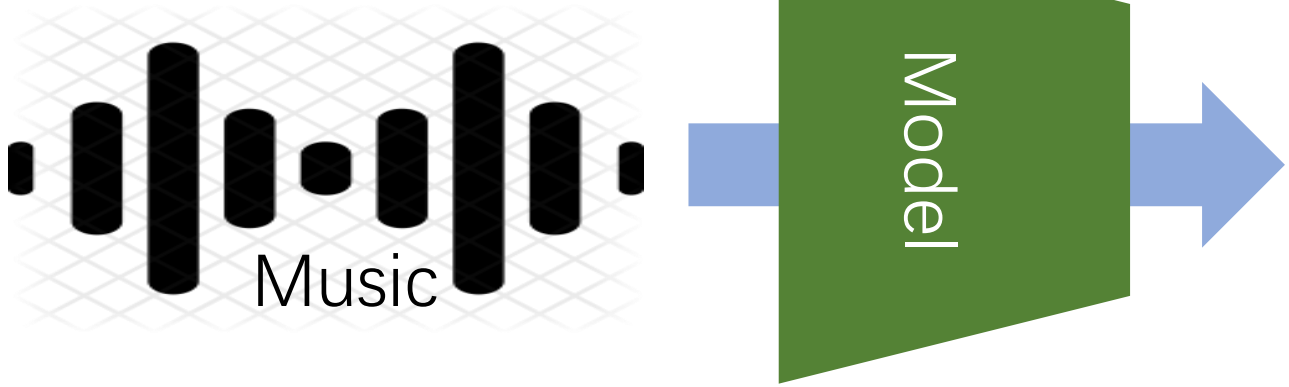
Task Definition

- How to construct the relationship between music and corresponding performance movements?
- New Task: Audio-driven Performance Video Generation (APVG)
- **Applications: Concert video generation, instrumental teaching**

Motivation

- Challenges:

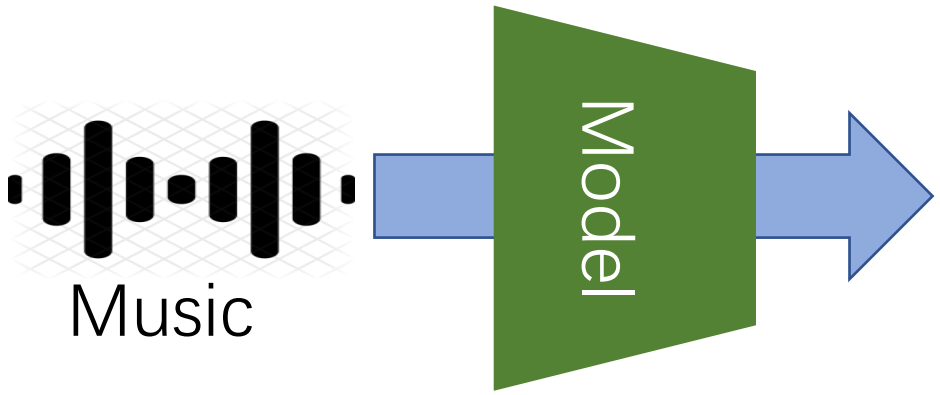
Generate **precise motion details** such as body and fingers from the **low-dimensional audio information**.



Motivation

Solutions:

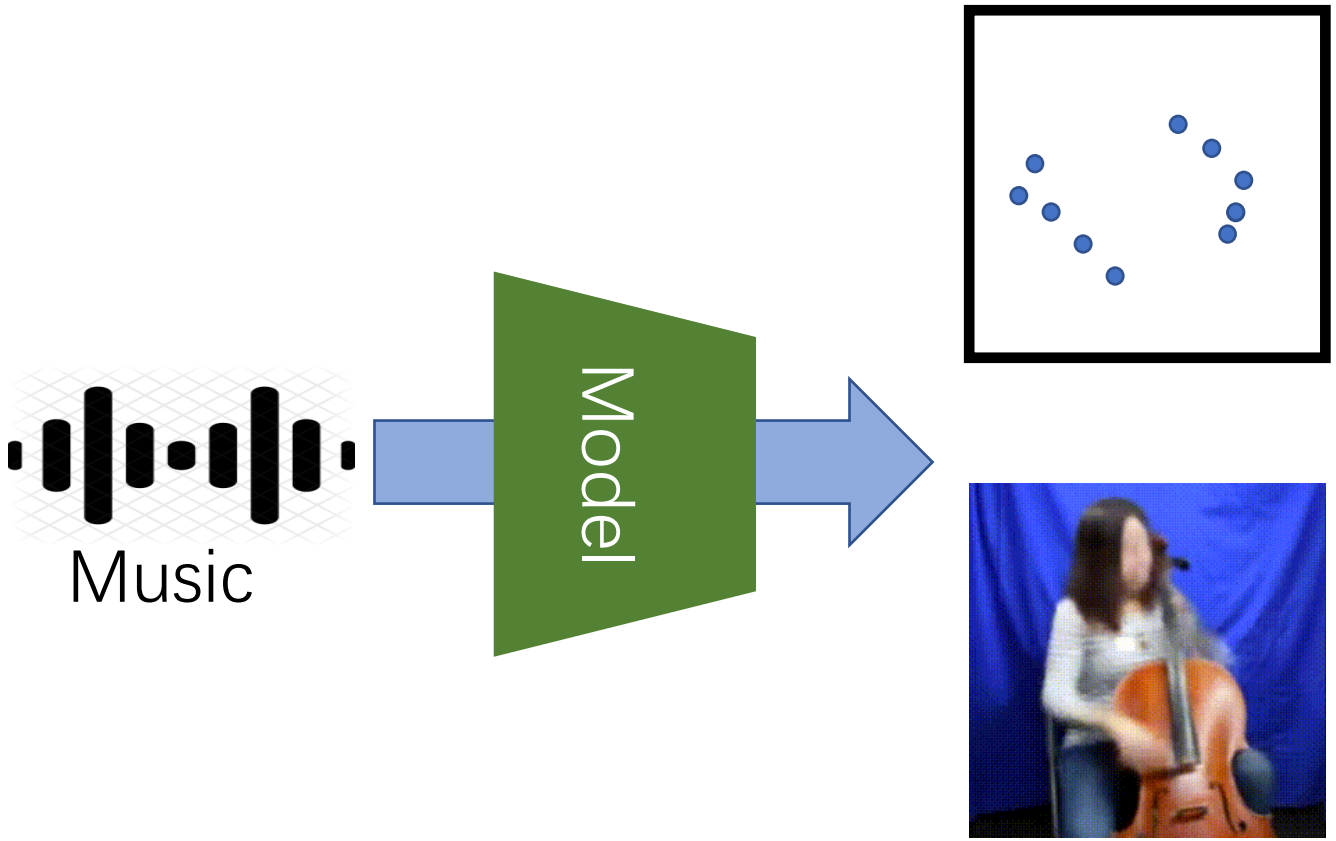
- **Multi-stage generation**
- Structure and temporal information



