個々の役割を指示可能な 入力言語に応じた 2 人のインタラクションの動作生成の検討

田中 幹大　　近藤 雅芳　　藤原 研人
LINE 株式会社
{mikihiro.tanaka, masayoshi.kondo, kent.fujiwara}@linecorp.com

1 概要

本研究では言語を入力とした，2 人のインタラクションを伴う動作生成に取り組む．既存研究では言語と動作の関係を学習し動作生成を行なってきた．よって，インタラクションを生成する際には 2 人分の動作を単一の記述で表現する必要がある．しかし，一部動作においてこの記述には一方は能動態，他方は受動態で表現される非対称な関係性を内包している．この見識に基づき本研究では，インタラクションの記述を受動態・能動態の言語に変換することによって，主体と受け手双方の動作を生成する手法を提案する．また，両者の位置や動作の関係性の考慮のために，動作間のクロスアテンションも導入する．更に，非対称な動作の場合に，一般に 2 人の動作の学習データ全てに主体・受け手のラベリングが必要となる．本研究では少数のアノテーションを用いて学習データ全体にこのラベリングを行う手法も提案する．既存手法を拡張する方法と比較し，文章を解釈し主体と受け手に分けて動作生成を行う提案手法の有効性を実験により示した．

2 はじめに

人間の動作のモデリングは，3 次元アバターやキャラクターを利用した高品質な動画像の制作において重要な要素になりつつある．このような動画制作では 2 人以上の人物を登場させる場合，1 人の時と比べて，個々の動作のモデリングに加えて，両者の位置関係やインタラクションを考慮した，より複雑なモデリングが必要となる．この複数人の動作を自在に作ることができれば，より多様な登場人物からなる映像制作を容易にできる．そこで本研究では，最も直感的なインターフェースの一つである言語に注目し，入力言語に応じた 2 人のインタラクションの動作生成という課題に取り組む．

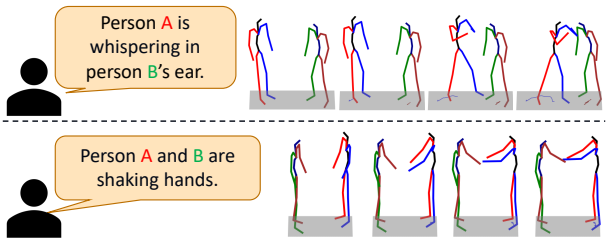

図 1: 言語入力で提案手法により生成されたインタラクション例を示す．上は非対称な動作例であり，話す側と聞く側が存在する．下は対称な動作例であり，両者共通の握手という動作をしている．

近年言語入力による 1 人の動作生成の研究が進展してきている [1, 2]．特に最近では拡散モデル [3, 4, 5] の発展に伴って，より言語に忠実かつ多様な動作生成が可能になってきた [6, 7]．ここで，2 人のインタラクションについて考えると，図 1 のように主体と受け手が存在する非対称な動作，また両者共に主体となって共通の行為を行う対称な動作が存在する．図 1 上の非対称な例では，2 人のインタラクションを描写した文章は，一方は能動態の囁くという表現，もう一方は受動態の囁かれるという表現で描写される関係性を内包している．[8] は 2 人の動作生成に取り組む研究だが，動作の主体と受け手の関係性を考慮しておらず，個々の役割を指示することも考えられていなかった．

そこで本研究では，先行研究の 1 人の動作生成モデル [6] を拡張し，2 人のインタラクションの対称・非対称性を考慮して生成する手法を提案する．まず，動作の対称・非対称性を扱うために，個々に対して動作が非対称な時は受動態・能動態の異なる言語を入力とし，動作が対称な時は同じ言語を入力とする．更に，両者の位置や動作の関係性を考慮するために，2 人の動作間のアテンションも導入する．

ここで，非対称な動作の生成において図 1 のようにどちらを主体とするか指示するためには，2 人まとめた動作の説明文に加えて，どちらの動作が主

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

体・受け手かのラベリングが必要となる．本研究では，アノテーションコストを抑えるべく，音源分離で知られる Permutation Invariant Training[9] を利用することで，後者の主体・受け手のアノテーションは少数のみを用いて，学習データ全体に擬似的にラベル付けするパイプラインを提案する．

