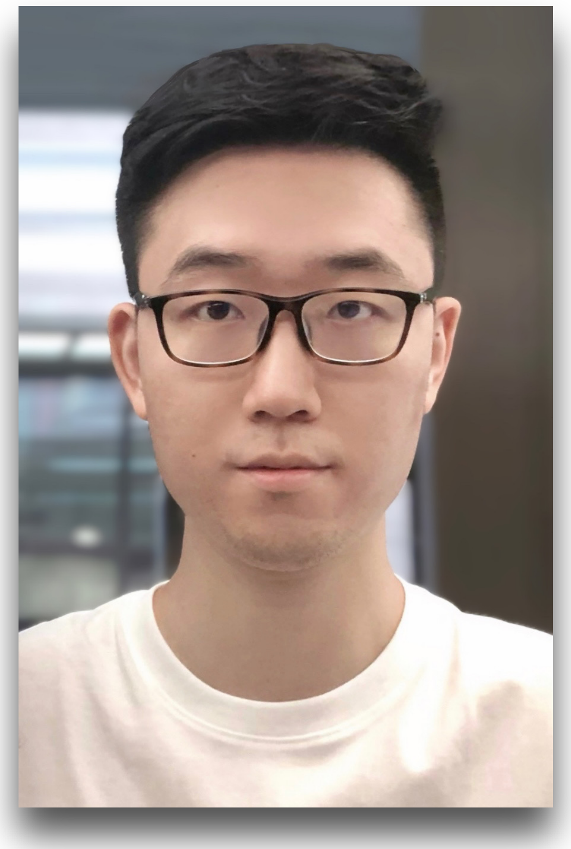
Diverse Human Motion Prediction Guided by Multi-Level Spatial-Temporal Anchors