Motivation

Solutions:

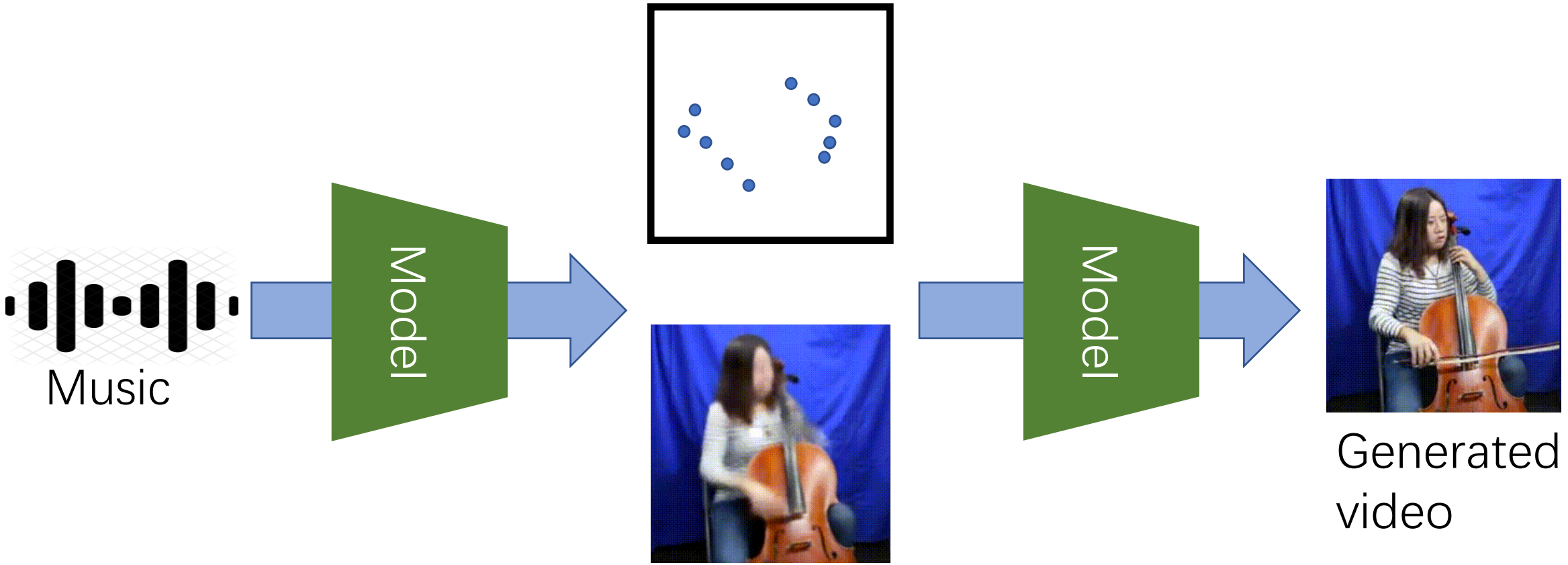
- **Multi-stage generation**
- Structure and temporal information



Motivation

Solutions:

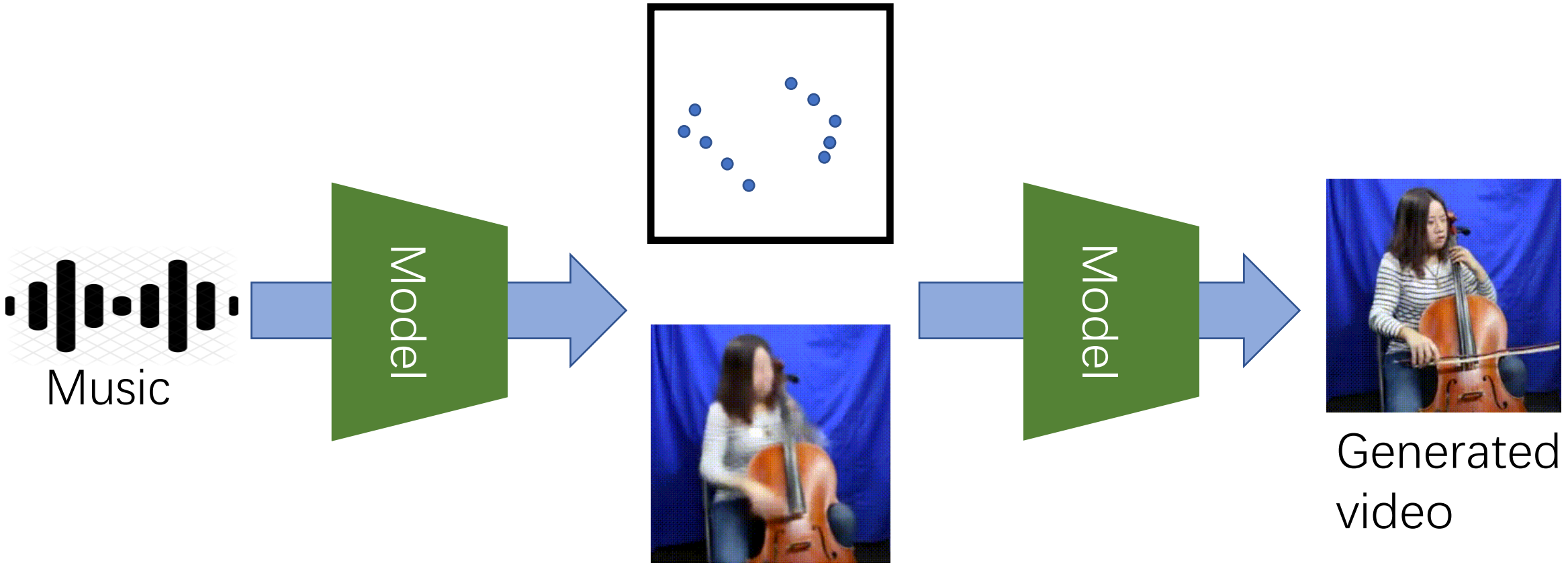
- **Multi-stage generation**
- Structure and temporal information



