スタイルを条件指定可能な音楽に合わせたダンス生成技術

石井 亮[1,a)]　永徳 真一郎[1]　伏尾 佳悟[1]　松尾 翔平[1]

概要：音楽とどのようなダンススタイルでダンスを生成するかを指定することで，指定された音楽とスタイルに合った新しいダンスの振付を自動生成可能なダンス生成技術 SDCGM（Style-specified Dance Choreography Generation from Music）を提案する．本技術は，音楽とスタイル情報を条件入力とする Transformer ベースの条件付き拡散モデルのフレームワークによって実現される．SDCGM を実装し，評価実験を実施した結果，従来の最先端モデルと比較して，SDCGM モデルで生成されたダンスは，自然さに関する品質が高く，さらに指定されたスタイルを適切に反映したダンスを生成できることが示唆された．さらに，アプリケーションのプロトタイプの実装とユーザ評価の初期検討を通じて，SDCGM が提供するダンス生成機能は，振付をおこなうプロダンサーにとって，ダンスの習熟や振付の支援に有益である可能性が示唆された．SDCGM を利用することで，誰もが好きな音楽とスタイルを指定して，魅力的な様々なダンスを生成することが可能になる．

1. まえがき

ダンスは，身体の動きを通じてコミュニケーションを図り，自己表現を行う行為である．ダンスは長い歴史を持ち，世界のほとんどの国や文化圏で見られ，ダンスのスタイルは非常に多様で独自の魅力的な表現方法を持つ．エンターテイメントとしてのダンスは，ダンサー自身だけでなく，観客に楽しんでもらえるものであり，現代ではショーやライブパフォーマンスに欠かせない存在となっている．また，動画配信サービスや SNS を通じて，私たちは日々，様々なダンスパフォーマンスの視聴を楽しむことができ，さらに自身のダンスを気軽に発信することも盛んにおこなわれている．このように，ダンスは人々にとって非常に身近で欠かすことのできない魅力的なエンターテイメントである．多くの人は，お気に入りの音楽（楽曲）やダンサーがいるであろう．時には，お気に入りのヒップホップのミュージシャン（ダンサー）に，カントリーミュージックなど全く異なるジャンルの音楽に合わせて踊る様子を見てみたいと思うこともある．また，ダンサーが魅力的なダンスの振付を創作することは，経験豊富なダンサーにとっても非常に難しい作業であり，一人で考えつくアイディアには限界がある．ダンスの振付において，あらゆる音楽に対して好みのダンススタイルで多くのダンスの振付を AI に自動生成させてシミュレーションすることができれば，振付におけるアイデアを大幅に広げることができ，振付にかかる膨大な作業量を軽減するだけではなく，本人だけでは考えられなかったより新しい振付を創造することの支援が可能であると考えられる．このような支援は，振付をおこなうダンサーに対してだけではなく，ダンスに馴染みのない一般ユーザにとっても，気軽にダンスの振付の創作活動を楽しむことができ，非常に有用であると考えられる．

このような背景を踏まえ，我々は，音楽とどのようなダンススタイルでダンスを生成するかを指定することで，音楽とダンススタイルに合ったダンスの振付を自動生成可能なダンス生成技術 SDCGM（Composition-specified Dance Choreography Generation from Music）を提案する．本技術は，Transformer ベースの条件付き拡散モデルのフレームワークをダンス生成に応用することによって実現される．本研究の貢献は以下の通りである．

- 最先端の Transformer ベースの拡散モデルを用いた，ユーザが選択した任意の音楽とダンススタイルを条件指定可能な新しいダンス生成技術を提案する．提案技術は原理的に，ダンススタイルとしてダンサー個人やダンスのジャンルなど，様々な粒度のスタイルを任意に指定可能である．本研究では，特にダンサー個人のスタイルを扱う．
- 客観評価と主観評価の結果から，提案するダンス生成技術は，指定されたダンススタイルを適切に反映したダンス振付を生成することが可能であり，さらに，生成されるダンスの自然さについての品質は先行研究の最先端モデルよりも高いものであることを示す．
- 提案技術を搭載したアプリケーションのプロトタイプ

[1] 日本電信電話株式会社 人間情報研究所
[a)] ryoct.ishii@ntt.com

を提案し，ユーザによる初期評価を通じてこの技術の実用での利用の可能性を示す．

2. 関連研究

音楽に合わせたダンス映像の生成を自動生成する試みは多くなされている．例えば，既存のダンス映像を再利用して新規の音楽に合った新たなダンス映像を生成する試みがある [1]．また，音楽を入力として，人間のような自然なダンスを CG キャラクタに動作させるために，ダンスの振付を自動生成する技術が盛んにおこなわれている．初期のアプローチでは，事前に登録されたダンスの動きから入力音楽にマッチするダンスを検索しつなぎ合わせるというものであった [2], [3], [4]．しかし，このアプローチでは，多様で複雑かつ滑らかな高品質な人間のダンスを生成することは困難であった．近年，大規模なダンスデータセットを活用し，機械学習を利用した生成モデルの提案がされている [5], [6], [7], [8], [9], [10], [11], [12], [13]. 特に，深層学習を用いた生成モデルによって，徐々に人間のような高品質なダンス生成ができるようになっており，目覚ましい発展を遂げている [5], [10], [14], [15].