Sirui Xu

Yu-Xiong Wang*

Liang-Yan Gui*

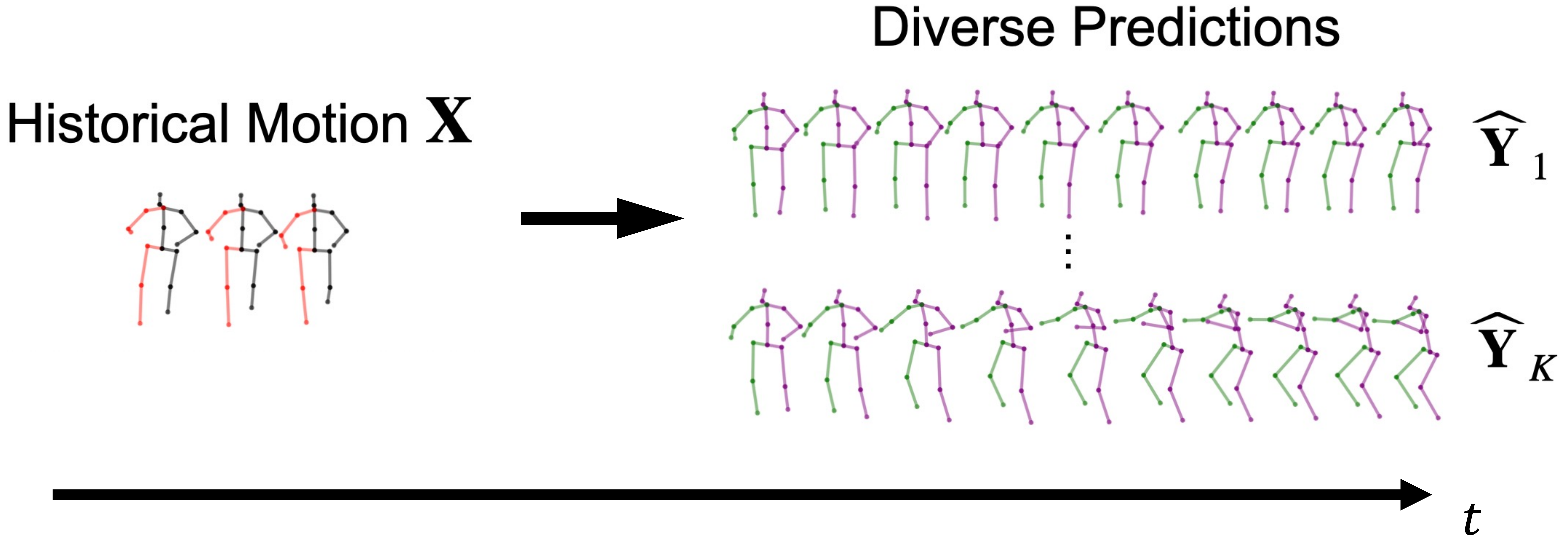
University of Illinois Urbana-Champaign

ILLINOIS
Computer Science
GRAINGER COLLEGE OF ENGINEERING

https://sirui-xu.github.io/STARS/

ECCV
TEL AVIV 2022

Diverse Human Motion Prediction



Historical Motion $\mathbf{X}$

Diverse Predictions

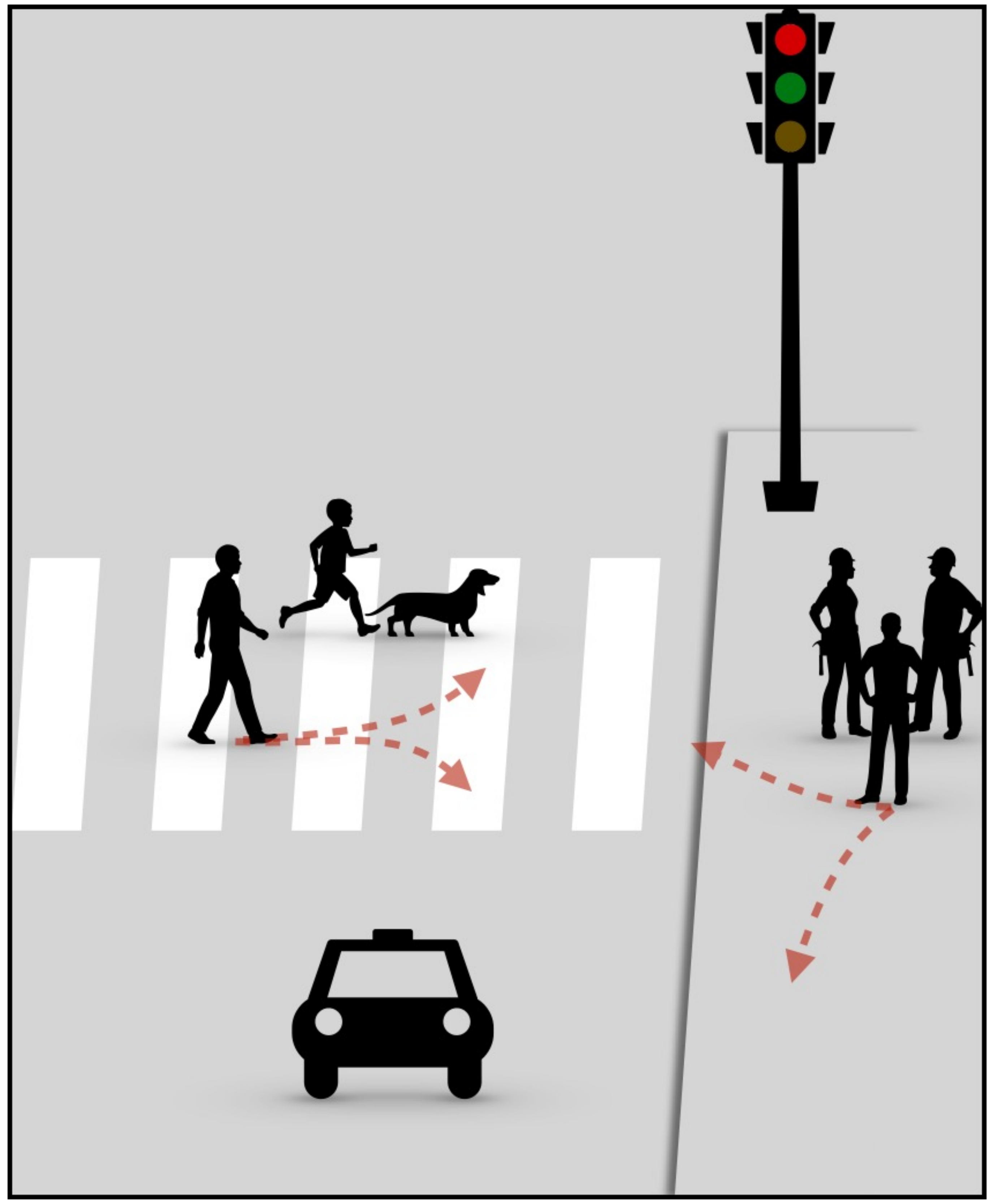
$\widehat{\mathbf{Y}}_1$

$\widehat{\mathbf{Y}}_K$

t

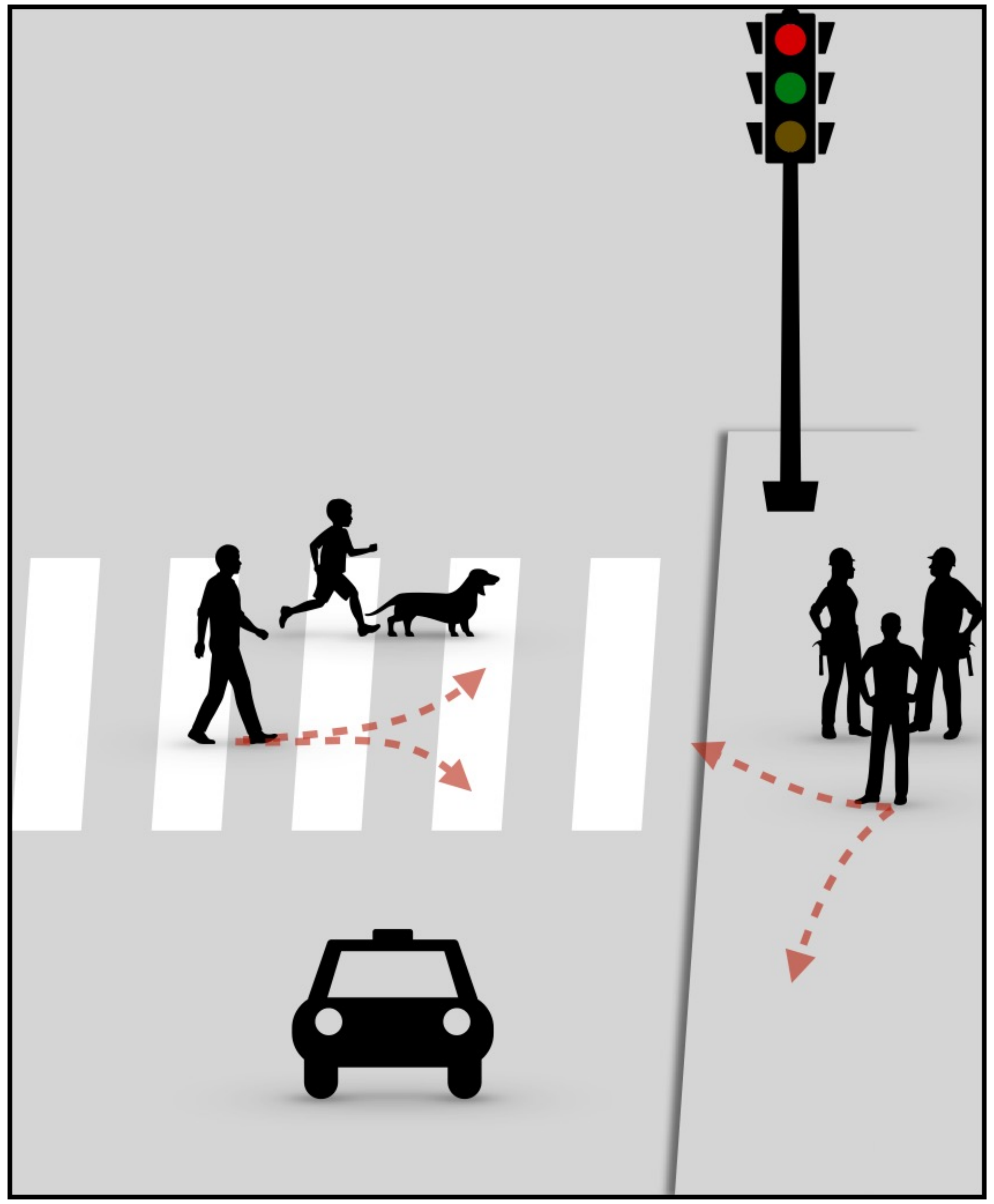
Diverse Human Motion Prediction

- Human future motion is inherently multi-modal, especially in long term



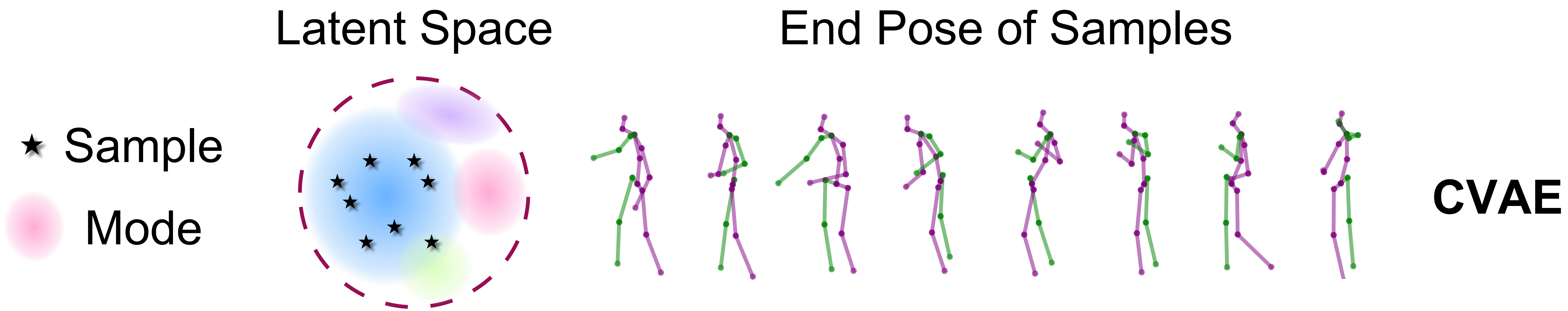
Diverse Human Motion Prediction

- Human future motion is inherently multi-modal, especially in long term
- Predicting a diverse set of human activities is critical for real-world applications



Limitation: likelihood-based sampling

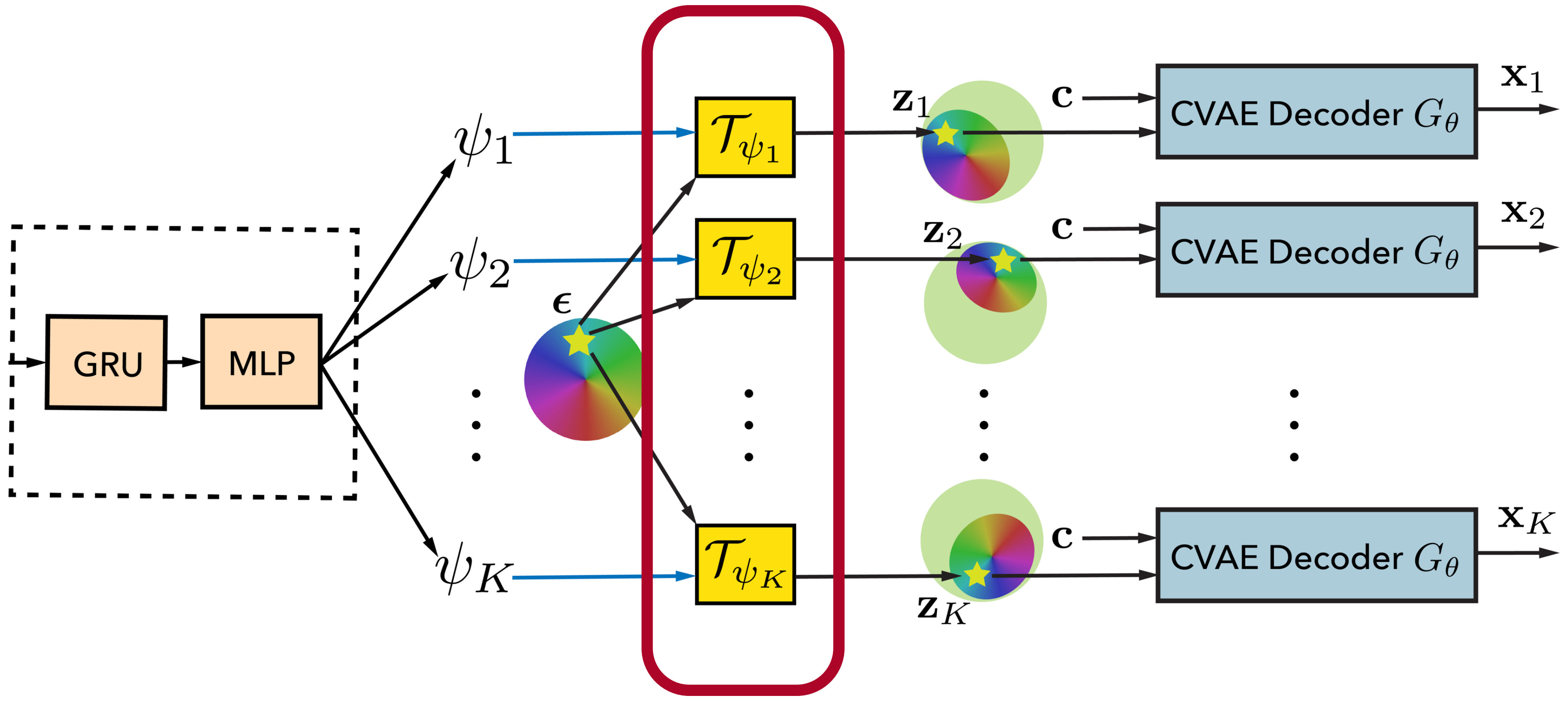
Challenges: predictions are often concentrated in the major mode with less diversity — Mode Collapse



Sohn et al. Learning structured output representation using deep conditional generative models, NeurIPS 2015