Motivation

Solutions:

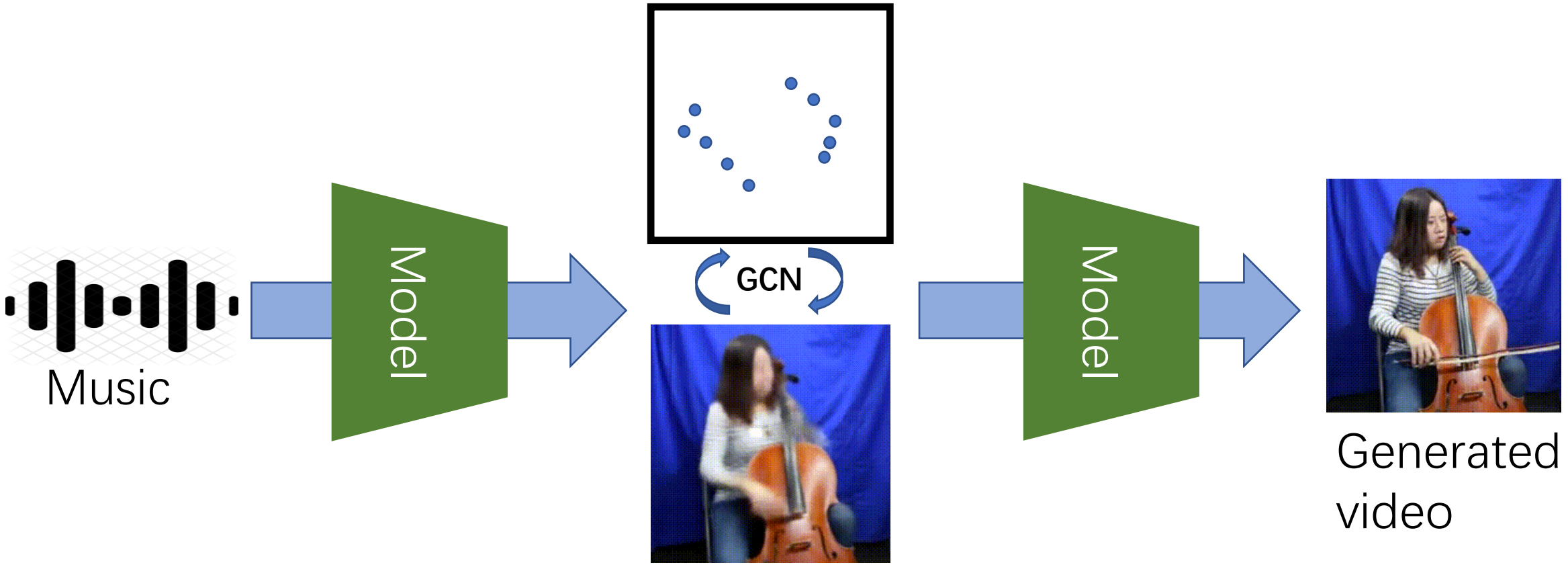
- Multi-stage generation
- **Structure and temporal information**



Motivation

Solutions:

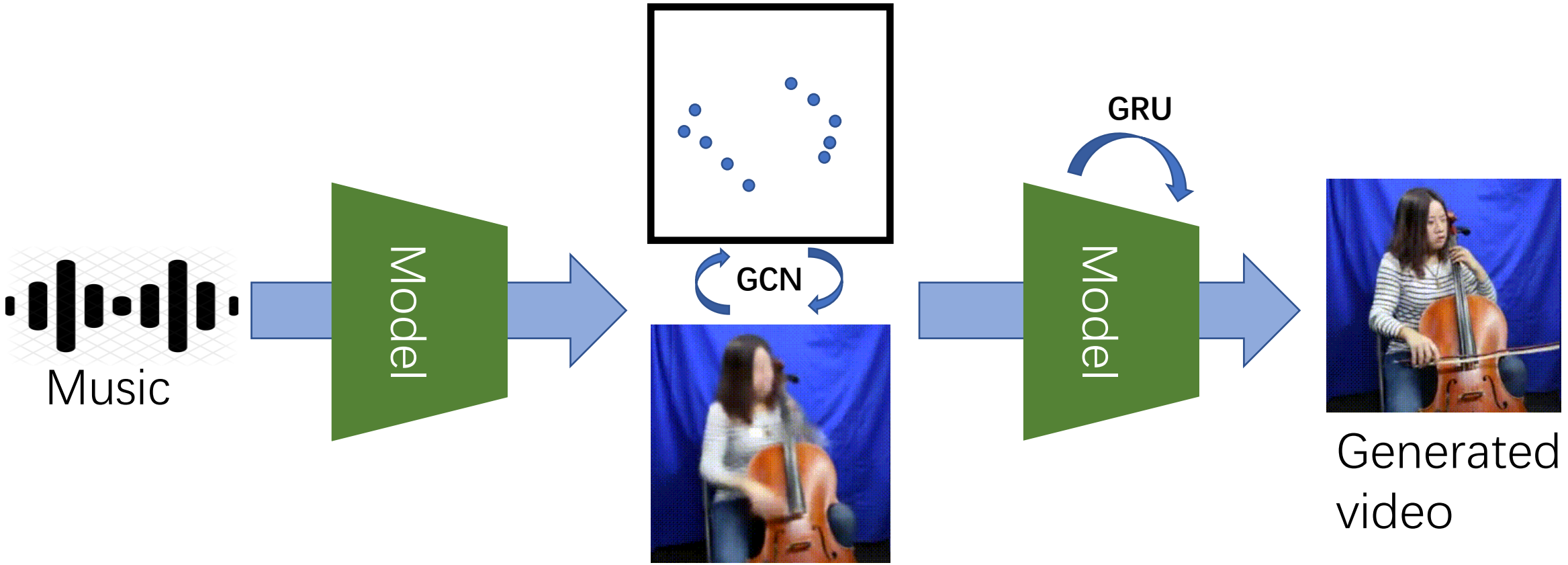
- Multi-stage generation
- **Structure and temporal information**