現在の最先端のダンス生成技術は，拡散モデルに基づく生成ネットワークと，強力な最先端の音楽特徴抽出器である Jukebox [16] を使用した，Editable Dance Generation From Music（EDGE） [6] である．Jukebox は，100 万曲でトレーニングされた大規模な GPT スタイルのモデルであり，音楽の予測タスク有用な強力な特徴表現を出力可能である [17], [18]. 拡散モデル（Diffusion models） [19], [20] は，元データにノイズを加えて全体をノイズ状態にする処理を逆に実行することでデータ分布を学習する深層学習ベースの生成モデルの一種である．近年，様々な画像や動画の生成タスクにおいて最先端の他の手法を上回る性能を実現し，生成モデリングの有効な手法として大きな期待が寄せられている [21], [22]．このような背景から，SDCGM においても，EDGE と同様に拡散モデルに基づくダンス生成ネットワークのアーキテクチャを基盤として活用する．

また，上述したように，音楽からダンスの振付を高い品質で生成できる生成技術が提案されているが，これらのモデルでは音楽に対して生成されるダンススタイルを考慮されていない．つまり，これまでのモデルでは，音楽に対して生成されるダンス振付の内容は，学習データに含まれるダンススタイルをもとに，完全にランダムなスタイルによるダンス生成がなされていた（スタイルを指定するという機構が無いため，生成結果にいずれかのダンスのスタイルが必ずしも保持される保証もない）．そのため，ユーザが好みのダンススタイルでダンスを生成させることは実現されていなかった．我々の提案する SDCGM の独自性は，ユーザが任意のダンススタイルを条件指定し，そのスタイルにマッチしたユニークで自然なダンスを自動生成できる点にある．

3. SDCGM（Style-specified Dance Choreography Generation from Music）

3.1 アーキテクチャ

SDCGM のネットワークアーキテクチャを図 1 に示す．このネットワークは，2 章にて前述した EDGE [6] で用いられた最先端の拡散モデルをベースにしている．我々の提案する SDCGM と EDGE の最も大きな違いは，SDCGM はダンススタイルの条件をネットワークへの入力として使用している点である．

SDCGM の主な入力情報は，拡散処理のための条件付け情報として利用される，音楽とダンススタイルの 2 つである．音楽情報はフリーズされた Jukebox モデル [16] を用いて埋め込まれ，ダンススタイルは one-hot ベクトルとして表現される．ダンススタイル情報は，設計者が様々な粒度で任意のスタイルを設計可能である．なお本研究では，ダンススタイルとして，ダンサー個人の粒度でスタイルを定義した．すなわち，どのダンサーのスタイルでダンスを生成するかのダンサー ID をスタイル情報とする．具体的に，ダンススタイルの表現ベクトルは，ダンサー数 N の次元数を持つ one-hot ベクトルとして設計した．拡散タイムステップ情報は，音楽条件と Feature-wise linear modulation（FiLM [23]）を組み合わせたトークンに組み込まれている．

オブジェクト $\boldsymbol{x}$ は，24 の関節の SMPL フォーマット [24] の姿勢シーケンスであり，各関節は 6 自由度回転表現と 1 自由度の並進表現を使用する．

拡散モデルの処理として，DDPM [19] の定義に従い，拡散を潜在変数 $\{\boldsymbol{z}_t\}_{t=0}^{T}$ を持つマルコフノイズプロセスと見なす．これは，データ分布からサンプリングされた $\boldsymbol{x} \sim p(\boldsymbol{x})$ の順方向ノイズプロセス $q(\boldsymbol{z}_t|\boldsymbol{x})$ に従う．順方向ノイズプロセスは次のように定義される．

$$q(\boldsymbol{z}_t|\boldsymbol{x}) \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}, (1-\bar{\alpha}_t)\boldsymbol{I}), \tag{1}$$

ここで，$\bar{\alpha}_t$ は単調減少スケジュールに従う定数であり，$\bar{\alpha}_t$ が 0 に近づくにつれ，$\boldsymbol{z}_T$ は $\boldsymbol{I}$ の分布に従う近似値として $\boldsymbol{z}_T \sim \mathcal{N}(0, \boldsymbol{I})$ と表すことができる．

音楽条件 c_M とダンススタイル条件 c_S は，モデルパラメータ θ を使用して，すべての t に対してを推定することを学習し，さらに目的関数 $\mathcal{L}_s$ を θ を最適化することで，順方向の拡散プロセスを反転させる．

$$\mathcal{L}_s = \mathbb{E}_{x,t}\left[\| \boldsymbol{x} - \hat{x}_\theta(\boldsymbol{z}_t, t, \boldsymbol{c_M}, \boldsymbol{c_S}) \|_2^2 \right]. \tag{2}$$

モデルはノイズを含むシーケンス $\boldsymbol{z}_T \sim \mathcal{N}(0, \boldsymbol{I})$ を受け取った後，最終シーケンス $\hat{\boldsymbol{x}}$ を推定し，それをノイズ除去して $\hat{\boldsymbol{z}}_{\boldsymbol{T-1}}$ へと戻す．これを，$t = 0$ になるまで繰り返し実施する．拡散ステップのネットワークは，Transformer の Decoder アーキテクチャに基づいており，Cross-Attention