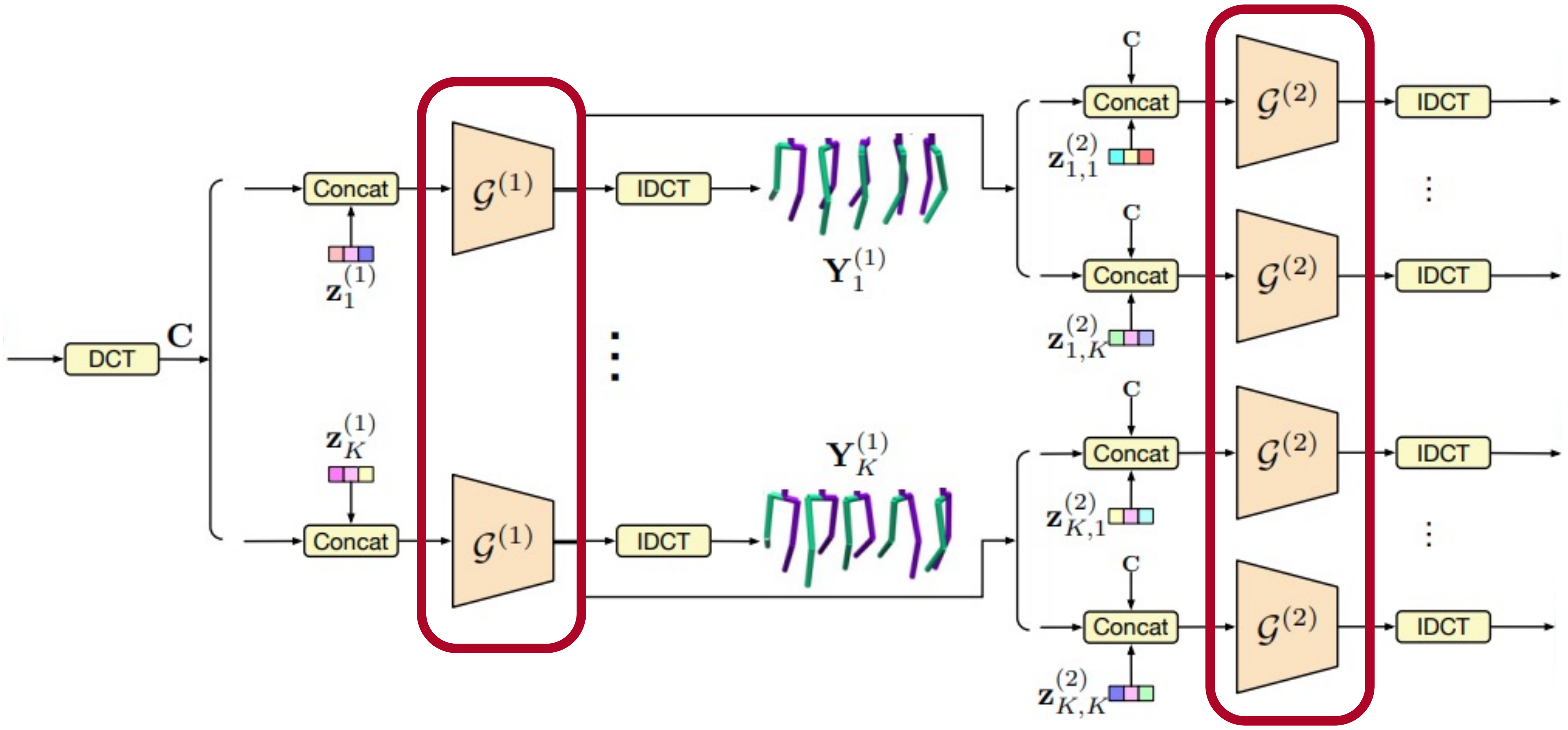
Prior Work

DLow



- Require K additional latent flows to diversify samples
- Need to train the predictor and latent flows in two separate stages

GSPS



- Need to generate different body parts in a sequential manner

Yuan et al. DLow: Diversifying latent flows for diverse human motion prediction, ECCV 2020
Mao et al. Generating smooth pose sequences for diverse human motion prediction, CVPR 2021

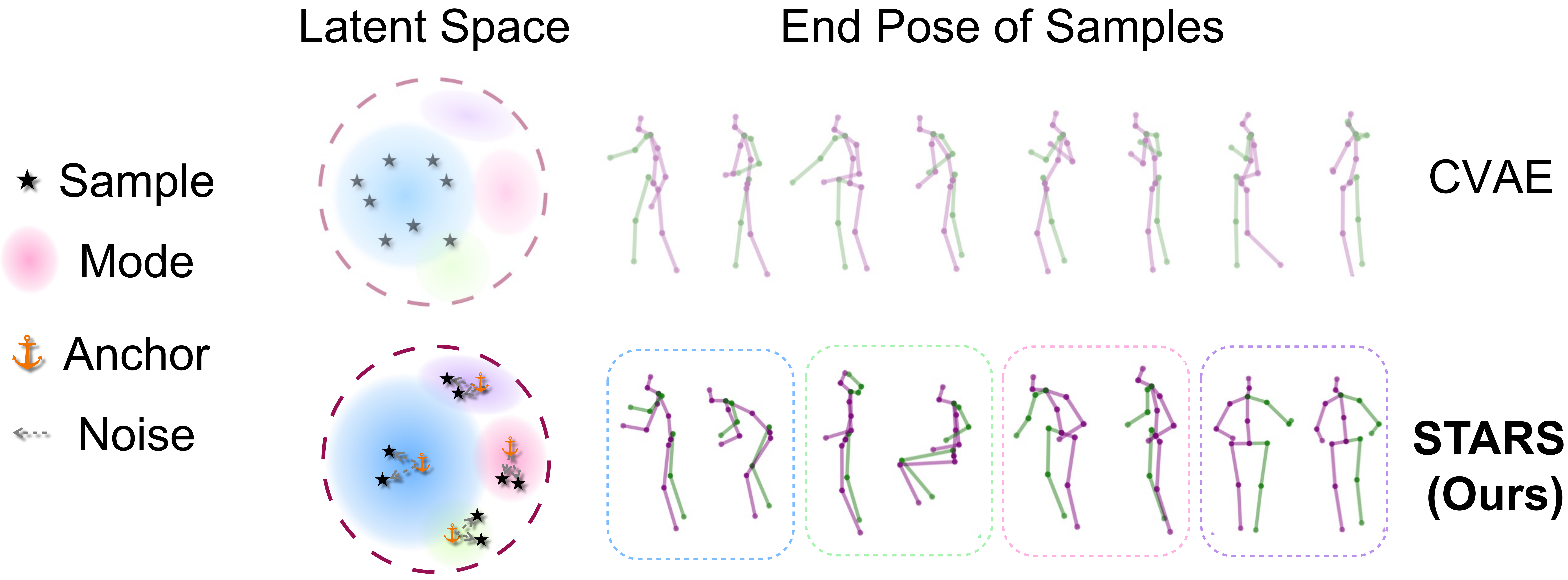
Motivation

- Future motions are not completely random or independent, following
  - Physical laws and body constraints
  - Trends in the history

Motivation

- Future motions are not completely random or independent, following
  - Physical laws and body constraints
  - Trends in the history
- Decompose future human motion in the latent space into
  - Deterministic learnable anchors
  - Stochastic noise

Our Approach: STARS



Latent Space

End Pose of Samples

★ Sample

Mode

⚓ Anchor

Noise

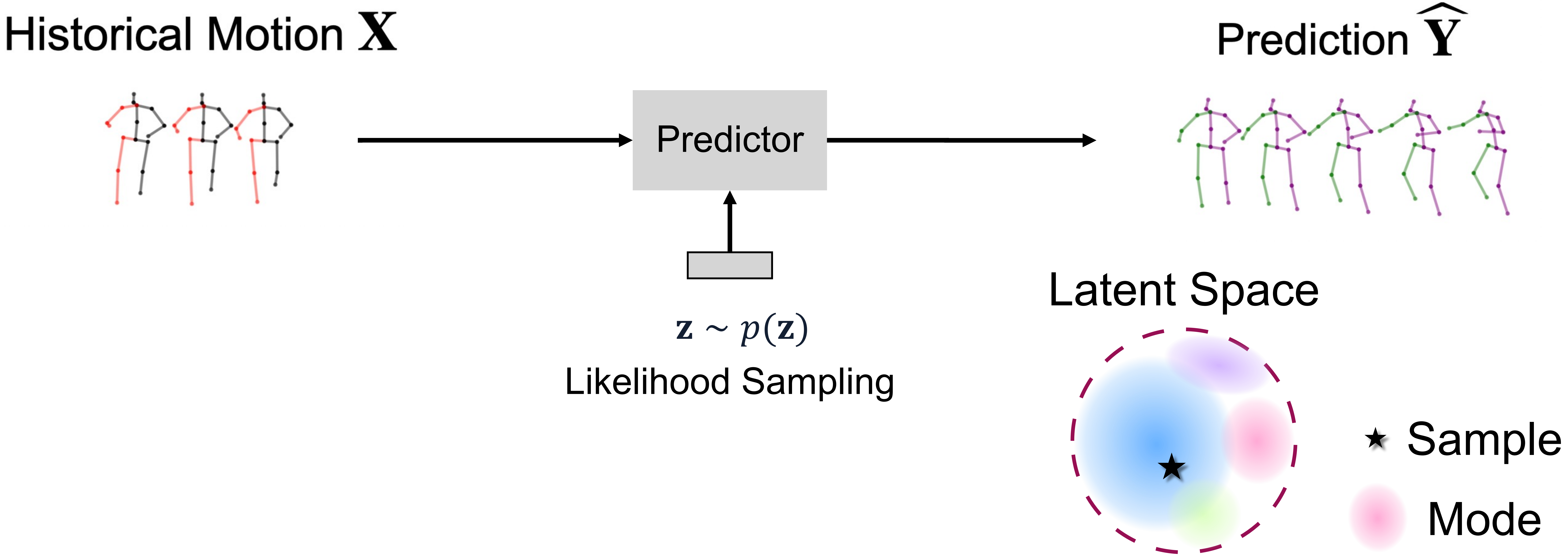
CVAE

STARS (Ours)

Sohn et al. Learning structured output representation using deep conditional generative models, NeurIPS 2015