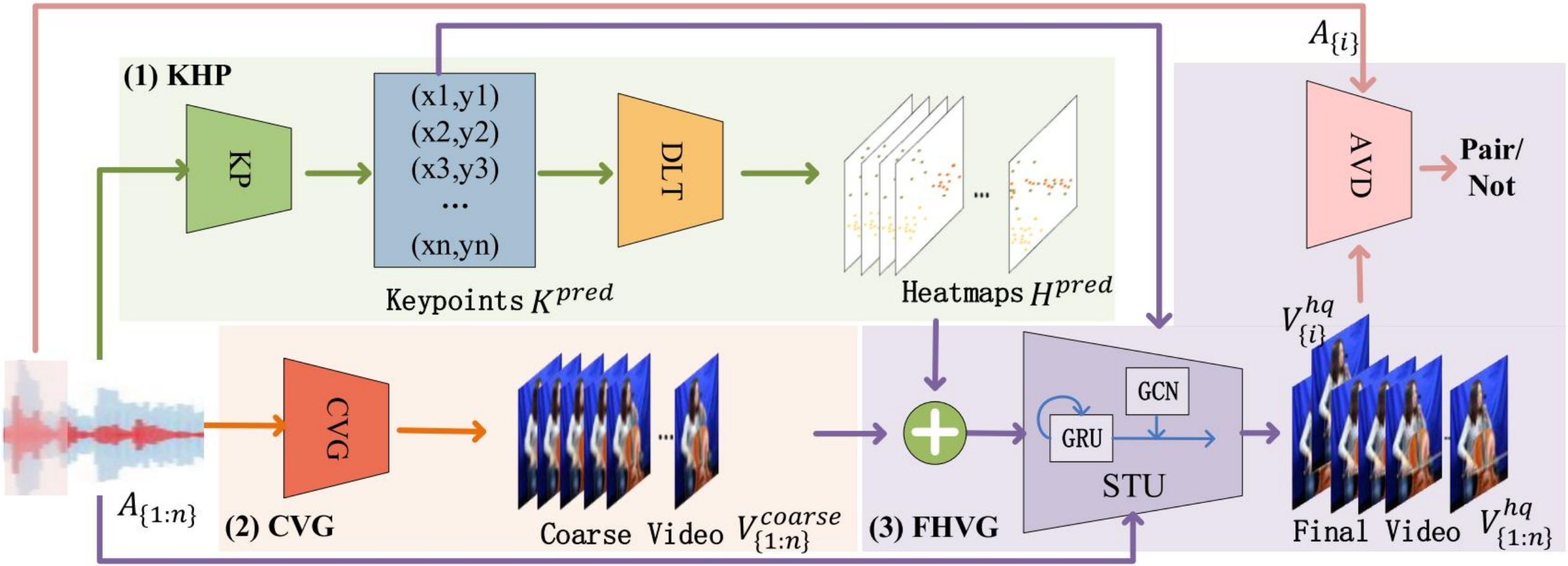
Motivation

Solutions:

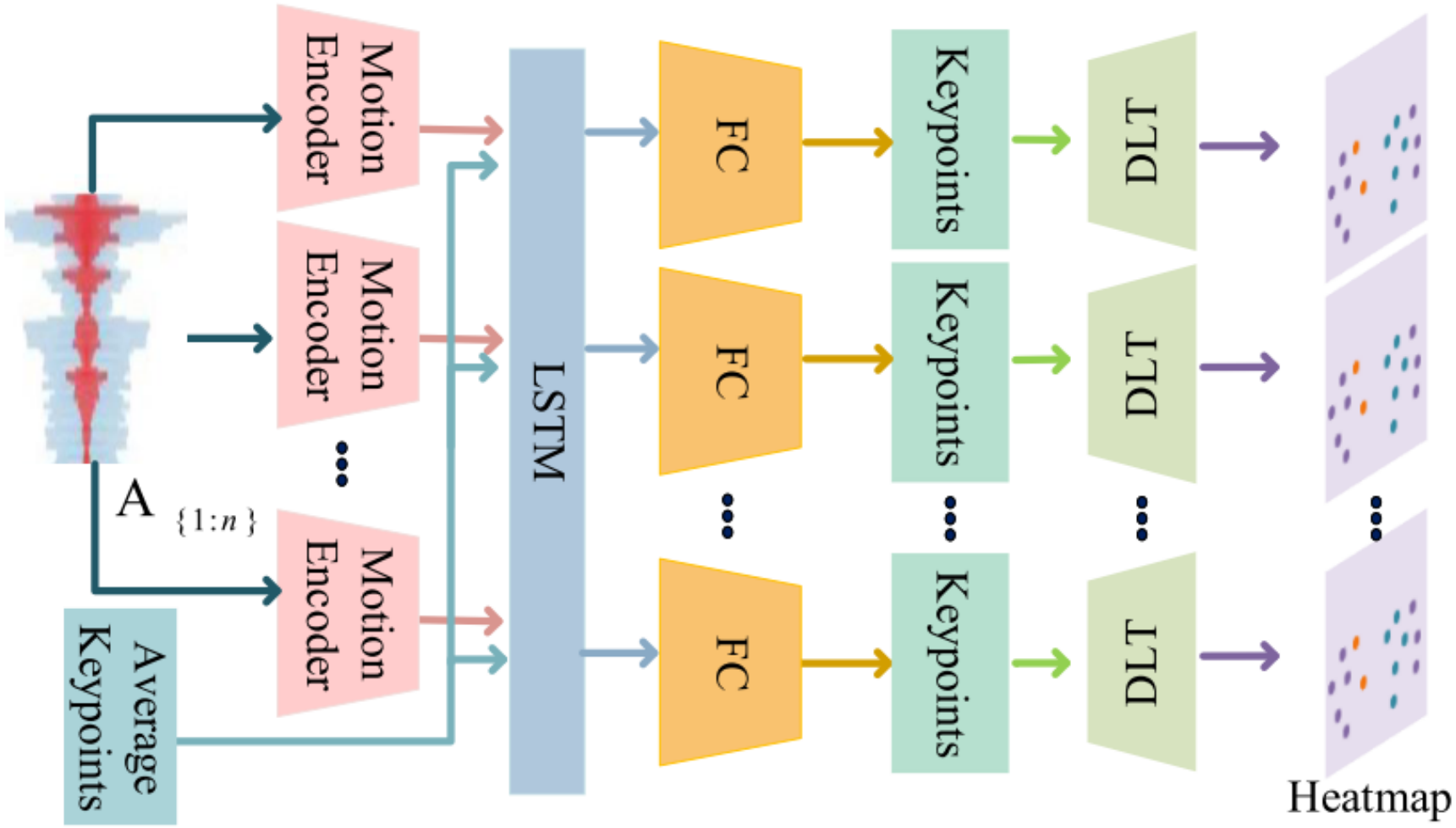
- Multi-stage generation
- **Structure and temporal information**



