時系列的整合性を考慮した テキスト指示による長系列な人物動作生成

天谷幸太郎 [1,2]，加藤駿 [1,2]，澁谷崇 [2,3]，Julian Tanke[2,3]，光藤祐樹 [2,3]，五十川麻理子 [1,3]
慶應義塾大学 [1]，慶応 AI センター [2]，Sony AI[3]

研究概要

背景

テキスト指示をもとにした人物の動作生成は
映画やゲームなどのクリエイション，ロボティクス，
将来的なインターフェースなどへの応用が期待される.

現状の課題

- Text-Motion ペアの既存データセットは十数秒の動作で構築され，時系列的整合性を学習できない
- 既存の長期動作生成手法で主流である Motion Stiching を用いる手法において，過去時系列の情報を保持，考慮していない

本研究の貢献

- 20s~30s ほどの時系列的整合性を考慮する必要のある動作を含む検証データセットの構築
- 潜在拡散モデルを用いた過去時系列情報を考慮した長期モーション生成モデルの提案

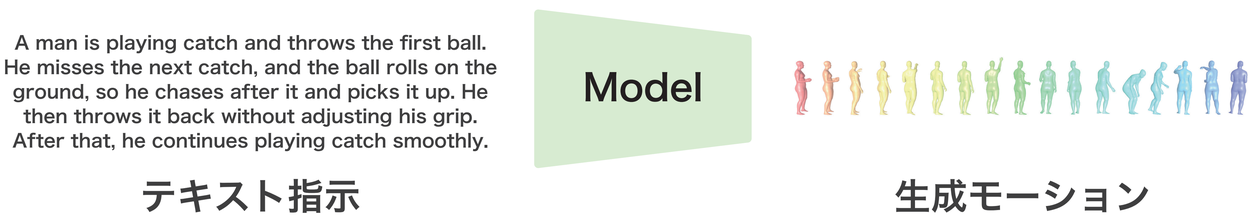

図 1. テキストによる人物動作生成

本研究の扱う課題設定

長期時系列生成における整合性

あるコンテキスト (Event) によって，以降に生成される動作 (Trigger) の自由度が制限される場合と定義する.

- **空間位置・向き**
  - Event：in, on, at, behind, in front of ...
  - Trigger：go to, return to, turn to ...
- **物体との関係 (接触)**
  - Event : hold, grip, catch, pick up ...
  - Trigger : set, go off, put off ...
- **物体・環境の状態**
  - Event : closed door, folded umbrella, a cup filled with water ...
  - Trigger : open the door, unfolds the umbrella, place the cup ...
- **身体部位・身体の状態**
  - Event : left hand raised, left foot injured ...
  - Trigger : raise right hand keeping left hand raised
    walk with a limp in the left leg ...

これらの状態は特に長期動作生成において頻出するため,

- 長期モーションデータセットの構築
- 整合性の評価指標の作成
- 長期モーション生成モデルの提案

が必要である

提案手法

手法概要

Motion-to-Motion で，Encoder-Decoder モデルを学習
　History motion を入力に Future motion を再構成
Encoder 部分を Diffusion に置き換えて,
　History motion と Text から潜在表現の生成を学習