3 関連研究

3.1 動作の認識

人間の動作の解析は，これまでは主に一般的な動画像を用いて行われてきた [10, 11]．しかし，近年の動作データ取得方法の発展に伴い [12, 13]，骨格の動作データが注目を集めている．背景の影響を受けない骨格データによる動作の認識は，現在活発に研究されているテーマの一つである [14, 15]．骨格の関節の座標値をベクトルデータ [16] や 2 次元のグリッドデータ [17] に変換したり，関節の連結性をグラフとして扱ったりして [18] 深層学習のモデルの入力として利用されている．さらには，実施された地点も特定する研究も行われている [19]．

動作解析の研究の多くは 1 人の人物を対象にしているが，近年になり，人同士のインタラクションの一般的なベンチマークとして，NTU RGB+D120 データセット [20] が構築された．このデータセットを用いて，骨格の動作データを用いた 2 人のインタラクションの認識の研究が始まっている [21, 22]．

3.2 動作の生成

骨格データの解析結果を活用し，新たな動作を生成しようとする研究も盛んに行われている．従来，動作生成は様々なパターンから最も整合性のある動作を接続することで行われてきたが [23]，深層学習の発展に伴い，データから学習することが現在主流になっている．行動ラベルが与えられた際に適切な動作を生成する方法や [24, 25]，音楽にあった動作を生成する研究 [26]，動作の周期性に着目し生成する手法 [27] などが提案されている．

CLIP[28] の登場に伴い，動作生成の研究においても，直感的なインターフェースとして言語が注目されている．CLIP の言語画像空間と動作の特徴空間を対応づけ，言語情報に合った動作データを生成する手法として MotionCLIP[29] が提案された．また，近年登場したテキストと動作の大規模なデータセット [30, 31]，及び拡散モデルと学習済み CLIP を活用

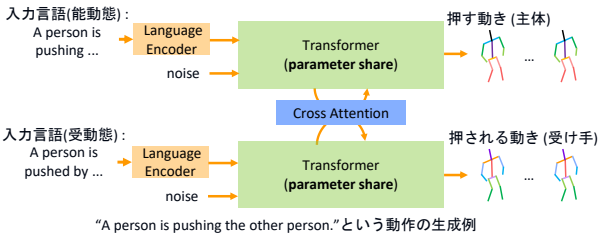

図 2: 提案モデルの図を示す．「押す」という 2 人の動作を生成する場合は，受動態・能動態の言語をそれぞれ入力し，対応した動作を出力する．対称な動作の場合は同じ言語を入力する．

することで，入力言語に忠実かつ多様な動作の生成を実現する手法も多く提案され始めている [6, 7]．本研究に近い研究として [8] らは 2 人の動作生成に取り組んでいるが，動作の対称・非対称性に注目し，主体と受け手といった役割を個々に指示して生成を行うのは本研究が初めての取り組みである．

4 提案手法

本研究では以下を満たすモデルを提案する．

- 対称な動作と非対称な動作ともに，個々の役割を指定してインタラクションを生成できる．
- 生成された 2 人の動作の辻褄が合っている．

2 人のインタラクションに対し，その動作カテゴリ c_i を説明する文章 y_i が付与されている状況を想定する．非対称な動作のカテゴリ c_i においては，本来個々の動作は，y_i を受動態・能動態の 2 つの描写に翻訳した文章 $(y_i^{passive}, y_i^{active})$ のどちらに相当するかのアノテーションが必要となる．本研究では，アノテーションコストを抑えるべく，少数のアノテーションから個々の動作が $(y_i^{passive}, y_i^{active})$ のどちらに相当するかラベリングを行う手法を提案する．