Approach

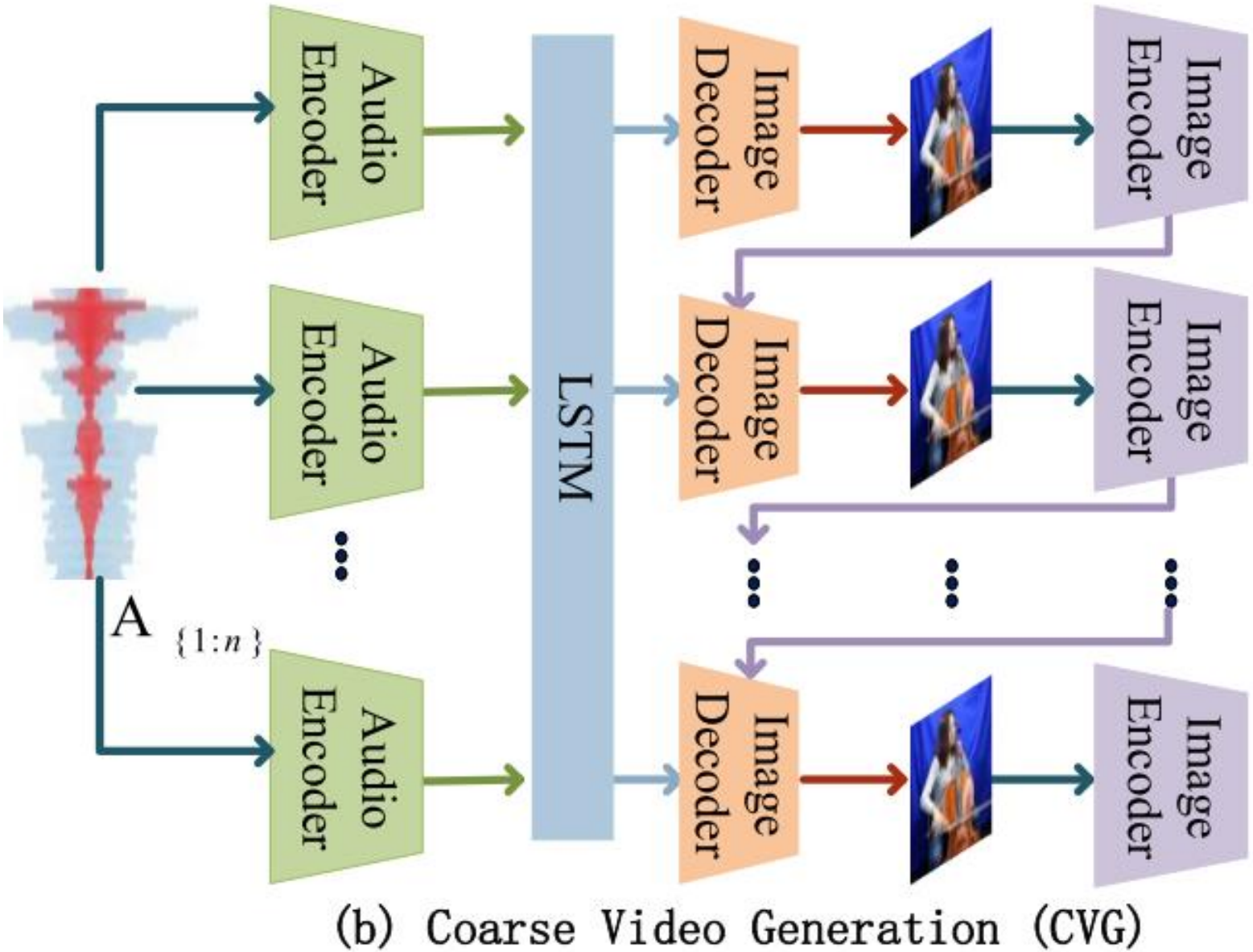

(1) KHP
KP
(x1,y1)
(x2,y2)
(x3,y3)
...
(xn,yn)
DLT
Keypoints K^{pred}
Heatmaps H^{pred}
$A_{\{i\}}$
AVD
Pair/
Not
$V^{hq}_{\{i\}}$
$A_{\{1:n\}}$
CVG
(2) CVG
Coarse Video $V^{coarse}_{\{1:n\}}$
(3) FHVG
GCN
GRU
STU
Final Video $V^{hq}_{\{1:n\}}$

Approach

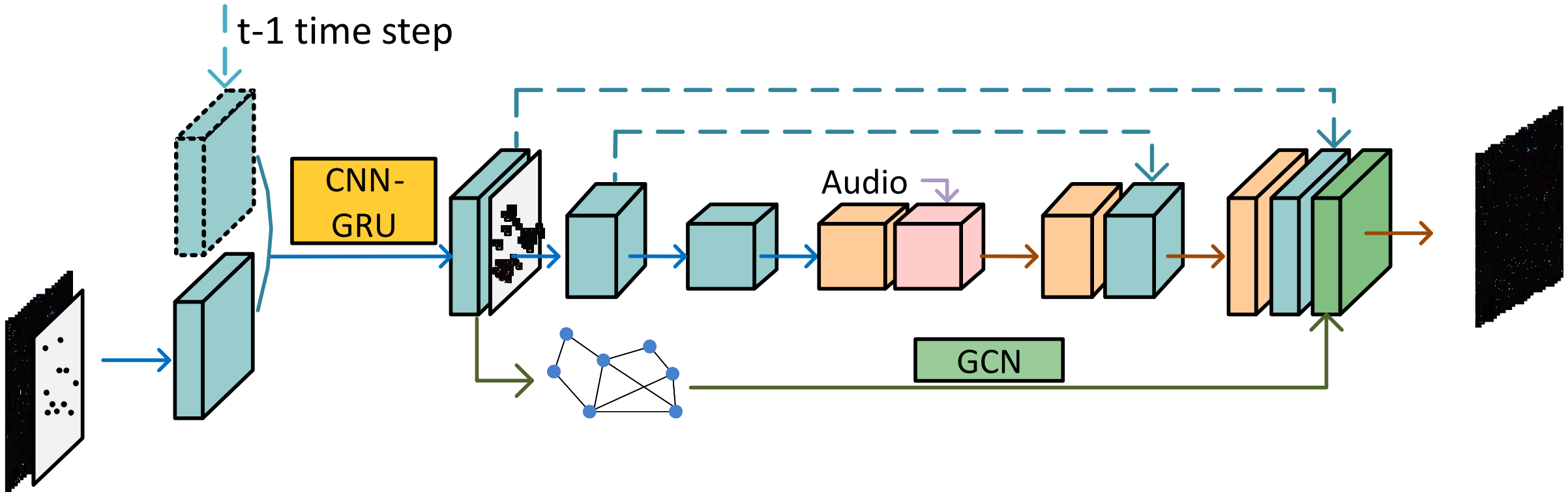


(a) Keypoint and Heatmap Prediction (KHP)