Context Encoder はキューに保存された潜在表現から,
次時刻に取るべき Motion の潜在表現を得るように学習

→ 長期一貫性が保たれるようになることを期待

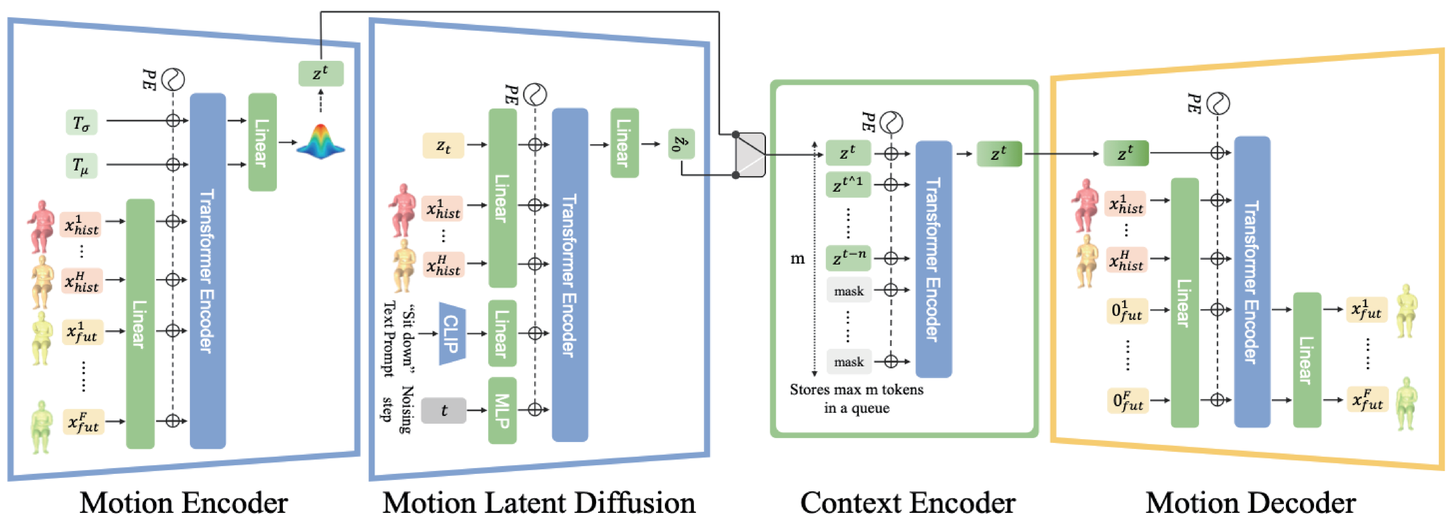

図 2. 提案モデル

データセットの作成

長尺なデータで構成され，時系列的一貫性評価に使用可能である
個別の評価項目によって一貫性の評価における評価揺れを軽減

以下の要素を含む検証用のデータセットを作成

- 約 30s の連続する人物動作
- 全体を説明するテキスト
- 細かい動作を説明するテキスト及び時間区分
- 一貫性を評価するための評価項目

表 1. 既存データセットとの比較

Dataset	Conference	Num Seq.	Total Time	Avg Length	Text/Action Classes
KIT Motion Language	MM 2018	3911	10.3h	9.68s	6.3k Text descriptions, 40 Action classes
UESTC	TPAMI 2019	25.6K	83h	11.66s	6.3k Text descriptions, 40 Action classes
NTU-RGB+D	MM 2020	114.4K	74h	2.33s	120 Action classes
HumanAct12	CVPR 2021	1191	6h	18.14s	12 Action classes
BABEL	CVPR 2022	-	43.5h	-	260 Action classes
HumanML3D	CVPR 2022	14.6K	28.5h	7.02s	44.9K Text descriptions
Ours	-	125	1.04h	**30.0s**	25 Long-term motions, Evaluation aspects

実験

定量評価において既存手法を下回る結果となった
一方で図 3 に示すように姿勢の維持において一定の優位性が確認できた.

表 2. 一貫性評価項目による評価

手法	Preference	平均スコア
DART	75	0.76
提案手法	61	0.64

表 3. 定量評価

手法	Motion				Transition			
	FID↓	MS↓	R-P@1↑	R-P@3↑	FID (Trans.)↓	Div.→	PJ→	AUJ Mean↓
FlowMDM [1]	6.47 ± 0.04	5.14 ± 0.01	0.35 ± 0.00	0.67 ± 0.00	2.52 ± 0.08	6.66 ± 0.10	**0.04** ± 0.00	**0.11** ± 0.00
DART [14]	**5.31** ± 0.10	**4.95** ± 0.00	**0.36** ± 0.00	**0.68** ± 0.00	**2.19** ± 0.05	6.66 ± 0.19	0.08 ± 0.00	0.14 ± 0.00
提案手法	5.47 ± 0.06	5.41 ± 0.03	0.31 ± 0.00	0.61 ± 0.01	2.85 ± 0.10	6.51 ± 0.09	0.06 ± 0.00	0.21 ± 0.00

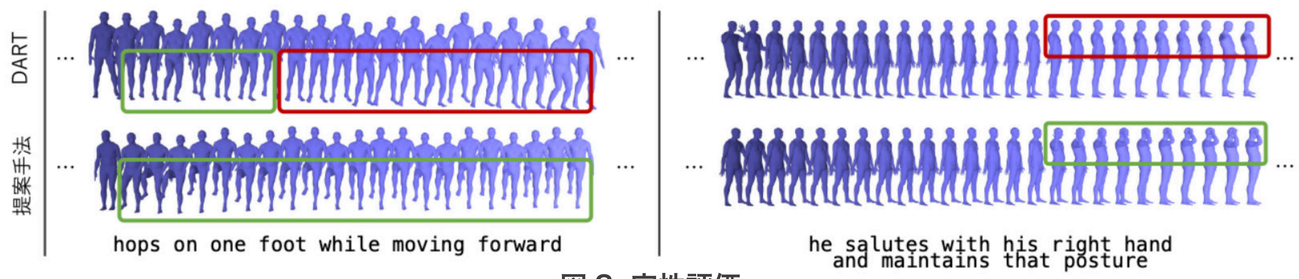

図 3. 定性評価

今後の展望

データセットの拡張

- 自動収集，自動キャプショニングによる，事前学習に十分な量の長期モーションの収集
- 評価項目の自動生成，自動評価

LLM ベースの手法による長期モーション生成

- 長期モーションデータを用いたパラメータ数の大きいモデルでの学習

Contact

天谷幸太郎：akot-ek@keio.jp
加藤駿：kato_shun1329@keio.jp
五十川麻理子：mariko.isogawa@keio.jp