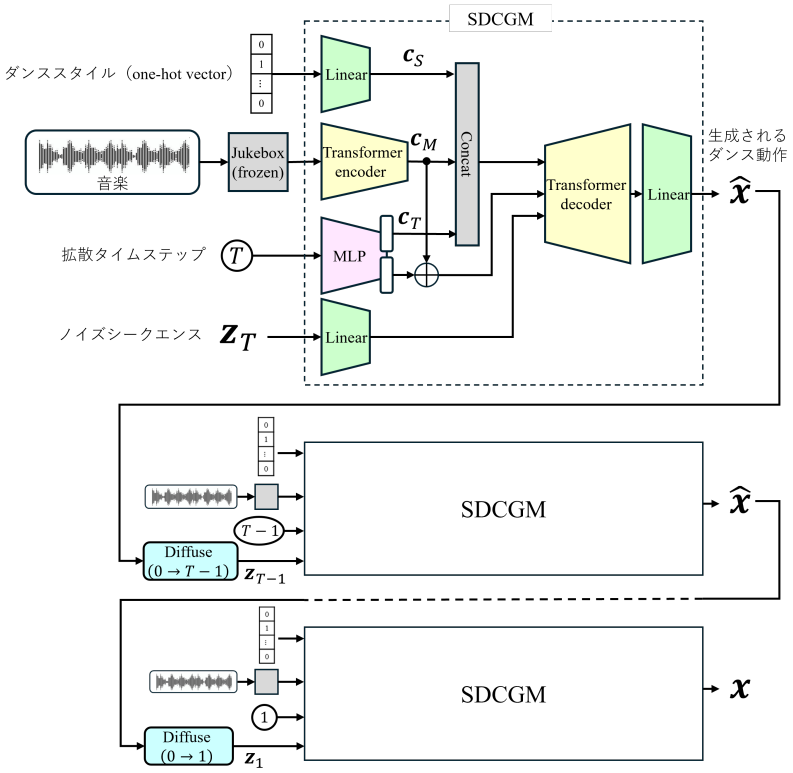

図 1　SDCGM のネットワークアーキテクチャ

メカニズムを利用して，Transformer の次元に投影された音楽条件 c_M およびダンススタイル条件 c_S を処理する．

損失関数には，単純な再構成損失と補助損失の 2 種類を使用した．補助損失は，物理的な特徴や制約に基づいた損失であり，運動生成の問題で一般的に使用されている．このような 2 種類の損失を考慮した統合的な損失関数は，x と $\hat{x}$ 間の関節位置，関節速度，足の速度，足と地面の接触の有無に関する差分の加重和として計算する．

各ノイズ除去時間ステップ t において，SDCGM はノイズ除去後のサンプルを予測し，時間ステップ $t-1$ におけるノイズとして返す．すなわち，$\hat{z}_{t-1} \sim q(\hat{x}_\theta(\hat{z}_t, t, c_M, c_S), t-1)$．これが $t=0$ に達すると終了する．これらのプロセスの訓練には，拡散ベースの様々なモデル [9], [22], [25], [26], [27] で一般的に利用されている分類器フリーガイダンス [19] を使用した．先行研究である [19] に従い，訓練中に低確率で条件付けをランダムに置き換えることで，分類器フリーのガイダンスを実装する．すなわち，$c_M = \emptyset$ および $c_S = \emptyset$ とする．ガイダンス付き推論は，無条件および条件付きで生成されたサンプルの重み付き和として表現される．

3.2　再生成プロセス

EDGE では，時系列的に連続的なダンスを生成するため，連続的に見たときに滑らかではない動き（例えば，人間が実施不可能な姿勢の飛び）が不適切に発生することがある．このような不適切な生成結果を是正するために，SDCGM では新たに，同じシーケンスに対して，再度，ダンスを生成する新しいメカニズムを導入する．具体的には，24 の関節の動作のそれぞれのフレーム間の速度が一定の閾値を超えると，滑らかではない動きが発生したと判断する閾値処理を導入するこの処理により，最終的に生成される時系列ダンス動作の品質が大幅に向上することが期待される．なお，後述する評価実験時には，このような再生成処プロセスは一つのシーケンスに対して最大で 30 回行われた．

3.3　ロングシーケンスの時系列ダンス生成

提案する SDCGM が 1 回の生成で扱うダンスシーケンス長は原理的に任意な設定が可能であるが，長くするほど一度に生成するダンスデータ長が増えるため，学習のためにより多くのデータ量が必要となる．使用できるダンスデータ量には限りがあるため，適度に短いシーケンス長に設定することで，学習に使用されるロングシーケンスなダンスデータを複数に分割して十分なデータ量を確保し，モデル学習をおこなうことが現実的である．SDCGM と同様な拡散モデルを使用した EDGE[6] では，モデルが扱うダンスシーケンス長を 5 秒に設定をしている．これを参考に，本研究でも後述する 4 章の実験において，5 秒と設定して SDCGM を学習している．

このようにモデルが，現実的に一度に生成可能なシーケ

ンス長には限りがあるため，ロングシーケンスな音楽に合わせた連続的なダンスを生成するためには，時系列的に連続した複数の生成結果をスムーズにつなぎ合わせる必要がある．このようなロングシーケンスな時系列ダンス作成をするために，モデルで扱うダンスシーケンス長を 5 秒と設定した本研究での具体的な処理を説明する．まず，5 秒間の音楽とダンススタイル情報を 2.5 秒でスライドしてモデルに入力としてダンス動作を順に生成する．これにより，2.5 秒ずつスライドされた 5 秒間の時系列ダンスシーケンスを取得する．これらの生成結果を，直線的に減衰する重み付け補間をすることで一貫性を維持したまま繋ぎ合わさる．この補間処理中，2 つの生成されたジョイントクォータニオンの積が 2.5 秒間，正と負の両方の値で混合された場合，クォータニオン間の境界で不連続な補間結果を生成される問題が発生する（EDGE ではこの問題を抱えている）．この問題を解決するために，我々は不連続なフレームが発生した区間の近傍で，球面線形補間することでこのような問題を抑止する処理を実施した．これにより，滑らかな生成結果が得られ，不適切な動作は発生しなかった．

4. 評価実験

SDCGM が，指定されたダンススタイルを適切に生成しているかどうかを評価するために，運動学および幾何学的な評価指標を用いた客観的な評価と人間による主観的な評価の 2 種類の実験を実施した．加えて，SDCGM はダンススタイルを条件としてダンス生成を行うため，これによって生成されるダンスの品質を低下させる可能性も考えられる．そのため，SDCGM が品質の高いダンスを生成できるのか否かを，既存技術と比較することで検証した．使用した既存技術は，これまでの最先端の性能を誇り，SDCGM と同様に拡散モデルを使用している EDGE [6] を採用した．

4.1 ダンスデータセット

評価実験のために，SDCGM，EDGE のモデル学習をする必要がある．本研究では，オープンデータセットである AIST Dance Video Database [28]，および AIST++データセット [29] を使用した．