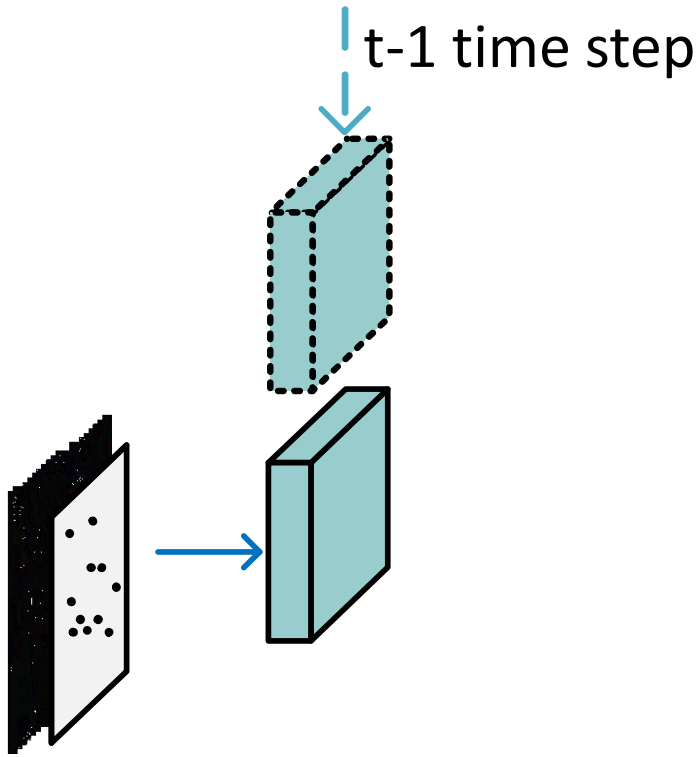
Approach


Audio Encoder
Audio Encoder
Audio Encoder
A {1:n}
LSTM
Image Decoder
Image Decoder
Image Decoder
Image Encoder
Image Encoder
Image Encoder
(b) Coarse Video Generation (CVG)