4.1 拡散モデルを用いた 1 人の動作生成

まず，言語から 1 人の動作生成で最も性能の良いモデルの一つである，[6] らの拡散モデルを用いた手法を説明する．動作生成は，ランダムにサンプリングされたノイズからの逆拡散過程として定式化することができる．動作系列を $X^{(0)} = [x_1, ..., x_F]$ (F はフレーム数，$x_1, ..., x_F$ は各時刻の姿勢) とした時，実データからサンプリングした動作系列 $X^{(0)} \sim q(X^{(0)})$ に対して，T ステップのノイズを順に加えたものを $X^{(1)}, ..., X^{(T)}$ のように表記する．まず，拡散過程では以下の式のようにガウシアンノイ

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

ズをマルコフ連鎖に従って付与する形で表せ，$X^{(T)}$ はおよそ $\mathcal{N}(\mathbf{0},\mathbf{I})$ に従う．(β_t はハイパーパラメータ)

$$q(X^{(1:T)}|X^{(0)}) = \prod_{t=1}^{t=T} q(X^{(t)}|X^{(t-1)})$$
$$q(X^{(t)}|X^{(t-1)}) = \mathcal{N}(X^{(t)}; \sqrt{1-\beta_t}X^{(t-1)}, \beta_t\mathbf{I}) \quad (1)$$

逆拡散過程ではランダムなノイズから始まり，推定された $\mu_\theta(X^{(t)},t), \Sigma_\theta(X^{(t)},t)$ に従うガウシアンノイズを付与する形でマルコフ連鎖に従ってノイズを除去していき，動作を生成する．

$$p_\theta(X^{(0:T)}) = p_\theta(X^{(T)}) \prod_{t=1}^{t=T} p_\theta(X^{(t-1)}|X_t)$$
$$p_\theta(X^{(t-1)}|X^{(t)}) = \mathcal{N}(X^{(t-1)}; \mu_\theta(X^{(t)},t), \Sigma_\theta(X^{(t)},t)) \quad (2)$$

[6] らは，$\Sigma_\theta(X^{(t)},t)$ は定数を用い，$\mu_\theta(X^{(t)},t)$ の推定のみ行なった．推定器には Transformer[32] をベースとしたモデルを用いた．$\mu_\theta(X^{(t)},t)$ は推定されたノイズ $\epsilon_\theta(X^{(t)},t,text)$ から計算できるため，ノイズ ϵ とステップ数 t，データ $X^{(0)}$ をランダムにサンプリングし，以下の式によってノイズ推定を学習することで，モデルパラメータの最適化を行った．

$$\mathcal{L} = E_{t\in[1,T],X^{(0)}\sim q(X^{(0)}),\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\|\epsilon - \epsilon_\theta(X^{(t)},t,text)\|] \quad (3)$$

4.2 インタラクション生成への拡張

2 人の動作の系列を $X_1 = [x_0^1,...,x_F^1], X_2 = [x_0^2,...,x_F^2]$ (F はフレーム数，x_0 は最初の位置や体の向き，$x_1,...,x_F$ は各時刻の姿勢) のように表す．個々の動作に対して，文章が付与されているデータセット $((X_1,y_1),(X_2,y_2))$ がある場合について考える．先行研究の単純な拡張によって 2 人のインタラクションを生成する方法として，一つの Transformer で 2 人分の動作を一度に生成する方法が考えられる．しかしこれでは，非対称な動作を扱う際に個々の役割を指示することができない．そこで，個別の動作をモデリングするために，図 2 のようにパラメータを共有した 2 つの Transformer を用い，非対称な動作の場合はそれぞれ受動態・能動態の言語を入力することにする．対称な動作については，双方共に能動態の言語を入力する．

次に，インタラクションでは互いの位置関係や動作のタイミングを合わせた生成が必要となる．本研究では互いの動作間のクロスアテンションを導入する．これにより，作用・反作用といった動作の関係性や，相手に合わせた協調的な動作の学習を図る．