AIST Dance Video Database には，10 種類のダンスジャンル（ブレイク、ポップ、ロック、ミドルヒップホップ、LA スタイルヒップホップ、ハウス、ワック、クランプ、ストリートジャズ、バレエジャズ）をカバーした 1,618 のダンス映像（ダンスを最大 9 台の複数視点カメラで同時に撮影した動画が 13,989 本）が収録されている．ダンスに使用された音楽として，10 種類の各ダンスジャンルに対して異なるテンポの音楽が 6 曲ずつで合計で 60 曲のデータが収録されている．収録されているダンスの内容は，1 名のダンサーによる Basic Dance（決められた 1 種類の基本的なステップを実施するダンス），Advanced Dance（自由におこなうフリーダンス）が全体の約 93.2%と主要な割合を占めている．その他にも，複数ダンサーによるグループダンスやダンスバトルが約 6.8%の割合で収録されている．

AIST++データセットには，AIST Dance Video Database に含まれる複数視点カメラの映像をもとに，3 次元画像処理技術により，ダンサーの骨格を抽出した後に，24 の関節の SMPL フォーマットで出力したダンスモーションデータ群が含まれる．このようなデータ処理は，1 名のダンサーによるダンスデータのみが対象となっている．

本研究では，AIST Dance Video Database に収録された 60 曲の音楽データと，AIST++データセットに収録されたダンスモーションデータを使用する．使用するデータに含まれる具体的なダンス内容は，10 種類のダンスジャンルごとに，3 名のダンサーが 6 曲の音楽を使用して，Basic Dance と Advanced Dance を実施したものである．Basic Dance では，1 ジャンル当たり 10 種類の基本的なダンス振付が用意されており，それぞれのダンス振付を 3 名の内の 2 名が踊ったデータがある．なお，ダンサー 2 名の割り当ては，3 名のダンサーのデータができる限り均等になるようになされている．Basic Dance には，10 ジャンル× 6 曲の音楽× 10 種類のダンス振付× 2 名のダンサー=1,200 シーンのダンスデータがある．Advanced Dance では，各ジャンルで，ダンサー 1 名ごとに 6 曲でフリーダンスが実施されており，さらに追加で 6 曲の内の 1 曲のみで 2 回目のフリーダンスが実施されている．よって，Advanced Dance は，10 ジャンル× 7 曲（6 曲の内の 1 曲を 2 回使用）× 3 名のダンサー=210 シーンのダンスデータがある．本研究では，各曲のデータ量を一律にするために，この内，10 ジャンル× 6 曲× 3 名のダンサー=180 シーンのダンスデータを使用する．よって，Basic Dance と Advanced Dance を合わせて，1,380 シーンのダンスデータを使用する．

4.2 学習・テストデータ

EDGE[6] での設定と同様に，本研究でも，EDGE および SDCGM モデルが生成するダンスシーケンス長は 5 秒間と設定した．学習およびテストデータとして，前節で示した Basic Dance と Advanced Dance の合計 1,308 シーンの音楽およびダンスデータを 5 秒間の窓幅で抽出し，0.5 秒間隔でスライドさせながら（30fps データ）データを構築した．その結果，合計で 151,060 シーケンスのデータを得た．学習データとテストデータの分割方法として，60 曲のデータを，各ジャンル 1 曲を含む 10 曲ずつに 6 分割した．その内の 50 曲 (各ジャンル 5 曲) を学習データ（曲によって長さが異なるため，平均 125,883 シーケンス）として使用し，残りの 10 曲（各ジャンル 1 曲）をテストデータ用（平均 25,177 シーケンス）として使用した．よって，テストデータに使用された音楽は，学習データに含まれてない未知の音楽であり，未知の音楽に対して新しいダンス振付を

生成する際の性能を評価可能である．これらの 6 分割された学習データとテストデータを入れ替えてモデルを学習して評価を行う施行を 6 回繰り返す，6 交差検証を実施した．データを入れ替えて全 60 曲をテストデータを利用し，評価の信頼性を高める．

なお, テストデータ用に割り当てられたデータの内, Basic Dance データは使用せずに，Advanced Dance のみをテストデータとして利用した．Basic Dance では，各ダンスシーケンスで決められた 1 種類の基本ステップのみを実施するため，Basic Dance は曲が変わっても同じ振付が含まれることがある．これを利用した場合，Ground truth では 1 種の基本ステップのみをおこなうため，EDGE, SDCGM も同様に 1 種の基本ステップのみを生成することで，客観評価や主観評価にてダンススタイルの再現性を測る際に，高評価となってしまう懸念がある．よって，テストデータとして Basic Dance を使用することは不適切であると考えた．一方で，Advanced Dance は，ダンサーが独自で多様なダンス振付をおこなっているため，同じ振付は学習データに含まれず，テストデータとして適切であると考えられる．実際にテストデータに使用された Advanced Dance のみのデータ数は，20,788 シーケンスである．

4.3 実装の詳細

SDCGM と EDGE のモデルは, 4 枚の NVIDIA RTX A4500 GPU を利用して学習した．EDGE は，公開済みコード[*1]を利用し，文献 [6] に記載されたモデルパラメータを参考に忠実に実装した．なお，文献 [6] では，学習・テストデータの設定に関する記載が無いため不明であり，本研究と異なる可能性がある．

4.4 客観評価

SDCGM によってダンススタイルを反映して生成されたダンス動作が，同様のスタイルを持つダンサー本人のモーションキャプチャデータ（Ground truth）にどの程度近いかを客観的に評価する．評価指標として, 先行研究 [6], [15], [29] で利用されている多様性メトリックの観点で計算される指標を用いた．これは，生成されたダンスの分布の広がりを，「運動学的（Kinetic）」特徴空間 ($Dist_k$) および「幾何学的（Geometric）」特徴空間 ($Dist_g$) で測定するものである．