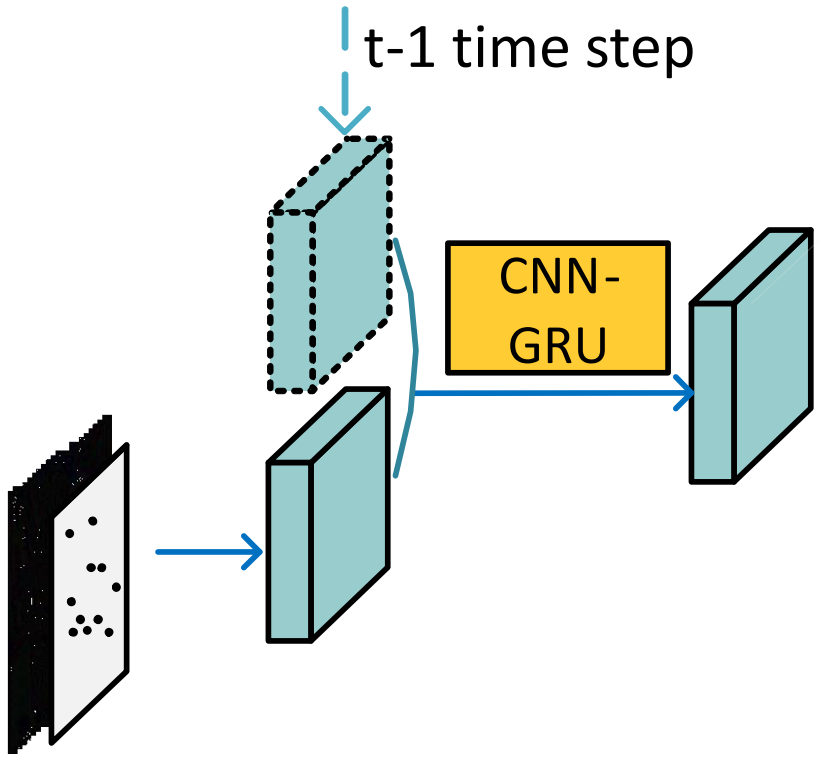
Approach



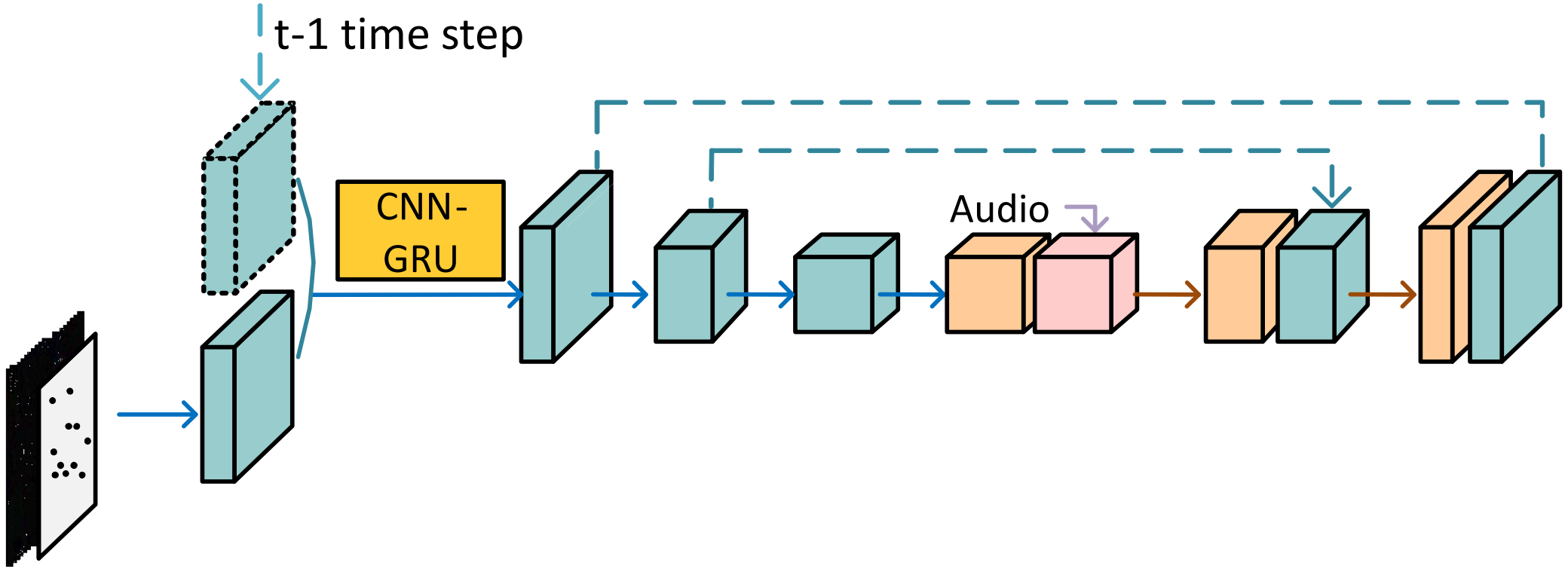
Approach



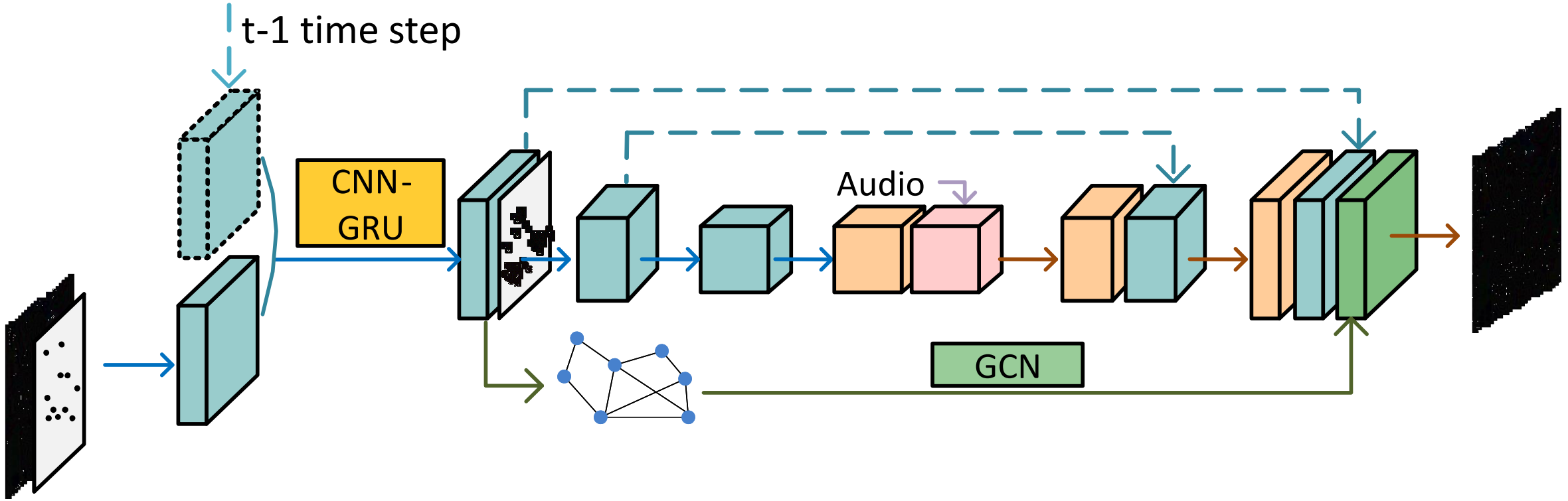
Approach



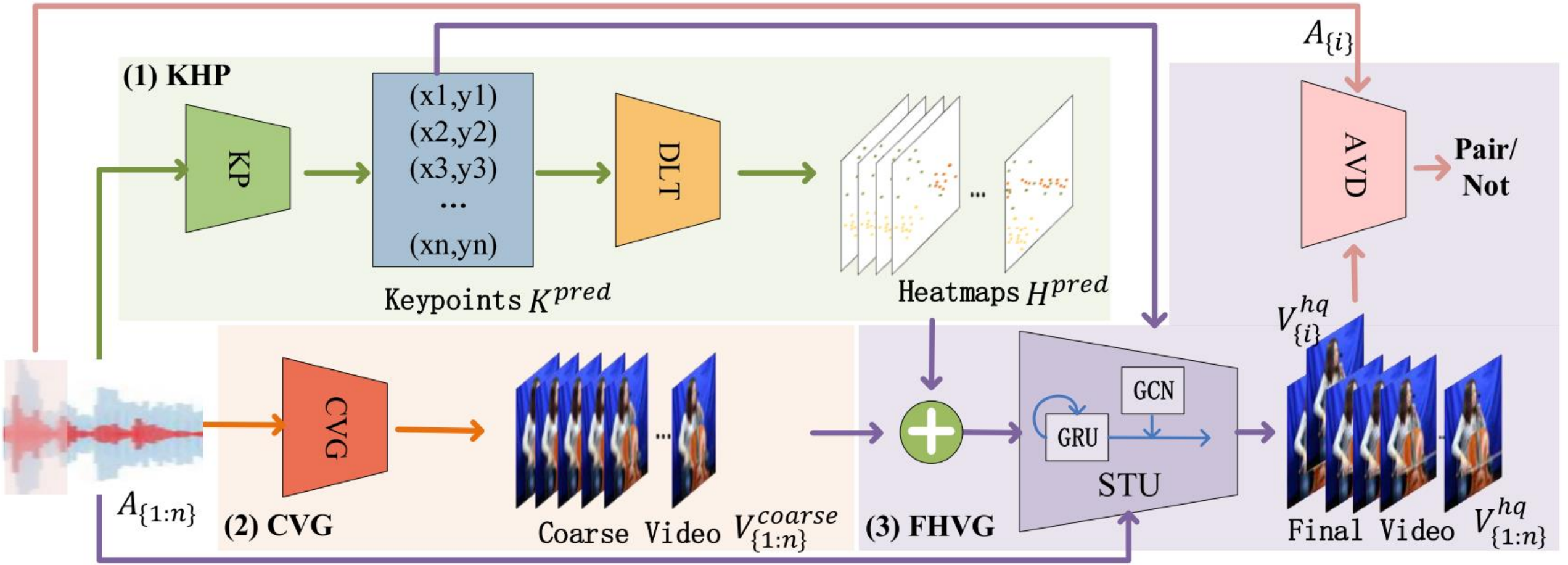
Approach



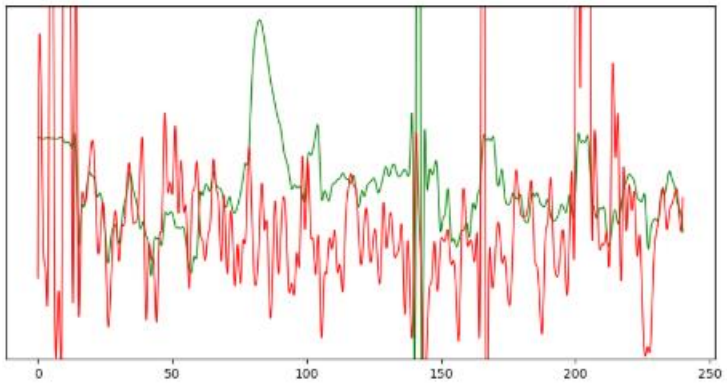
Approach



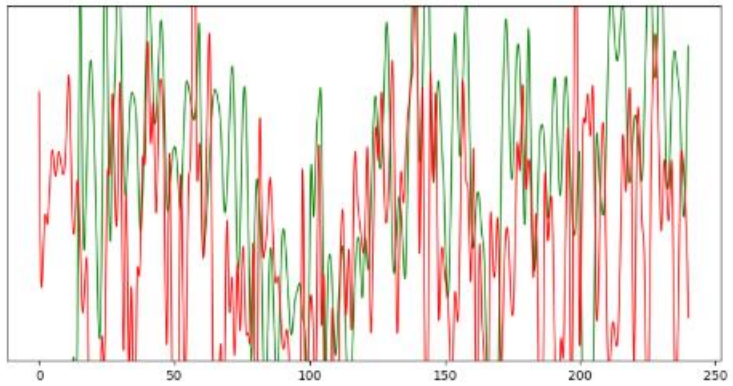
Approach


(1) KHP
KP
(x1,y1)
(x2,y2)
(x3,y3)
...
(xn,yn)
DLT
Keypoints K^{pred}
Heatmaps H^{pred}
$A_{\{i\}}$
AVD
Pair/
Not
$A_{\{1:n\}}$
(2) CVG