最終的なロス関数は式 3 と同様に，以下で表せる．

$$\mathcal{L} = E_{t\in[1,T],(X_1^{(0)},X_2^{(0)})\sim q(X^{(0)}),\epsilon_1\sim\mathcal{N}(\mathbf{0},\mathbf{I}),\epsilon_2\sim\mathcal{N}(\mathbf{0},\mathbf{I})} [\|\epsilon_1 - \epsilon_\theta(X_1^{(t)},t,y_1)\| + \|\epsilon_2 - \epsilon_\theta(X_2^{(t)},t,y_2)\|] \quad (4)$$

生成時は式 2 に従って，T ステップ分ノイズを除去していくことで，同時に 2 人の動作を生成する．

4.3 主体・受け手の擬似ラベルの付与

4.2 節の手法のアノテーションコストを削減するべく，非対称な動作カテゴリ c^i における個々の動作が，$(y_i^{passive}, y_i^{active})$ のどちらに相当するかのラベリングを半自動的に行う手法を説明する．

ここで，非対称なクラス c^i について，二つの学習可能なパラメータ w_1^i, w_2^i を用意する．式 4 において，言語特徴の代わりに w_1^i, w_2^i を用いて 2 通りのガイダンスによってノイズを推定し，以下の式のように小さい方を選択して誤差逆伝播を行う．

$$\mathcal{L} = E_{t\in[1,T],(X_1^{(0)},X_2^{(0)})\sim q(X^{(0)}),\epsilon_1\sim\mathcal{N}(\mathbf{0},\mathbf{I}),\epsilon_2\sim\mathcal{N}(\mathbf{0},\mathbf{I})}$$
$$[\min(\|\epsilon_1 - \epsilon_\theta(X_1^{(t)},t,w_1^i)\| + \|\epsilon_2 - \epsilon_\theta(X_2^{(t)},t,w_2^i)\|,$$
$$\|\epsilon_1 - \epsilon_\theta(X_1^{(t)},t,w_2^i)\| + \|\epsilon_2 - \epsilon_\theta(X_2^{(t)},t,w_1^i)\|)] \quad (5)$$

これは音源分離の分野で用いられる Permutation Invariant Training[9] と同様の手法であり，これによって w_1^i, w_2^i は主体か受け手のどちらか一方の特徴を得るように学習することが期待できる．学習した w_1^i, w_2^i のどちらが主体・受け手に相当するかは，少数のラベル付きデータと照合することで確かめることができる．そして最後に，学習データそれぞれにランダムなノイズを付与し，X_1, X_2 のどちらが w_1^i, w_2^i に対応するかを式 5 を用いて計算することで，擬似的に個々の動作が主体か受け手かを特定し，$(y_i^{passive}, y_i^{active})$ をそれぞれに付与できる．

5 実験

5.1 実験設定

データセット: NTU RGB+D120 データセット [20] から，2 人のインタラクションを扱っている 26 クラスの動作を利用した．9 クラスの対称な動作，17 クラスの非対称な動作が含まれている．Appendix A にクラスの例を載せる．これらの動画像から 3 次元骨格

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

を BEV[33] によって推定して実験に用いた．登場する被験者が異なるように学習/検証/テストデータを分け，それぞれ 20,306，3,493，3,044 件利用した．

評価指標: 1 人の動作生成の先行研究に倣って以下の評価指標によって定量評価を行う．生成された動作が入力言語と対応しているかの正解率 (Accuracy) を用いる．また，生成されたものと真のデータの分布の一致具合によって評価する，Frechet Inception Distance(FID) を用いる．そして，生成される動作の多様性 (Diversity) を評価する．これらの評価を行うには，インタラクションの動作の特徴を抽出するモデルが必要となる．しかしインタラクション認識の標準的なモデルはないため，本研究では同じ学習データでインタラクションを認識する Transformer をベースとしたモデルを学習させて評価に用いた．

これらに加えて，本研究では生成された 2 人の動作が相互に矛盾のないものになっているかを表す，MutualConsistency と呼ぶ指標を提案する．より詳細を AppendixC で説明する．