STARS Formulation

Basic prediction framework



Historical Motion $\mathbf{X}$

Predictor

Prediction $\widehat{\mathbf{Y}}$

$\mathbf{z} \sim p(\mathbf{z})$

Likelihood Sampling

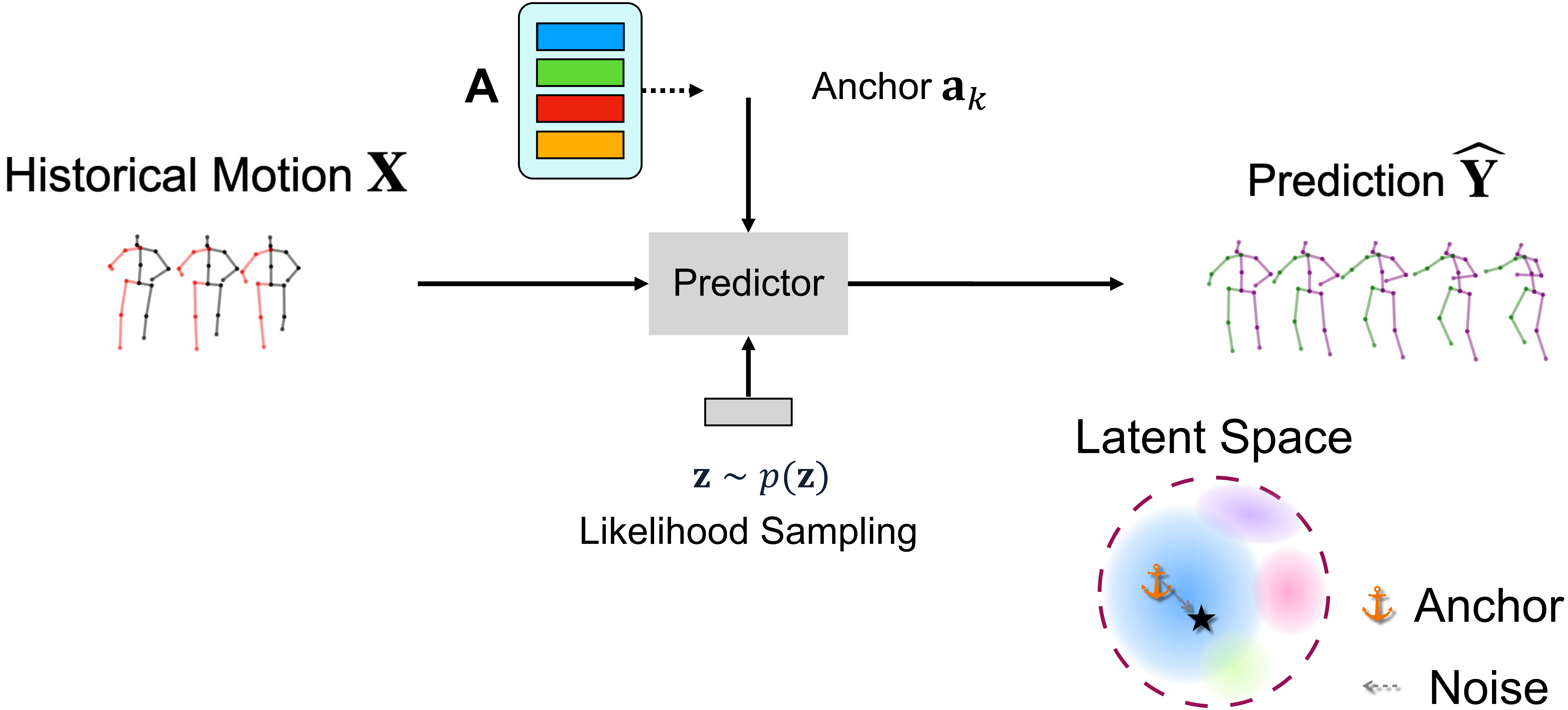
Latent Space

★ Sample

Mode

STARS Formulation

Sampling



$\mathbf{A}$

Anchor $\mathbf{a}_k$

Historical Motion $\mathbf{X}$

Predictor

Prediction $\widehat{\mathbf{Y}}$

$\mathbf{z} \sim p(\mathbf{z})$

Likelihood Sampling

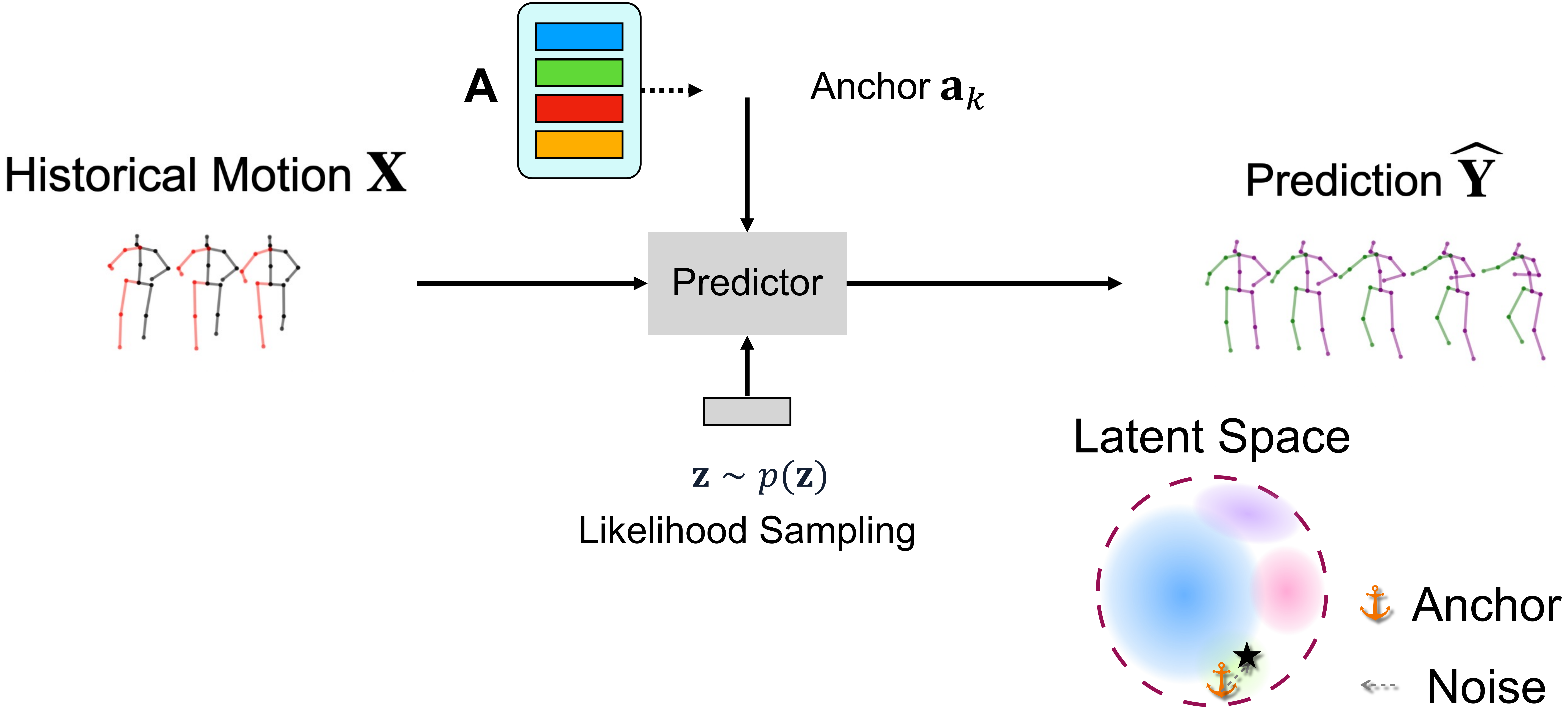
Latent Space

⚓ Anchor

⇠ Noise

STARS Formulation

Sampling



A

Anchor $\mathbf{a}_k$

Historical Motion $\mathbf{X}$

Predictor

Prediction $\widehat{\mathbf{Y}}$

$\mathbf{z} \sim p(\mathbf{z})$

Likelihood Sampling

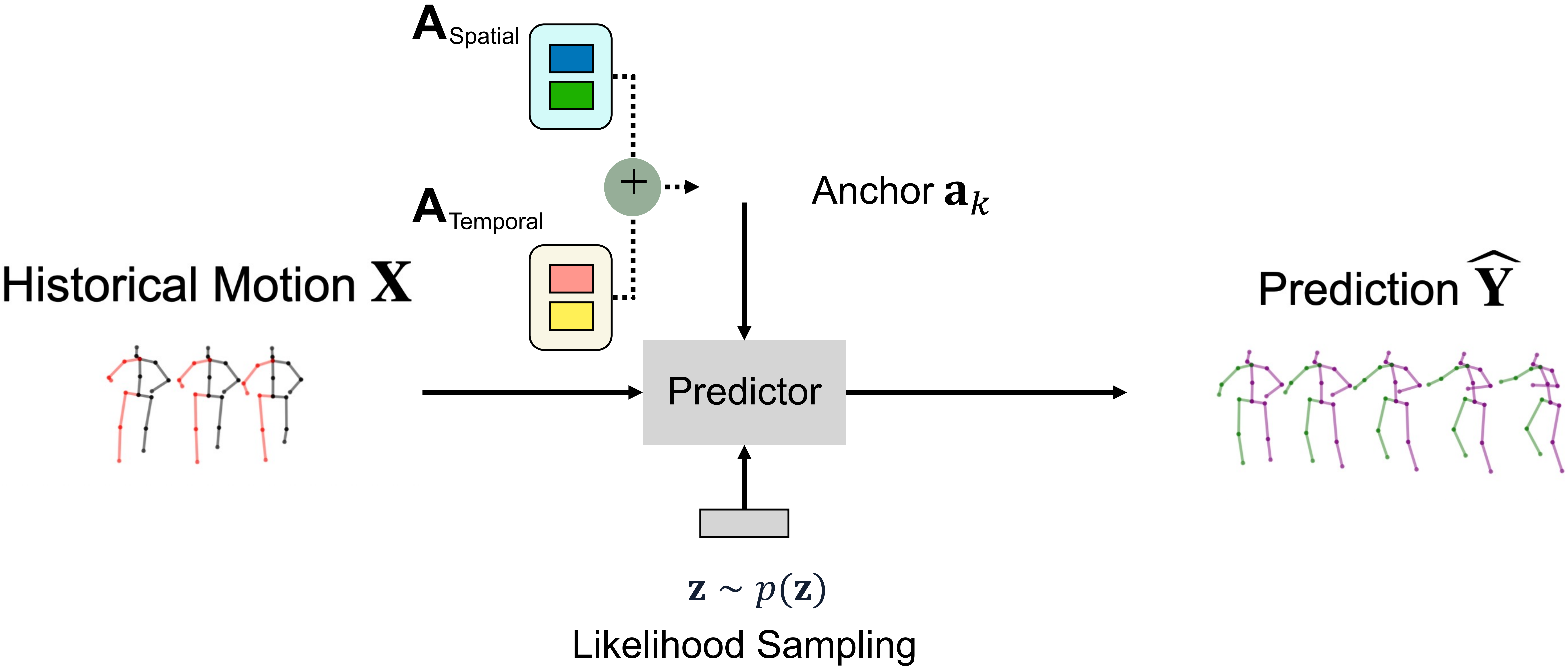
Latent Space

⚓ Anchor

⇠ Noise

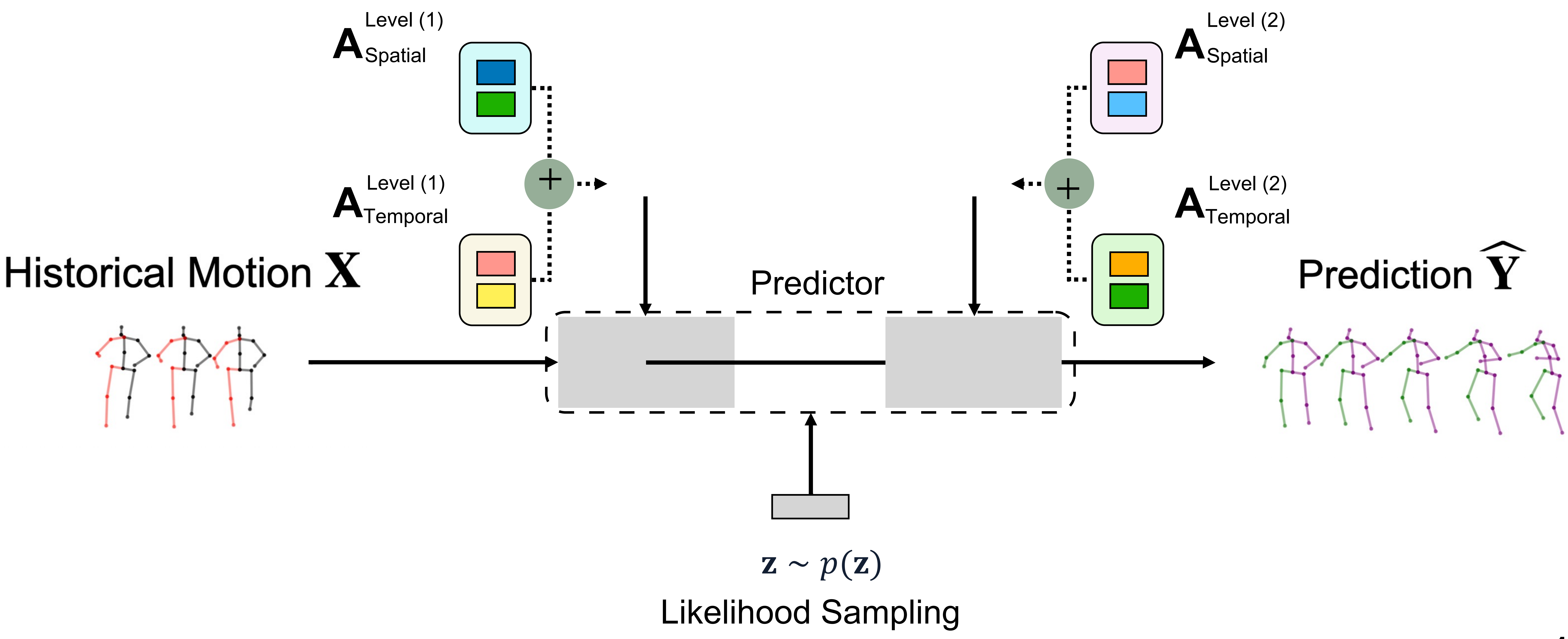
STARS Formulation

Sampling: spatial-temporal decomposition


$\mathbf{A}_{\text{Spatial}}$
$\mathbf{A}_{\text{Temporal}}$
Anchor $\mathbf{a}_k$
Historical Motion $\mathbf{X}$
Predictor
Prediction $\widehat{\mathbf{Y}}$