具体的な算出方法として，まず Kinetic 特徴ベクトルと Geometric 特徴ベクトルをそれぞれ算出する．Kinetic 特徴ベクトルは，各関節において，水平方向の速度の 2 乗平均，垂直方向の速度の 2 乗平均，平均加速度の 3 次元の特徴量ベクトル (24 関節× 3 の 72 次元ベクトル）を求める．Geometric 特徴ベクトルは，1 フレームでの姿勢や速度に関する幾何学的な条件に関して，True/False の 2 値を各次元の値とするバイナリ特徴ベクトルである．例えば，あるフレームで，首，両腿で張られる平面の法線方向について，手首の速度成分が閾値以上か否かといった 2 値の特徴である．このようなバイナリ特徴が 32 種類用意されており，Geometric 特徴ベクトルは 32 次元ベクトルで表現される．これを全フレームにて実施し，平均をとることで連続値の 32 次元特徴ベクトルを最終的な Geometric 特徴ベクトルとして求める．取得された Kinetic 特徴ベクトルと Geometric 特徴ベクトルの各次元は平均 0、分散 1 にスケールされる．次に，Kinetic 特徴ベクトルと Geometric 特徴ベクトルそれぞれで，全ての生成シーケンスのベクトル同士について，ユークリッド距離の平均値を求めたものが，$Dist_k$，$Dist_g$ となる．

表 1 EDGE - Ground truth 間と SDCGM - Ground truth 間の $Dist_k$ と $Dist_g$ の平均絶対誤差の結果

	$Dist_k$ (↓)	$Dist_g$ (↓)
EDGE - Ground truth 間	0.13	0.26
SDCGM - Ground truth 間	**0.06**	**0.01**

SDCGM の評価観点として，指定されたダンススタイルと同じダンサーの Ground truth データの分布を模倣するダンスを自動生成することができていれば，指定したダンススタイルのダンスが正確に生成できたものと考えられる．なおこれについては，先行研究 [6] においても，Ground truth データの分布を模倣することの評価の妥当性が示されている．まとめると，SDCGM では，ダンススタイルを Ground truth のダンサーと同様に設定して生成を行い，その生成ダンスと Ground truth 間での $Dist_k$ と $Dist_g$ のスコアが近くなることで，指定したダンススタイルが適切に反映されたダンスを生成できたものと評価する．

$Dist_k$，$Dist_g$ は，EDGE 文献 [6] と同様に，5 秒間のダンスシーケンスデータ単位で算出した．本評価では，SDCGM が Ground truth のダンサーのスタイルを適切に反映できているかを評価することが狙いであるために，4.2 節で示した 20,788 シーケンスのテストデータを 30 人のダンサーごとに分けて，EDGE，SDCGM，それに対応する Ground truth における，$Dist_k$ と $Dist_g$ をそれぞれ算出した．その後，EDGE，SDCGM と Ground truth との誤差を算出した．これらの絶対誤差の平均値を表 1 に示す．$Dist_k$ と $Dist_g$ の SDCGM - Ground truth 間の絶対誤差は 0.06 と 0.01 であり，EDGE - Ground truth 間（0.13 と 0.26）よりも小さい値となった．対応のある t 検定の結果，有意差が認められた（$Dist_k$ と $Dist_g$ 共に $p < 0.05$）．

4.5 主観評価

4.1 節のデータセットに参加していない，10 ジャンルのいずれかのジャンルを得意とするプロダンサー 24 名（各ジャンル少なくとも 1 名以上）を被験者とし，ダンス生成