比較手法: まず，提案モデルで 4.3 節の手法により，個々の動作に対して言語ラベルを付与する．このデータを用いて以下の手法により比較実験を行う．

[6] らの手法を個別の役割を指示可能な形で拡張し，2 つの Transformer を用いて動作を行う手法をベースラインとし，Ours w/o cross attention と記す．Ours と記す手法はこれに対して 2 人の動作間のクロスアテンションを加えたものである．更に，提案手法は 1 人の動作生成手法である [6] らのモデルに新たなパラメータを加える形で設計されており，[6] らの言語特徴抽出器や動作生成モデルの共通部分のパラメータを初期値として学習することができる．これを Ours+pretrained と記す．

また，[6] らの手法を基にして，2 人分の動作を一つの系列として扱うことで一つの Transformer で生成する手法を Single Transformer と記す．この手法は [8] 同様 2 人まとめた指示しかできず，個別に役割を指示できないため，参考値として比較を行う．

5.2 定量評価

表 1 に，定量評価の結果を示す．提案手法はベースラインの Ours w/o cross attention に比べていずれの指標でも良い性能を示し，特に Accuracy と MutualConsistency で差が出る結果となった．Ours w/o cross attention は Accuracy が 71.3%あるが，MutualConsistency が 41.7%となった．これはそのカテゴリらしい動作をある程度生成できているが，生成されたインタラクションの半分以上が辻褄の合わないものになっていることを示している．個々に付与した擬似ラベルを用いない Single Transformer と比べても，特に 1 人の言語と動作の対応関係の事前知識を用いる Ours+pretrained の Accuracy が高く，個別の動作のモデリングによってより入力言語に忠実なインタラクションを生成できていることが分かる．Single Transformer の FID が小さいのは，擬似ラベルのエラーの影響を受けないためと考える．

表 1: 定量評価の結果を示す．生成する動作の系列長は可変であり，テストデータと同じ長さとした．4 回評価実験を行い，95%信頼区間を ± で示した．

Methods	Accuracy↑	FID ↓	Diversity→	MutualConsistency↑
Ground Truth	$84.6^{\pm 0.1}$	$0.003^{\pm 0.00}$	$30.20^{\pm 0.24}$	$99.8^{\pm 0.0}$
Single Transformer	$78.7^{\pm 0.8}$	$6.67^{\pm 0.58}$	$29.70^{\pm 0.67}$	$99.9^{\pm 0.1}$
Ours w/o cross attention	$71.3^{\pm 2.0}$	$17.78^{\pm 1.00}$	$29.03^{\pm 0.52}$	$41.7^{\pm 2.3}$
Ours	$78.9^{\pm 0.7}$	$9.62^{\pm 0.21}$	$29.39^{\pm 0.59}$	$98.7^{\pm 0.2}$
Ours+pretrained	$\mathbf{84.1}^{\pm 1.2}$	$\mathbf{9.41}^{\pm 0.59}$	$\mathbf{29.80}^{\pm 0.56}$	$\mathbf{99.0}^{\pm 0.1}$

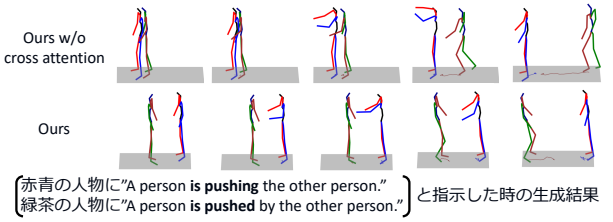

図 3: 生成結果を示す．提案手法は「押す」という動作で，辻褄の合った動作を生成できている．

5.3 定性評価

図 3 に「押す」というインタラクションを生成した時の結果を示す．Ours w/o cross attention に比べて，提案手法は一方が手を伸ばして相手を押し，もう一方は押された結果として後ろによろける動作が生成されている．このように，定性的にも提案手法は言語に忠実かつ 2 人の辻褄の合ったインタラクションが生成できていることを確認できる．

6 おわりに

本研究では，個々に役割を指示可能な，言語による 2 人のインタラクション生成に取り組んだ．互いの動作間のクロスアテンションを導入し，動作の関係性を考慮したモデルを提案し，言語に忠実かつ相互に矛盾のない 2 人のインタラクションを生成可能であることを実験によって示した．物理環境の考慮及び，より自由度の高い言語入力によるインタラクション生成が今後の課題として挙げられる．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

参考文献

[1]Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *ECCV*, 2022.