CVG
Coarse Video $V^{coarse}_{\{1:n\}}$
(3) FHVG
GCN
GRU
STU
$V^{hq}_{\{i\}}$
Final Video $V^{hq}_{\{1:n\}}$

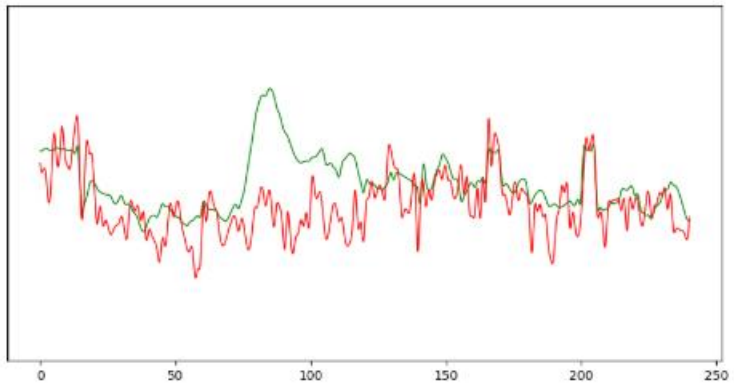
Experiments



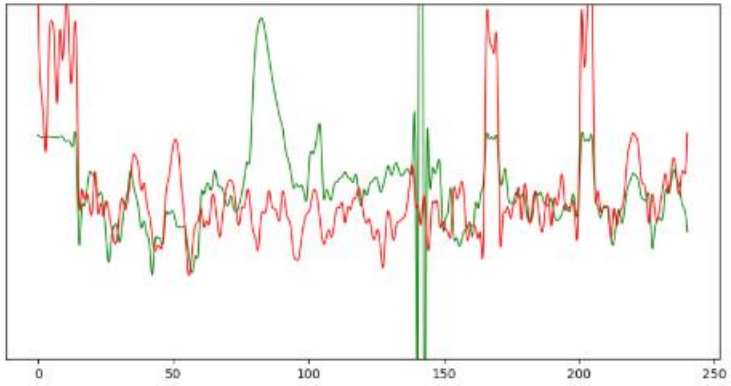
(a) Baseline: left hand



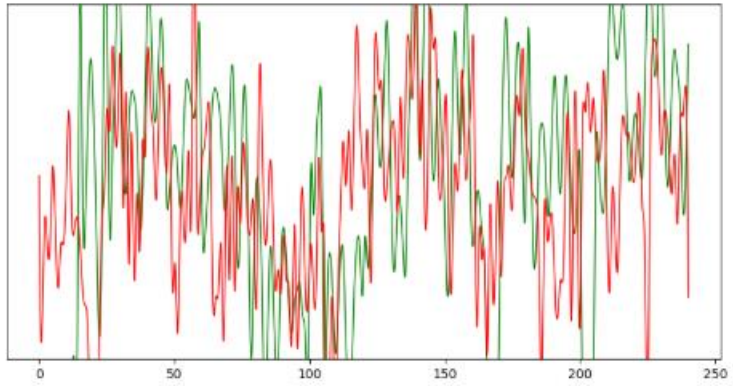
(b) Baseline: right hand



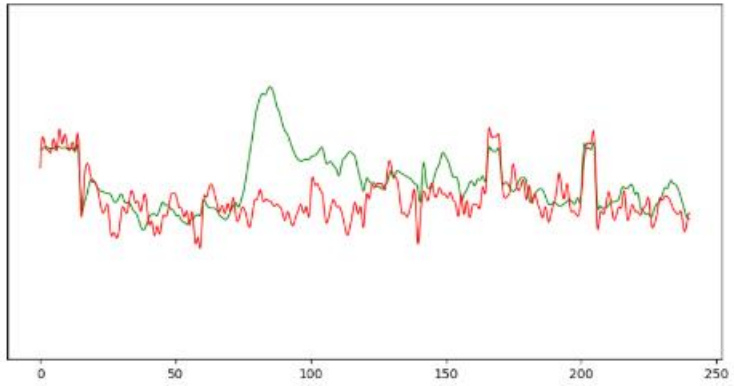
(c) Baseline: body



(d) Ours: left hand



(e) Ours: right hand



(f) Ours: body

Thanks for watching