$\mathbf{z} \sim p(\mathbf{z})$
Likelihood Sampling

STARS Formulation

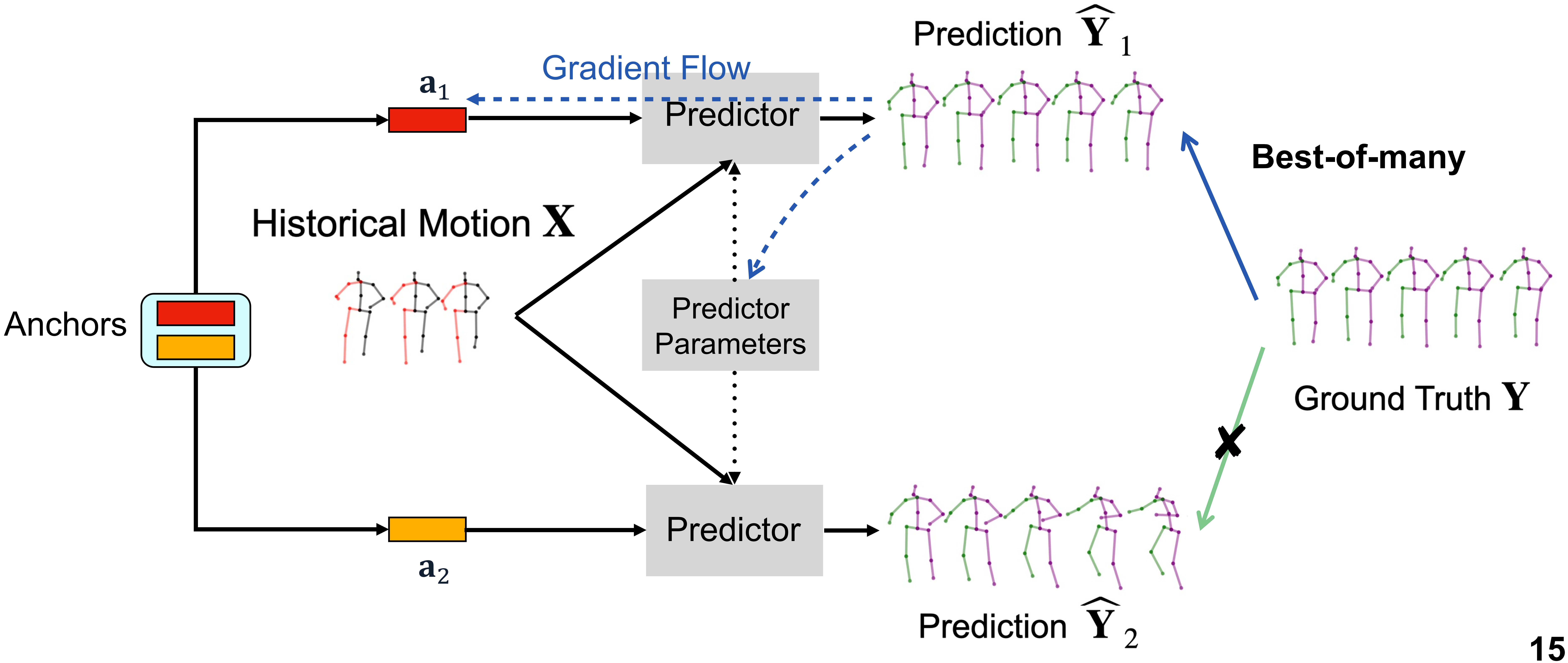
Sampling: multi-level decomposition



$\mathbf{A}_{\text{Spatial}}^{\text{Level (1)}}$
$\mathbf{A}_{\text{Temporal}}^{\text{Level (1)}}$
$\mathbf{A}_{\text{Spatial}}^{\text{Level (2)}}$
$\mathbf{A}_{\text{Temporal}}^{\text{Level (2)}}$
+
+
Historical Motion $\mathbf{X}$
Predictor
Prediction $\widehat{\mathbf{Y}}$
$\mathbf{z} \sim p(\mathbf{z})$
Likelihood Sampling

STARS Formulation

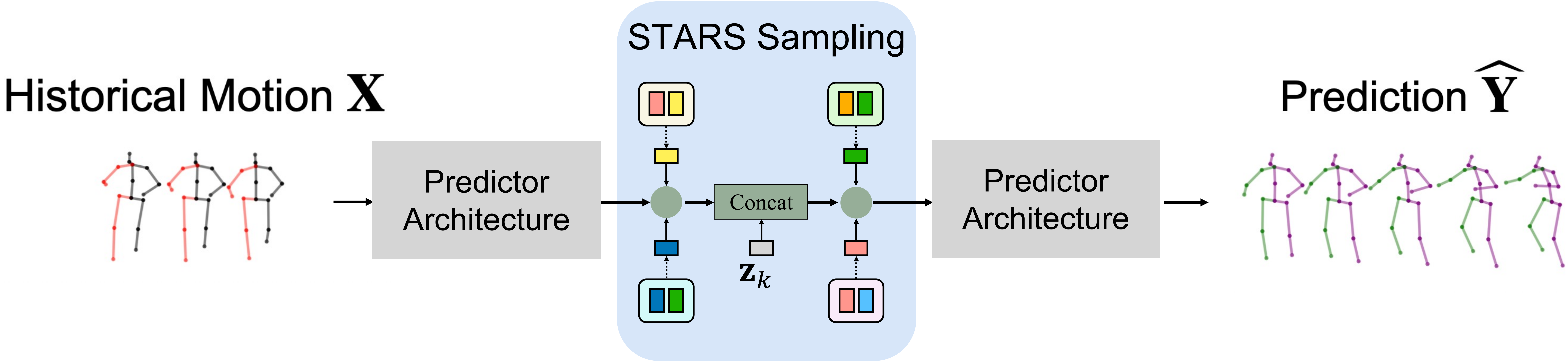
Training


Prediction $\widehat{\mathbf{Y}}_1$
Gradient Flow
$\mathbf{a}_1$
Predictor
Best-of-many
Historical Motion $\mathbf{X}$
Anchors
Predictor Parameters
Ground Truth $\mathbf{Y}$
Predictor
$\mathbf{a}_2$
Prediction $\widehat{\mathbf{Y}}_2$

Predictor Architecture

Plug-in anywhere

- STARS sampling is general, agnostic to predictor architectures



Predictor Architecture

- Using Discrete Cosine Transform (DCT) to convert motions to the frequency domain
- Using Spatial-Temporal Graph Convolutional Network (STGCN)