[2]Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *CVPR*, 2022.

[3]Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[4]Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *NeurIPS*, 2020.

[5]Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *NeurIPS*, 2021.

[6]Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022.

[7]Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Amit H Bermano, and Daniel Cohen-Or. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

[8]Shubh Maheshwari, Debtanu Gupta, and Ravi Kiran Sarvadevabhatla. Mugl: Large scale multi person conditional action generation with locomotion. In *WACV*, 2022.

[9]Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen. Multi-talker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *TASLP*, 2017.

[10]Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019.

[11]Yutaro Shigeto, Yuya Yoshikawa, Jiaqing Lin, and Akikazu Takeuchi. Video caption dataset for describing human actions in japanese. In *LREC*, 2020.

[12]Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Transactions On Graphics*, 2019.

[13]Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *CVPR*, 2020.

[14]Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *CVPR*, 2014.

[15]Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*, 2015.

[16]Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *ECCV*, 2016.

[17]Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3d action recognition. In *CVPR*, 2017.

[18]Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, 2019.

[19]Qing Yu and Kent Fujiwara. Frame-level label refinement for skeleton-based weakly-supervised action recognition. In *AAAI*, 2023.

[20]Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. In *TPAMI*, 2020.

[21]Yunsheng Pang, Qiuhong Ke, Hossein Rahmani, James Bailey, and Jun Liu. Igformer: Interaction graph transformer for skeleton-based human interaction recognition. In *ECCV*, 2022.

[22]Yoshiki Ito, Quan Kong, Kenichi Morita, and Tomoaki Yoshinaga. Efficient and accurate skeleton-based two-person interaction recognition using inter- and intra-body graphs. In *ICIP*, 2022.

[23]Michael Büttner and Simon Clavet. Motion matching-the road to next gen animation. *Proc. of Nucl. ai*, 2015.

[24]Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[25]Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *ICCV*, 2021.

[26]Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *CVPR*, 2021.

[27]Sebastian Starke, Ian Mason, and Taku Komura. Deepphase: periodic autoencoders for learning motion phase manifolds. *ACM Transactions on Graphics (TOG)*, 2022.

[28]Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[29]Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. *ECCV*, 2022.

[30]Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 2016.

[31]Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *CVPR*, 2021.

[32]Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[33]Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J. Black. Putting people in their place: Monocular regression of 3d people in depth. In *CVPR*, 2022.

[34]Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019.

[35]Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

A　データセットの詳細

対称な動作は，hugging，shaking hands，walking towards，walking apart，high-five，cheers and drink，carry object，exchange things，rock-paper-scissors の 9 カテゴリ，非対称な動作は，punch/slap，kicking，pushing，pat on back，point finger，giving object，touch pocket，hit with object，wield knife，knock over，grab stuff，shoot with gun，step on foot，take a photo，follow，whisper，support somebody の 17 カテゴリからなる．本研究ではこれらを文章に変換して用いた．

B　提案モデルの詳細

図 4 にモデルの詳細な図を示す．モデルの入力の initial position は最初の体の位置と向きを表す 4 次元のベクトルである．pose representation は $(r^{va}, r^{vx}, r^{vz}, r_h, \mathbf{j}^p, \mathbf{j}^v, \mathbf{j}^r, \mathbf{f})$ からなる，263 次元のベクトルである．y 軸を垂直方向とした時，$r^{va}, r^{vx}, r^{vz}, r_h$ はそれぞれ腰の，y 軸回りの角速度，x 方向の速度，z 方向の速度，高さを表す．$\mathbf{j}^p \in \mathbb{R}^{(J-1)\times 3}, \mathbf{j}^v \in \mathbb{R}^{J\times 3}$ は J 個の関節の位置と速度を表す．$\mathbf{j}^r \in \mathbb{R}^{(J-1)\times 6}$ は各関節の 6 次元の回転 [34] を表す．$\mathbf{f} \in \mathbb{R}^4$ は足の 4 点の関節が地面に接しているかを表している．拡散のステップ数は 1,000 とし，式 1 の β_t は，0.0001 から 0.02 に線形に増やした．