[*1] https://github.com/Stanford-TML/EDGE

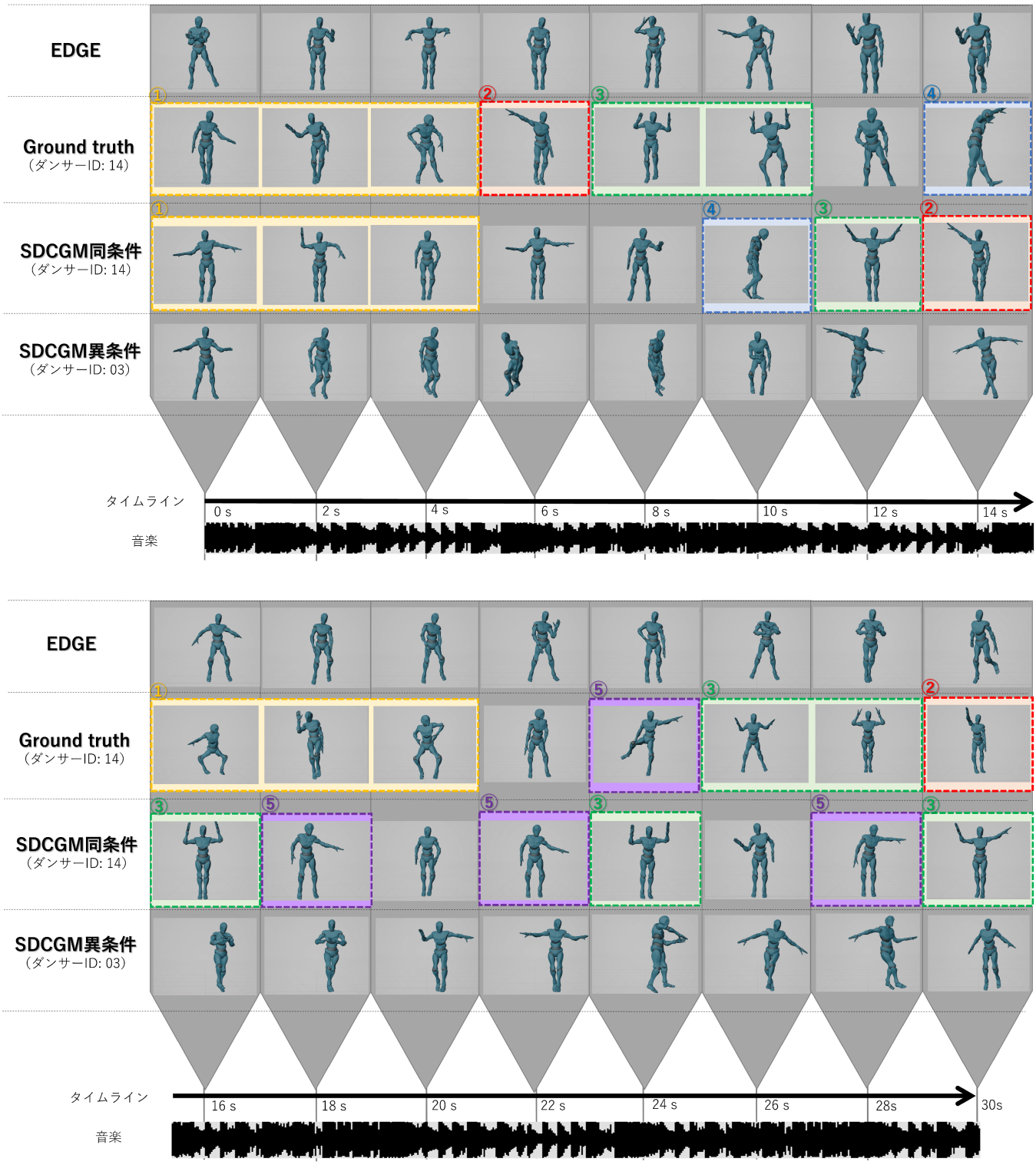

図 2 30 秒間のダンスアニメーションビデオのスクリーンショット．最上段：EDGE 条件，二段目：Ground truth 条件（ダンサー ID：14），三段目：SDCGM 同条件（Ground truth と同じスタイル [ダンサー ID：14] を指定），最下段：SDCGM 異条件 (Ground truth と異なるスタイル [ダンサー ID：03] を指定)．SDCGM 同条件，Ground truth と同じスタイルが条件として指定されてダンスが生成されており，CG アニメーションは適切にこのスタイルを反映している．スナップショット中の①〜⑤の番号は，同じダンスルーティンまたは技が実施されていることを示している．EDGE ではスタイルがまったく指定されず，SDCGM 異条件では Ground truth 条件と異なるスタイルが指定されたため，どちらもダンス動作のスタイルが Ground truth 条件と大きく異なる．

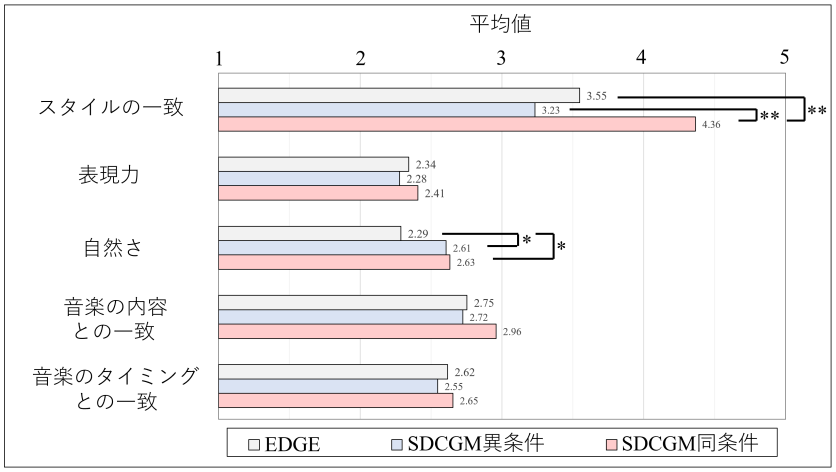

図 3 主観評価の平均スコア．**，*は多重比較後に，Bonferroni 補正を適用した t 検定の結果，$p < .001$，$p < .05$ であったことを示す．

結果に基づいてダンスを踊る CG アニメーションの映像を視聴してもらい，その印象について主観評価をおこなう実験を実施した．

評価項目として，指定されたダンススタイル（ダンサーのスタイル）が，CG アニメーションのダンス動作に適切に反映されているかどうかを評価する項目を設定した．SDCGM または EDGE で生成した 30 秒間のダンス生成結果に基づく，CG アニメーション映像と，実際のダンサーのモーションキャプチャデータ（Ground truth）の映像を左右に並べて比較し，2 つのダンススタイルが一致しているか否かを 5 段階のリッカート尺度（1:「非常に一致していない」〜5:「非常に一致している」）で評価した．これに加えて，生成されたダンスの品質を評価するために，個々のダンス映像に対して，「表現力」，「自然さ」，「音楽の内容との一致」，「音楽のリズムとの一致」の項目について，5 段階のリッカート尺度で評価した．評価項目の詳細は下記のとおりである．

- **スタイルの一致：** 2 つの動画を見比べて，ダンススタイルが一致しているか?
- **表現力：** ダンスの表現力が優れているか?
- **自然さ：** ダンスは自然か?
- **音楽の内容との一致：** ダンスが音楽の内容に合っているか?
- **音楽のリズムとの一致：** ダンスが音楽のリズムに合っているか?

