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

ILLINOIS Computer Science GRAINGER COLLEGE OF ENGINEERING

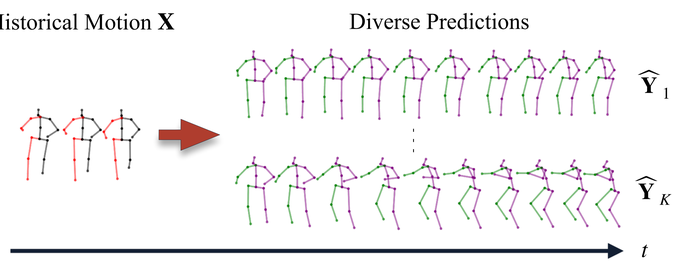
Sirui Xu **Yu-Xiong Wang*** **Liang-Yan Gui***

University of Illinois at Urbana-Champaign

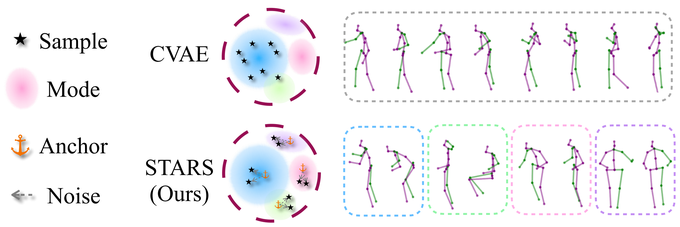
ECCV TEL AVIV 2022

Motivation

- **Problem Statement:** diverse human motion prediction

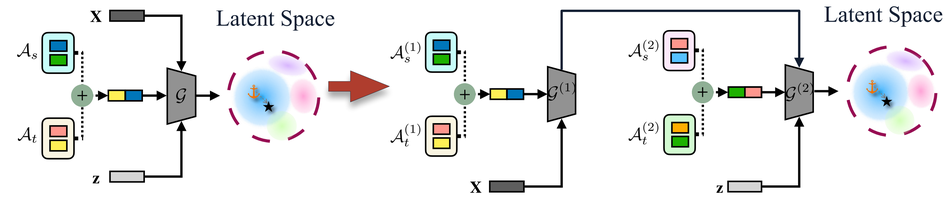


- **Limitation & challenge:** predictions are often concentrated in the major mode with less diversity
- **Key insight:** future motion (diversity) is not completely random or independent, following
  - Physical laws and body constraints
  - Trends in the history



- **Our approach:** decompose future human motions
  - Deterministic shareable and learnable latent codes named *Anchor*
    - Share across all historical motions
    - Easy to optimize and diversify
    - Factorize spatial and temporal (frequency) components
  - Stochastic noise

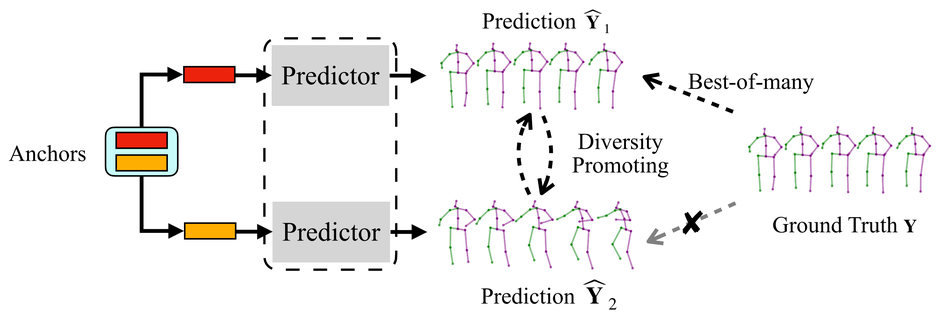
Multi-Level Spatial-Temporal Anchors



- **Sample with spatial-temporal anchor:**
- $\mathbf{z} \sim p(\mathbf{z})$, $\widehat{\mathbf{Y}}_k = \mathcal{G}(\mathbf{a}_i^s + \mathbf{a}_j^t, \mathbf{z}, \mathbf{X})$
- Spatial anchor: vary at spatial dimension
- Temporal anchor: vary at frequency dimension

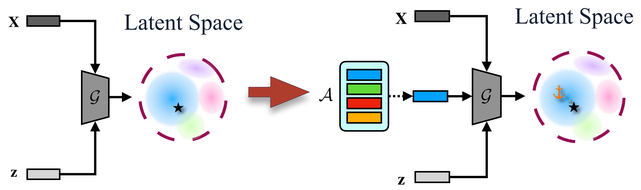
- **Sample with multi-level anchor:**
- $\mathbf{z} \sim p(\mathbf{z})$, $\widehat{\mathbf{Y}}_k = \mathcal{G}^{(2)}(\mathbf{a}_i^{s_2} + \mathbf{a}_j^{t_2}, \mathbf{z}, \mathcal{G}^{(1)}(\mathbf{a}_i^{s_1} + \mathbf{a}_j^{t_1}, \mathbf{X}))$
- $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}$: different network parts