Historical Motion $\mathbf{X}$
DCT
STGCN
STARS Sampling
Concat
$\mathbf{z}_k$
STGCN
IDCT
Prediction $\widehat{\mathbf{Y}}$

Predictor Architecture: IE-STGCN

Bottleneck spatial-temporal interactions

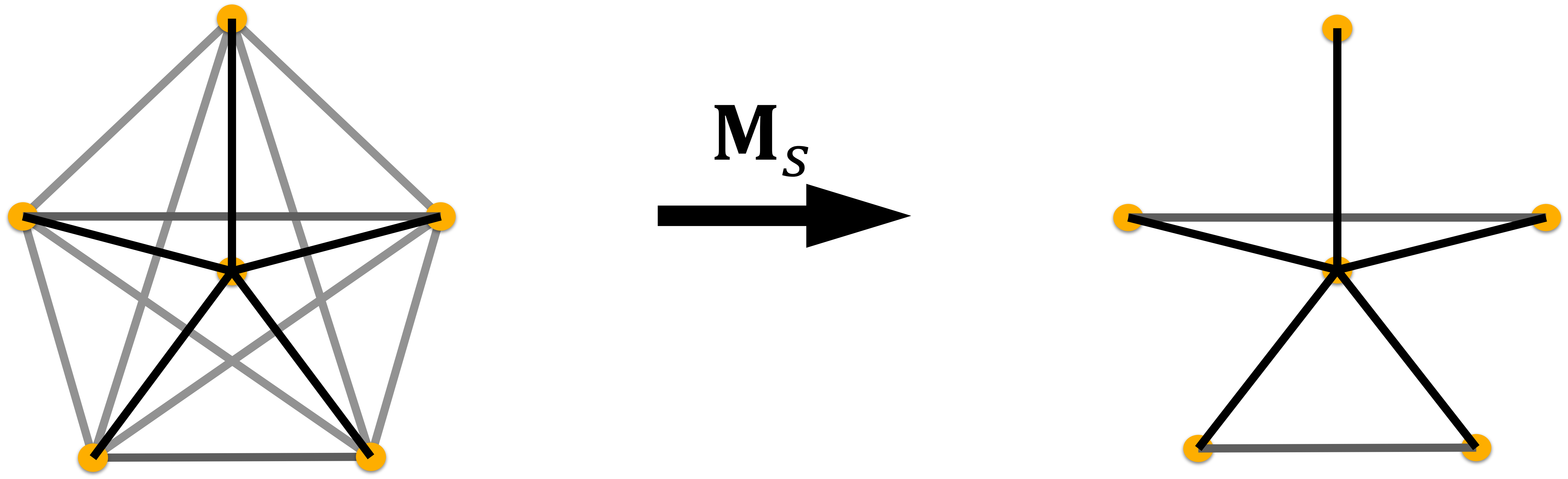
- Spatial-Temporal Graph Convolutional Network (STGCN): $\mathbf{H}_k^{(l+1)} = \sigma(\mathbf{Adj}^{(l)}\mathbf{H}_k^{(l)}\mathbf{W}^{(l)})$
  - Factorizing spatial-temporal connectivity: $\mathbf{Adj}^{(l)} = \mathbf{Adj}_s^{(l)}\mathbf{Adj}_f^{(l)}$
  - Incorporating spatial-temporal anchors

Predictor Architecture: IE-STGCN

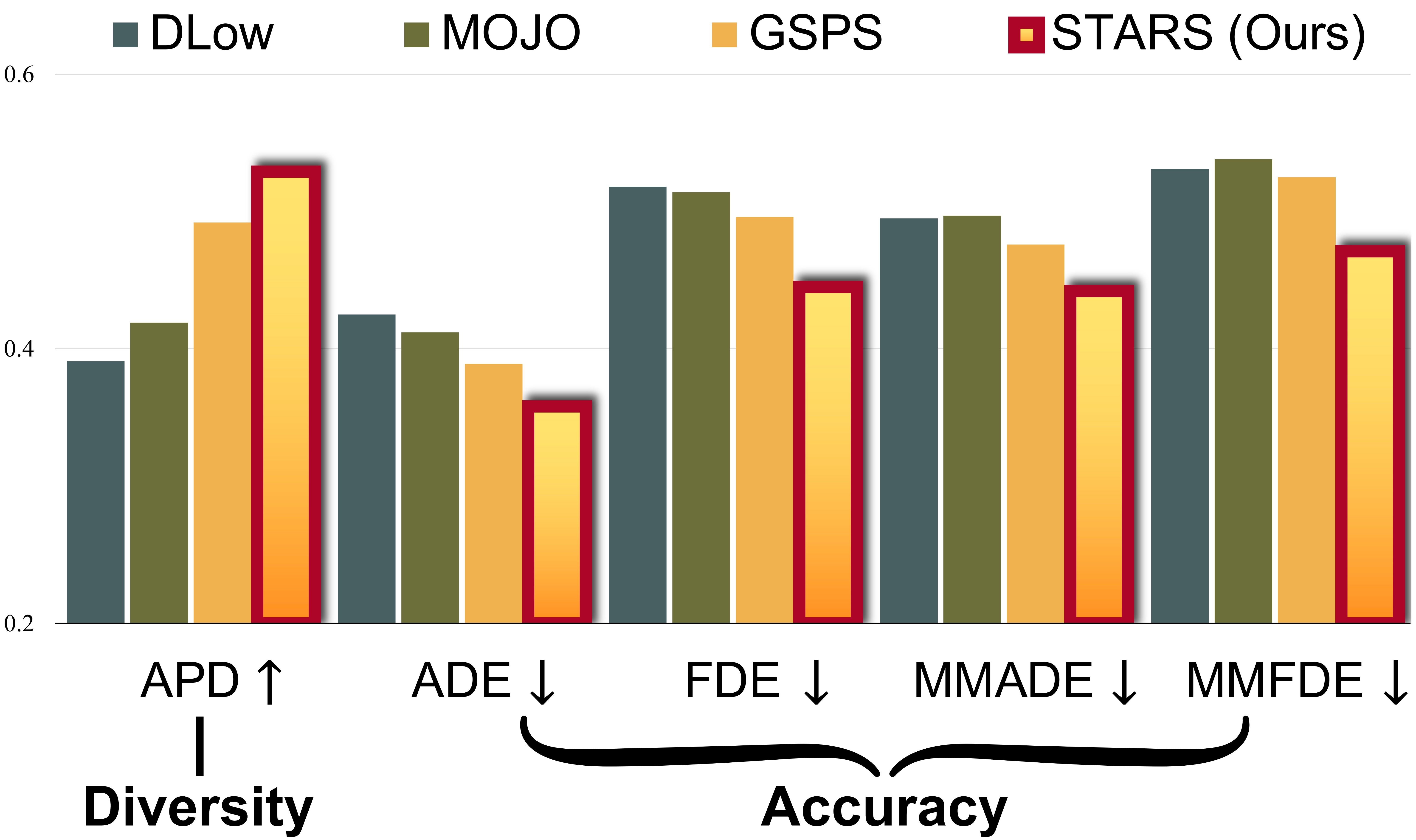
Bottleneck spatial-temporal interactions

- Spatial-Temporal Graph Convolutional Network (STGCN): $\mathbf{H}_k^{(l+1)} = \sigma(\mathbf{Adj}^{(l)}\mathbf{H}_k^{(l)}\mathbf{W}^{(l)})$

- Spatial Interaction Pruning: $\hat{\mathbf{Adj}}_s^{(l)} = \mathbf{M}_s \odot \mathbf{Adj}_s^{(l)}$



