時間的整合性を考慮したテキストからの 2 次元モーション生成

EP22058 川本寛和　　指導教授：山下隆義

1. はじめに

テキストからの 2 次元モーション生成の先行研究として，2CM-GPT[1] が提案されている．2CM-GPT は，3 次元モーション生成モデルと異なり，収集が容易な 2 次元モーションのデータセットを学習に利用できる．そのため，生成に失敗した 2 次元モーションを収集して，データセットを動的に拡張することにより，効果的なファインチューニングが実現できる．一方で，2CM-GPT はいくつかの課題もある．1 つ目は，各フレームのモーションを独立で生成するため，時間的な整合性が低い．2 つ目は，テキストとモーションを対応付けることなく混在して学習するため，テキストとモーションの整合性が損なわれる．これらの課題を解決するために，本研究ではテキストとモーションをアテンション機構で関連付けるとともに，時間的な整合性を考慮する手法を提案する．

2. 2 次元モーション生成

2CM-GPT は，2 次元のモーションを生成する代表的な手法である．2CM-GPT のモデル構造を図 1 に示す．2CM-GPT は，Motion Tokenizer で人間のモーションを離散的なトークンに変換する．そして，テキストも同様に Text Tokenizer でトークンに変換する．これらを連結させた Mixed Tokens を Language Encoder に入力して潜在ベクトルを獲得する．潜在ベクトルをもとに Language Decoder が出力した Output Tokens を Motion Tokenizer に入力して 2 次元のモーションを生成する．

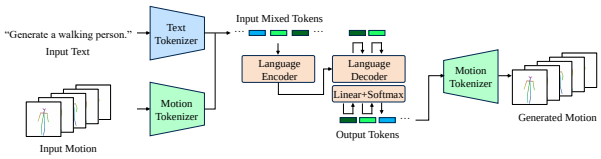

図 1：2CM-GPT のモデル構造

3. 提案手法

本研究では，2CM-GPT の学習アプローチが抱える「時間的な整合性」と「テキストとモーションの整合性」の不一致を解決する手法を提案する．提案手法のモデル構造を図 2 に示す．なお，提案手法がテキストと 2 次元モーションの関係性を学習する過程を Training Phase，テキストから 2 次元モーションを生成する過程を T2M Phase とする．

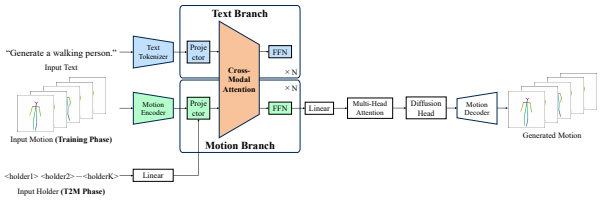

図 2：提案手法のモデル構造

3.1. 学習

Training Phase では，VAE で学習された Motion Encoder を用いて人間の連続的な動作を離散化しない形で潜在空間に埋め込む．この埋め込み特徴を Motion Branch に入力する．テキストは，Text Tokenizer で埋め込み特徴とし，Text Branch に入力する．各ブランチにおいて，埋め込み特徴を Projector に入力し，Query, Key, Value ベクトルを抽出する．そして，Motion Branch と Text Branch の各ベクトルを同じ Cross-Modal Attention に入力して，Motion Branch の Value ベクトルにテキスト特徴を反映させる．Motion Branch の出力を Multi-Head Attention に入力して潜在ベクトルを獲得する．潜在ベクトルを Diffusion Head に入力し，後述の T2M Phase のための逆拡散過程を学習する．損失関数には，生成モーションと実モーションの差を評価する特徴再構成損失，対応関係にあるモーションとテキストがクロスモーダルな特徴空間上で近接するよう制約する分類損失，逆拡散過程における潜在表現の再構成誤差を評価する拡散損失を用いる．

3.2. モーション生成

T2M Phase では，Motion Branch と Text Branch の Cross-Modal Attention によって，Motion Branch に入力した Holder にテキスト特徴を反映させる．Motion Branch の出力を Multi-Head Attention に入力して，潜在ベクトルを獲得する．潜在ベクトルを Diffusion Head に入力して，逆拡散過程によるノイズ除去を行う．その後，VAE で学習された Motion Decoder を用いて，潜在ベクトルから 2 次元のモーションを生成する．

4. 評価実験

2CM-GPT との比較実験により，提案手法の有効性を示す．

4.1. 定量的評価

表 1 より，提案手法は FID が低いことから，2CM-GPT と比べて 2 次元モーション生成精度の向上を確認した．一方，2CM-GPT の Diversity は提案手法よりも高い．その要因として，モーション生成精度が十分でないために，類似の指示文に対しても多様なモーションを生成することが考えられる．

表 1：テキストからの 2 次元モーション生成精度の比較

Method	FID ↓		Diversity ↑	
	real	gen	real	gen
2CM-GPT	-1.37×10^{-9}	32.36	16.96	19.28
提案手法	-5.24×10^{-9}	12.15	16.69	12.92

そこで，多様性の要因を検証するために，表 1 の Diversity と，後述の定性的評価で用いる図 3 の指示文のみを与えて生成されたモーションの Diversity を比較する．表 2 より，2CM-GPT は同一の指示文のみを与えた場合でも Diversity が高い傾向を示したことから，上記の可能性を裏付ける傾向が確認された．

表 2：指示文ごとの生成モーションの Diversity の比較

Method	Multi Instruction	Single Instruction
2CM-GPT	19.28	18.85
提案手法	12.92	6.07

4.2. 定性的評価

2CM-GPT と提案手法に同じ指示文を与えて生成させた 2 次元モーションを図 3 に示す．図 3 の手の動きに注目すると，提案手法の生成モーションは 2CM-GPT と比較して，指示文の動作内容を正確に反映していると判断できる．

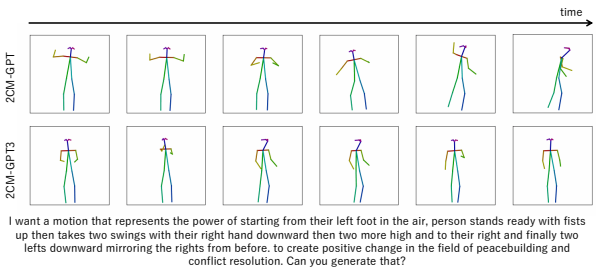

図 3：テキストからの 2 次元モーション生成結果の可視化

5. おわりに

本研究では，2CM-GPT と提案手法の評価実験を行い，提案手法の有効性を示した．今後は，提案手法で生成したモーションを用いてポーズ誘導による人物動画生成を実施し，実用性を検証する．

参考文献

[1] R. Inoue, *et al.*, "2D Motion Generation Using Joint Spatial Information with 2CM-GPT", VISIGRAPP, vol.2, pp.582-590, 2025.