C　評価指標の詳細

Accuracy: 生成された動作が入力された言語に対応するかを評価する指標．図 5 に示す，入力されたインタラクションのカテゴリを予測するモデルを学習させ，評価に用いる．以下 FID，Diversity の評価にも，このモデルによって抽出された特徴量を用いる．本研究では，図 5 の Global Average Pooling の後の出力を特徴量として用いた．

FID: 生成モデルの評価で最も用いられる，生成されたものと真のデータの分布の一致具合を評価する指標．生成された動作とテストデータの動作の特徴量の平均と共分散をそれぞれ $(\mu, \Sigma), (\mu', \Sigma')$ と表したときに，以下の式により分布間距離を求める．

$$\mathrm{FID} = \|\mu - \mu'\|_2 + \mathrm{Tr}(\Sigma + \Sigma' - 2\sqrt{\Sigma\Sigma'}) \tag{6}$$

Diversity: 生成される動作の多様性を評価する指標．ランダムな文章入力から生成された S_d 個の動作の集合を二つ用意し，それぞれから抽出された特

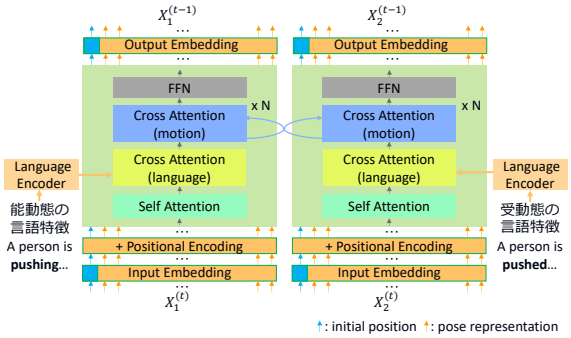

図 4: 詳細な提案モデルの図を示す．パラメータを共有する transformer は N 個のアテンションブロックによって構成され，1 ステップずつノイズを除去し 2 人の動作を生成する．アテンションブロック内では順に，セルフアテンション，言語のクロスアテンション，動作間のクロスアテンション，順伝播型ニューラルネットワークによって処理が行われる．

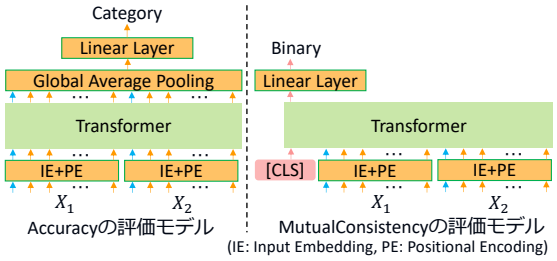

図 5: Accuracy と MutualConsistency の評価に用いたモデルの図を示す．IE と PE は図 4 と同様．

徴量 $[v_1, ..., v_{S_d}]$，$[v'_1, ..., v'_{S_d}]$ を用いて，以下の式によって生成される動作の多様性を評価する．

$$\mathrm{Diversity} = \frac{1}{S_d}\sum_{i=1}^{S_d} \|v_i - v'_i\|_2 \tag{7}$$

MutualConsistency: 本研究で提案する，生成された 2 人の動作が辻褄の合ったものになっているかを評価する指標．図 5 に示すモデルを学習させ，評価に用いる．学習時は BERT[35] で用いられる next sentence prediction タスクの学習と同様に，[CLS] と呼ぶトークンを用意し，そこからの出力を用いて動作が正しいペアかを予測させる．例えば押すという動作において，誤ったペアでは，押す・押されるの方向やタイミング，位置関係が誤っているデータとなるため，正しいペアとの判別を学習することにより，入力された 2 人の動作が辻褄の合っているものかを判定できるようになる．本研究では正しいペアと誤ったペアを 1 対 1 の比率で学習させた．このモデルを用いて，生成された 2 人の動作が正しいペアとなっている割合によって評価する．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).