STARS significantly improves diversity and accuracy



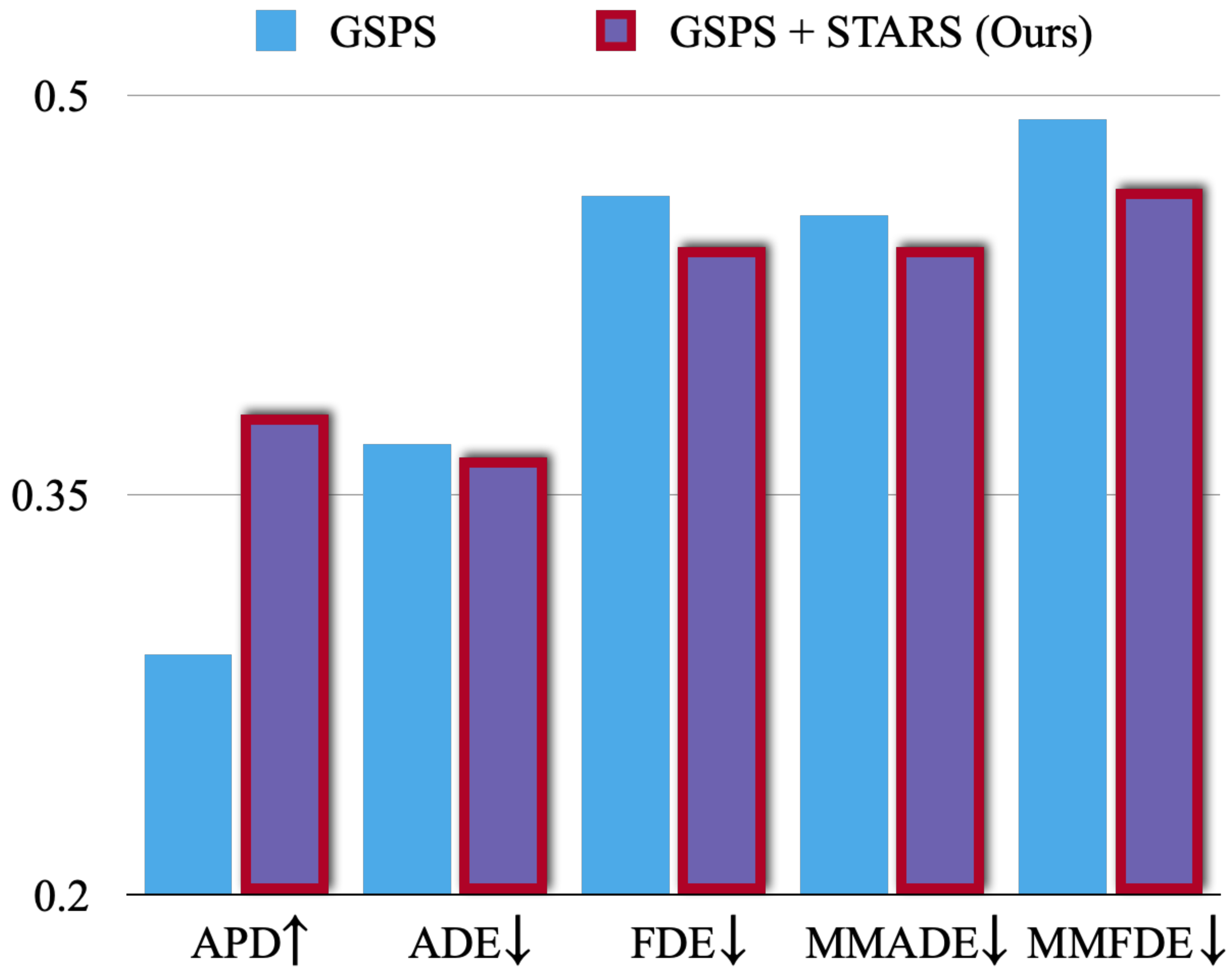
Diversity

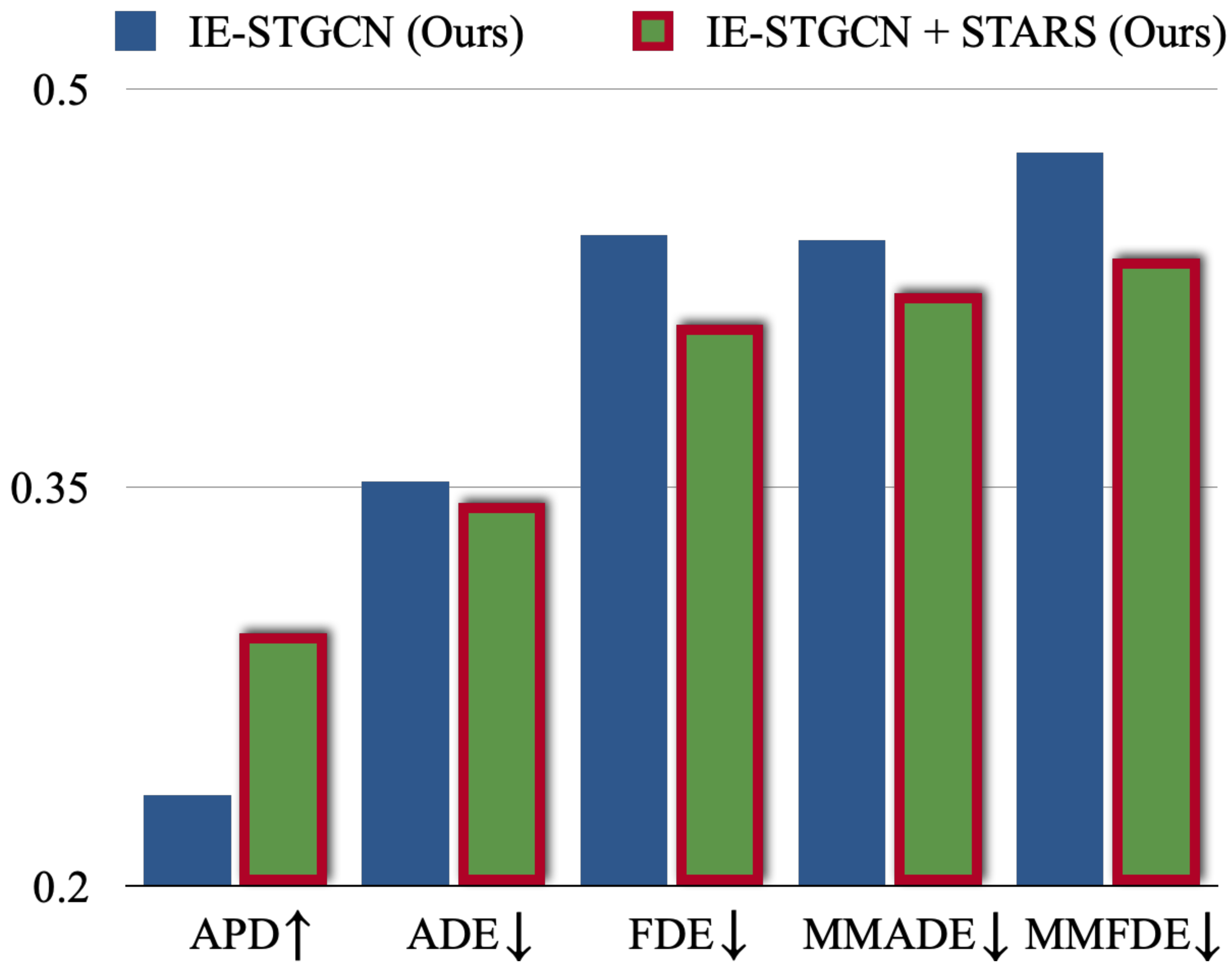
Accuracy

Human3.6M, #predictions = 50

Yuan et al. DLow: Diversifying latent flows for diverse human motion prediction, ECCV 2020
Zhang et al. We are more than our joints: Predicting how 3D bodies move, CVPR 2021
Mao et al. Generating smooth pose sequences for diverse human motion prediction, ICCV 2021