次に，評価映像の作成方法を示す．4,1，4.2 節で前述した通り，テストデータである長さ 5 秒の 20,788 シーケンスデータは，各 30 名のダンサーがジャンルごとに割り当てられた 6 曲でフリーダンスをおこなった計 180 のダンスシーンデータをもとに，スライドしながら 5 秒間の窓幅で抽出されたものである．このテストデータから EDGE，SDCGM にて生成されたダンスシーケンスを 3.3 節で示した方法でつなぎ合わせて，音楽に合わせたロングシーケンスの時系列ダンスを作成した．これによって，ダンサー 30 名に対する 6 曲ずつ相当のダンスシーケンスが得られる．複数の条件によって生成された全てのダンスシーケンスを視聴して評価することは時間的に困難であったため，この中から，各 30 人のダンサーごとに 1 曲のダンスシーケンスをランダムに選択し，さらにランダムな開始位置から 30 秒間のダンスシーケンスを抽出した．このとき，30 本のダンスシーケンスでは全て異なる音楽を選択するように配慮した．これにより，EDGE，SDCGM は，それぞれ 30 秒間のダンスシーケンスが 30 本抽出される．

SDCGM における，スタイル条件の情報は，1) Ground truth と同じダンサー ID を使用した場合 (「SDCGM 同条件」と呼ぶ)，2) Ground truth とは異なるダンサー ID をランダムに選択する場合（「SDCGM 異条件」と呼ぶ）の 2 つの条件を設定した．Ground truth として，上記の方法で選択された 30 秒間の区間に相当する正解データを利用した．

CG モデルのアニメーション映像は，生成された 24 の関節の SMPL フォーマットの姿勢情報を人体の CG モデルに反映しレンダリングすることで作成した．

以上より，全 30 人のダンサーに対して，EDGE 条件，SDCGM 同条件，SDCGM 異条件，Ground truth 条件の 4 つの条件で 30 秒の映像を作成した．よって，映像は合計 120 本作成された．

実験で使用した映像サンプルをスクリーンショットを図 2 に示す．図中の①〜⑤の番号は，同じダンスルーティン，または技が実施されていることを示している．Ground truth 条件では，ロックダンスを得意とするダンサー（ダンサー ID：14）が，30 秒間，ロックダンスのさまざまなテクニックやステップを組み合わせたダンスを踊っている．ロックダンスのルーティンは，両腕を横に伸ばしてパンチを繰り出し，右手を上げて手首を円運動させ，前かがみになって腕を上げてロックするといった動き（図中では①が該当）や，手と指によるポインティング（両腕を上げて伸

ばし，どこかを指差す）（図中②が該当）が代表的的であり，30 秒間の中で多く生成されている．他にも，両手で手首を回す動き（図中③が該当），足踏みしながら体を横に回転させる動き（図中④が該当），左手を横に伸ばしてパンチを繰り出しそれを固定する動き（図中⑤が該当）についても典型的なルーティンである．

EDGE 条件では，生成されたダンスはロックダンスとは異なり，はっきりと断定することはできないが，ポップ，ヒップホップ，ハウスなど，複数のジャンルの動きが混ざっているようなダンスが生成されているように見える（スタイルを指定していないため，このような複数のスタイルが交じり合ったダンスが生成されやすい）．したがって，これは Ground truth とはまったく異なるスタイルである．SDCGM 異条件においては，Ground truth 条件と異なるスタイル（ダンサー ID：03）が指定されたため，同様に生成されたダンス動作のスタイルが Ground truth 条件と大きく異なる．一方，SDCGM 同条件では，Ground truth と同じダンサーのスタイル（ダンサー ID：14）が指定されているため，前述した①〜⑤のダンスルーティンを多く伴うダンスが生成されていることが確認できる．よって，SDCGM 同条件ではダンスのスタイルは Ground truth と非常に類似していることが観察される．

24 名の被験者が全てのアニメーション映像を視聴し，動画ごとにそれぞれの項目についての評価を実施した．EDGE 条件，SDCGM 異条件，SDCGM 同条件の評価値の平均を図 3 に示す．

「スタイルの一貫性」の平均値は，EDGE 条件は 3.55 と低く，SDCGM 異条件も同様に 3.23 と低い値であった．どちらの条件も，Ground truth のスタイルと一致していると感じられるダンスの生成がなされていなかったを結果となった．一方，SDCGM 同条件では，4.36 と高い値であった．多重比較および Bonferroni 補正を適用した t 検定により，SDCGM 同条件と EDGE 間，および SDCGM 同条件と SDCGM 異条件間に有意差が認められた ($p < .001$).

「自然さ」の平均値については，SDCGM 異条件で 2.61 と SDCGM 同条件で 2.63 であり，EDGE 条件の 2.29 に比べて統計的に有意に高い値であった（$p < .05$）．「表現力」，「自然性」，「音楽との関連性」，「音楽とのタイミングの一致」においては，EDGE，SDCGM 異条件，SDCGM 同条件の間で，有意差は認められなかった．

5. 議論

客観評価の結果, SDCGM の $Dist_k$ と $Dist_g$ の値は, EDGE に比べて，Ground truth により近い値を示した．また，主観評価の結果においても，SDCGM 同条件は「スタイルの一致」の項目で 4.36 と高い評価値を示した．この結果から，SDCGM が指定されたダンススタイル条件に従って，ダンススタイルを適切に反映してダンスを生成できていたと示唆された．一方，EDGE の「スタイルの一致」の項目の評価値は 3.55 と低く，ダンススタイルが反映されていると感じられない結果となった．これは当然の結果であり，EDGE ではダンススタイルを条件指定してダンスを生成することはできないからである．また，SDCGM 異条件においても，評価値は 3.23 と低い値となった．これは，比較される Ground truth と異なるダンサーをスタイルとして指定しため，期待通りの結果である．本結果も，SDCGM がスタイル条件を適切に反映したダンスが生成できることを支持するものである．