STARS: Training



- Forward pass: generating future motions with **each anchor explicitly**
- **Backward pass:** anchors are based on separate losses; Backbone is based on fused losses
  - **Best-of-many**: optimize the best predictions
  - **Diversity-promoting**: promote pair-wise distance between predictions

Anchor-Based Sampling

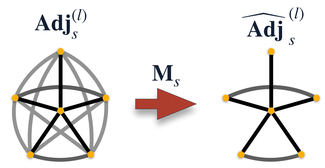


- **Conventional sampling:**
- $\mathbf{z} \sim p(\mathbf{z})$, $\widehat{\mathbf{Y}} = \mathcal{G}(\mathbf{z}, \mathbf{X})$

- **STARS sampling (Ours):**
- $\mathbf{z} \sim p(\mathbf{z})$, $\widehat{\mathbf{Y}}_k = \mathcal{G}(\mathbf{a}_k, \mathbf{z}, \mathbf{X})$
- $\mathbf{a}_k \in \mathcal{A} = \{\mathbf{a}_k\}_{k=1}^K$: K *learnable* codes

Predictor Architecture: Interaction-Enhanced Spatial-Temporal GNN

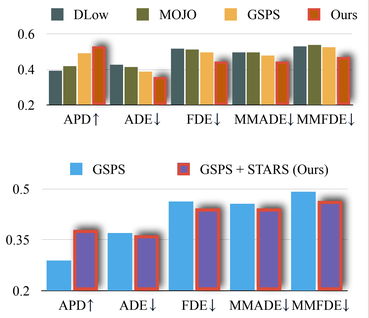
- **Motivation:** Better incorporate multi-level spatial-temporal anchors
- Use DCT to convert motions to the **frequency domain**: $\mathbf{H}_k^{(0)} = \mathbf{C}\mathbf{X}$
- Use Graph Convolutional Network: $\mathbf{H}_k^{(l+1)} = \sigma(\mathbf{Adj}^{(l)}\mathbf{H}_k^{(l)}\mathbf{W}^{(l)})$
  - **Bottleneck spatial-temporal interactions**
    - Factorize the adjacency matrix: $\mathbf{Adj}^{(l)} = \mathbf{Adj}_s^{(l)}\mathbf{Adj}_f^{(l)}$
    - Cross-Layer Interaction Sharing: $\mathbf{Adj}_s^{(l)} = \mathbf{Adj}_s^{(l+2)}$
    - Spatial Interaction Pruning: $\widehat{\mathbf{Adj}}_s^{(l)} = \mathbf{M}_s \odot \mathbf{Adj}_s^{(l)}$



Quantitative Evaluation

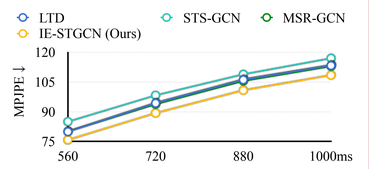
Diverse Motion Prediction

- Compare to prior work
  - Significantly outperform in diversity (APD↑) and accuracy (Others↓)
- A **general** framework, agnostic to backbones
  - Benefit GSPS across all metrics



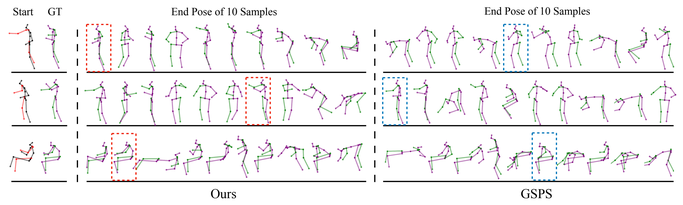
Deterministic Motion Prediction

- Significantly outperform baselines in accuracy



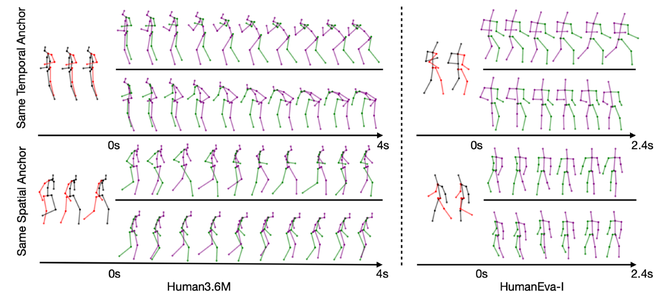
Qualitative Evaluation

Diverse Motion Prediction



Controllable Motion Prediction

- Prior work: Only body parts control → Ours: control of space and time



Conclusions

- **STARS**: a simple yet effective sampling framework that leverages learnable anchors to represent human motion
- Enable novel controllable motion prediction with spatial-temporal anchors
- **Future work**: extend STARS for human-scene interaction prediction