STARS is general with different predictor architectures

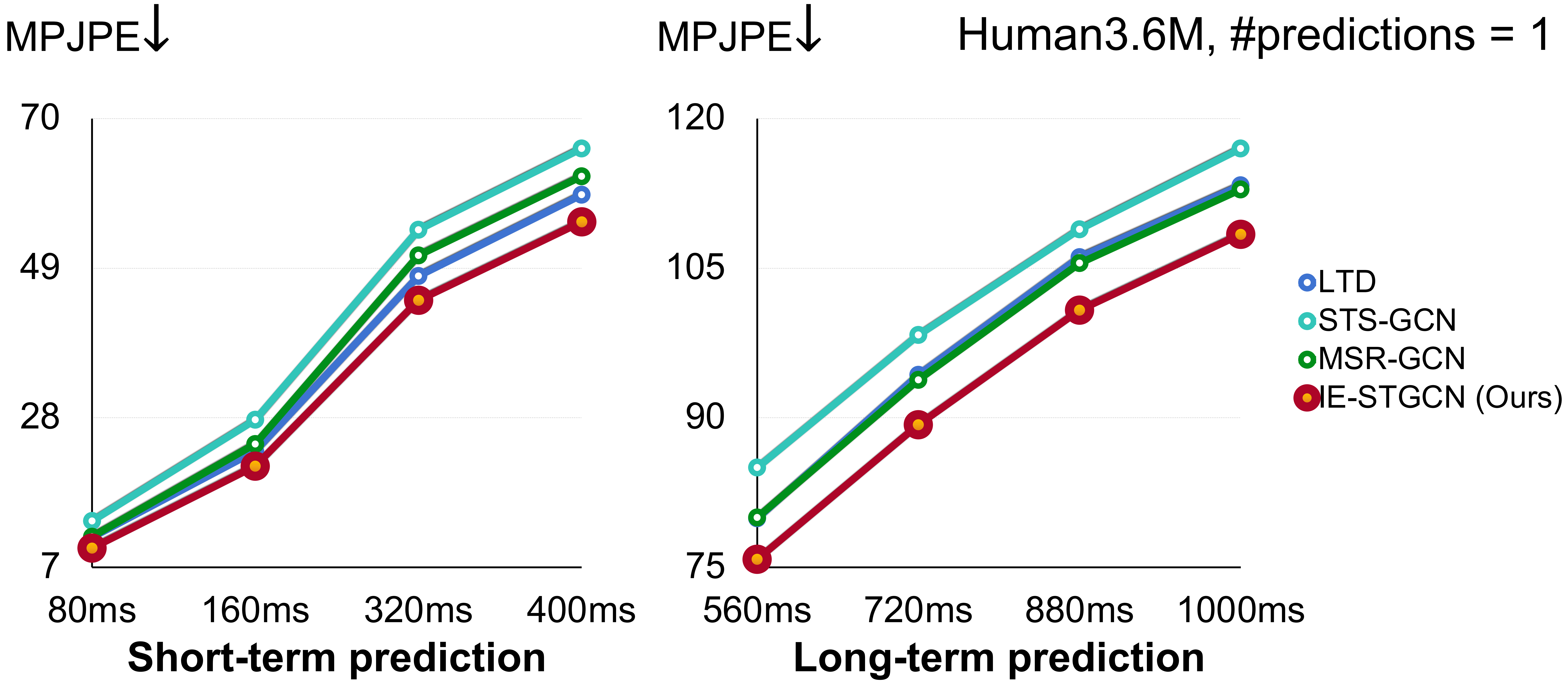




Human3.6M, #predictions = 100

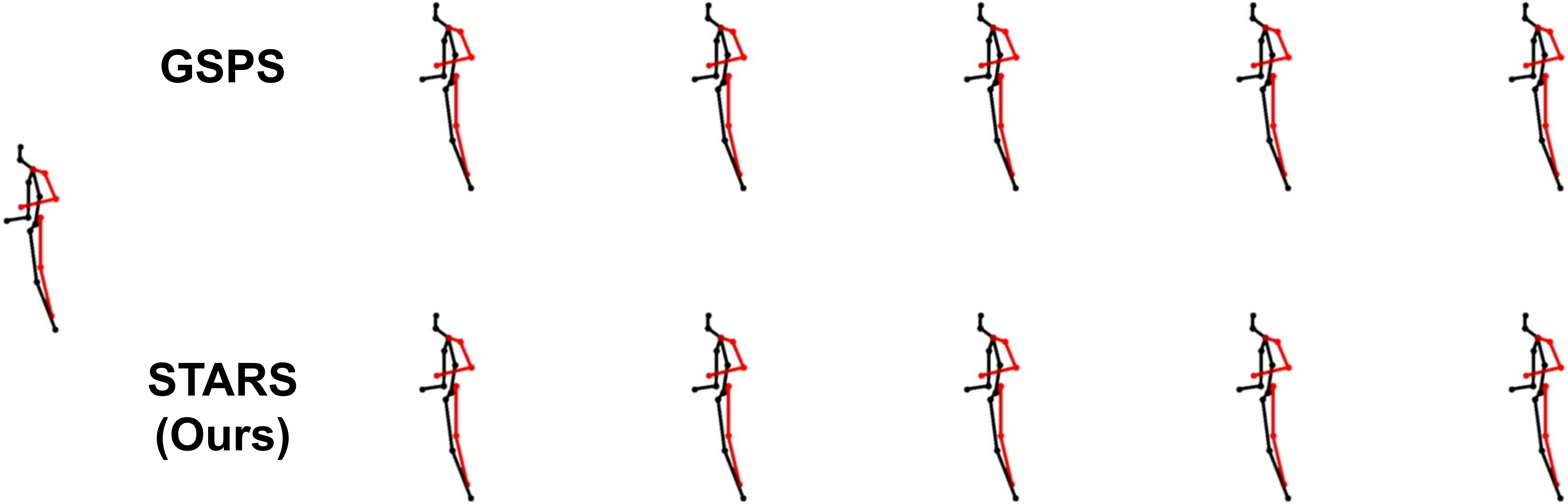
Mao et al. Generating smooth pose sequences for diverse human motion prediction, ICCV 2021

Generalizable to Deterministic Motion Prediction



Mao et al. Learning trajectory dependencies for human motion prediction, ICCV 2019
Sofianos et al. Space-time-separable graph convolutional network for pose forecasting, ICCV 2021
Dang et al. MSR-GCN: Multi-scale residual graph convolution networks for human motion prediction, ICCV 2021

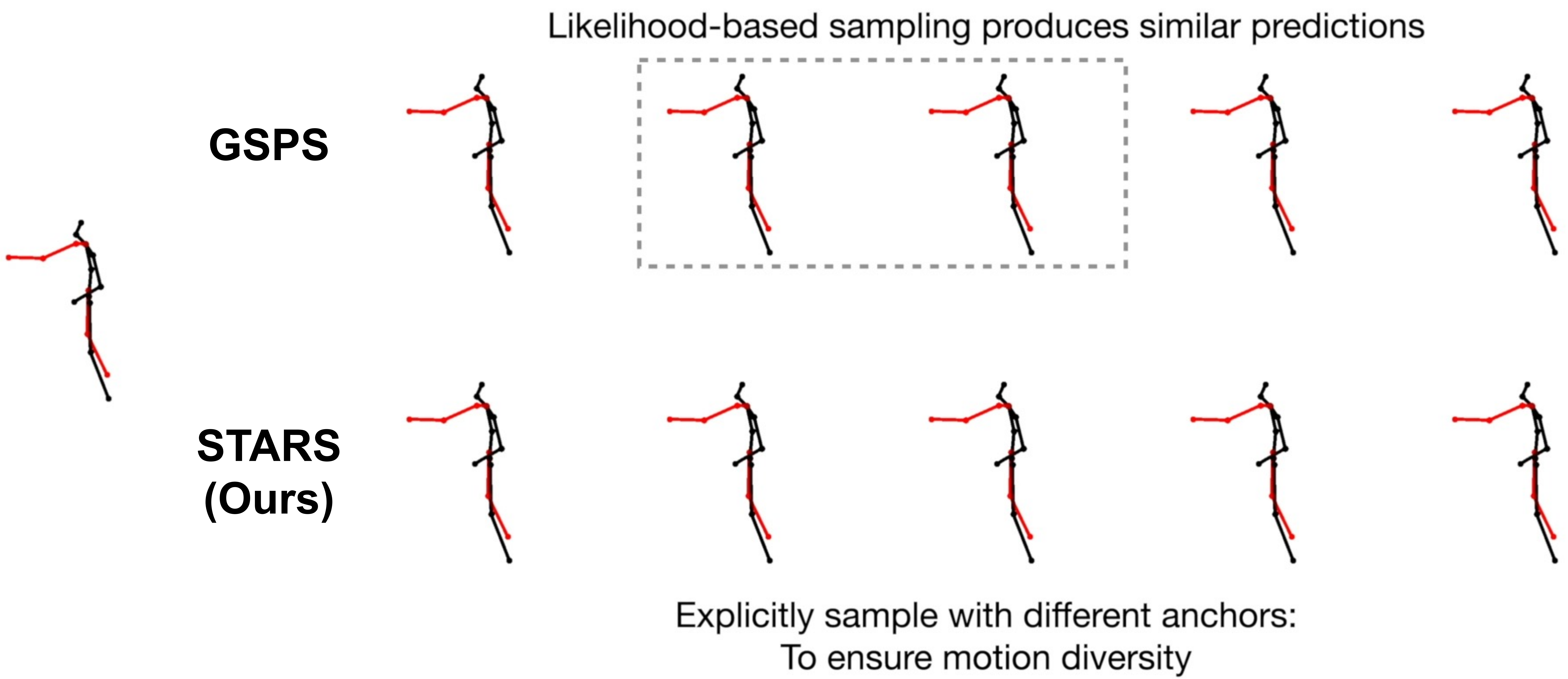
Diverse Motion Prediction



GSPS

STARS (Ours)

Diverse Motion Prediction



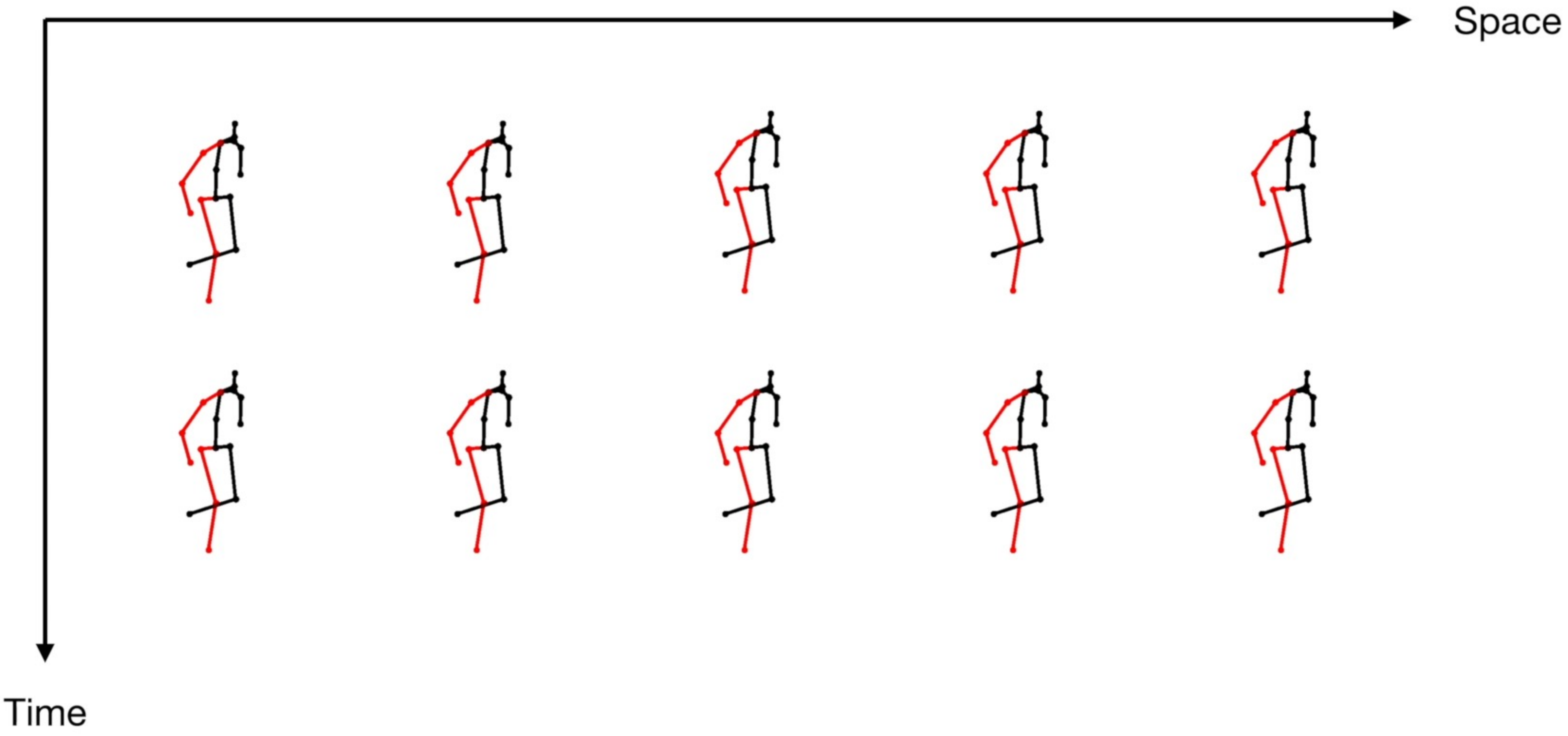
Likelihood-based sampling produces similar predictions

GSPS

STARS
(Ours)

Explicitly sample with different anchors:
To ensure motion diversity

Controllable Motion Prediction



Conclusions

- **STARS:** a simple yet effective and general framework that leverages **learnable anchors** to **diversify** predictions
- Enable controllable motion prediction in native space and time with spatial-temporal anchors
- Future work: extend STARS for other prediction tasks

Thank you

And welcome to

Poster 1.A, 49
25-Oct-22

ILLINOIS
Computer Science
GRAINGER COLLEGE OF ENGINEERING

https://sirui-xu.github.io/STARS/

ECCV
TEL AVIV 2022