SDCGM では，スタイル条件を入力としない EDGE 条件に比べて，ダンススタイルの条件指定によって生成されるダンスの品質が低下するのではないかという懸念があった．しかしながら，SDCGM の両条件（SDCGM 同条件と SDCGM 異条件の両方）は主観評価にて，「表現力」，「音楽の内容との一致」，「音楽のリズムとの一致」の項目において，EDGE 条件と差が見られなかった．予想外の結果として，「自然さ」の項目については，SDCGM の両条件は EDGE 条件よりも有意に評価値が高かった．これには 2 つの理由が考えられる．1 つ目は，ダンススタイルを条件指定することで生成されるダンス動作に一貫性が確保されたことである．EDGE のようにスタイルが条件指定されていない場合，図 2 に示すように，様々なスタイルの要素が混在したダンスが生成され，動作に一貫性がないばかりか，動作が突然に変化する傾向がある．これに対して，ダンススタイル条件を指定した場合は，指定されるスタイルの条件内に生成結果が限定されるため，動作に一貫性が生まれ自然さが高いと評価された可能性が考えられる．2 つ目の理由は，3.2 節で示した再生成プロセスによるものである．ロングシーケンスのダンスを生成する場合には，EDGE では滑らかでない動きが不適切に発生してしまうことがあるが，再生成プロセスはそれを抑止できていたものと考えられる．この新たな処理が，自然さの品質を高めるのに有用であった可能性が考えられる．

以上をまとめると，我々の提案する SDCGM は，先行研究における最先端のダンス生成技術である EDGE よりも，自然さの高いダンスの生成を可能にし，指定されたダンススタイルを適切に反映したダンス動作を生成できることが示唆された．

6. アプリケーションのプロトタイプと初期評価

SDCGM を利用して，ユーザが容易に任意の音楽とダンススタイルを指定して，ダンス生成結果を視聴可能な GUI アプリケーションのプロトタイプシステムを開発した．GUI 画面のサンプルを図 4 に示す．ユーザが「音楽ファイル」と「スタイル」の選択をおこなった後，ダンス生成のボタンを押下すると生成結果がすぐに CG キャラクタに反映され，音楽と一緒にダンスアニメーションを視聴可能で

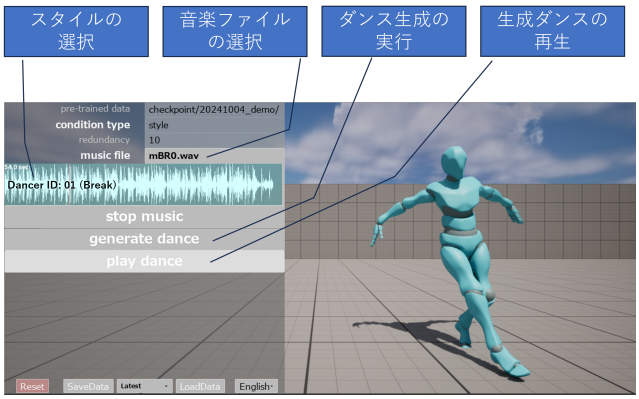

図 4　SDCGM を組み込んだダンス生成 GUI アプリケーションのプロトタイプ画面

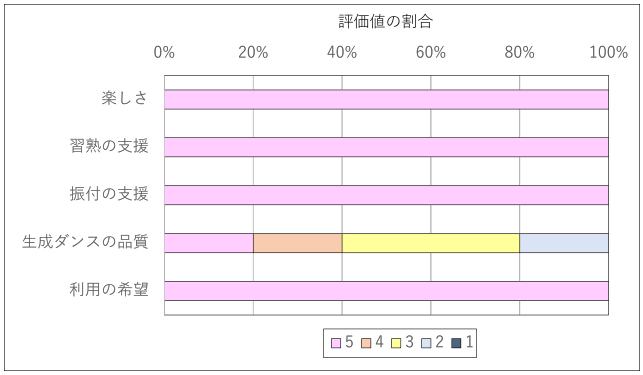

図 5　GUI アプリケーションの主観評価結果

ある．

この GUI アプリケーションの有用性や課題を抽出するための初期評価実験を実施した．被験者として，プロのダンサー（日常的にアーティストやダンスショーの振付を実施している方）5 名に 30 分程度，自由に操作してもらい，使用後にアプリケーションの利用用途や有用性を評価するために，「楽しさ」，「発想の支援」，「振付の支援」，「生成ダンスの品質」、「利用の希望」の項目について，5 段階のリッカート尺度（1:「非常にそう思わない」～5:「非常にそう思う」）で評価してもらった．評価項目の詳細は下記のとおりである．

- **楽しさ：** 本ダンス生成アプリケーションを使用することは楽しいか？
- **習熟の支援：** 本ダンス生成アプリケーションを使用することでダンスを習熟する支援ができるか?
- **振付の支援：** 本ダンス生成アプリケーションを使用することでダンスの振付の支援ができるか?
- **生成ダンスの品質：** 本ダンス生成アプリケーションで生成されるダンスの品質は十分であるか？
- **利用の希望：** 本ダンス生成アプリケーションを日常的に使用したいか？

さらに，主観評価後に，本アプリケーションの感想や改善点について，聞き取り形式で意見を収集した．

5 名の被験者の GUI アプリケーションの主観評価結果（評価値の割合）を図 5 に示す．「楽しさ」，「習熟の支援」，「振付の支援」，「利用の希望」については，5 名全員が 5 点（とてもそう思う）の高い評価値であった．本項目に関する，聞き取り調査で得られた意見として，様々な異なるダンサーのスタイルを指定して，ダンス振付を生成できる体験はこれまでになく，非常に楽しく 30 分では物足りないという意見が多くあった．同じ曲で様々な異なるスタイルでダンスを生成しシュミレーションすることは，ダンスを学び始めた初心者からプロの振付師まで幅広く，ダンスの教則映像として利用できる価値が高い．また，プロダンサーでも得意なジャンル，ダンス動作が限られており，自身が得意としないスタイルで様々なダンス振付を生成できることで新たな振付のヒントになるため，振付時の発想支援に有用であると，全員から肯定的な意見があった．まとめると，様々なスタイルを指定して生成したダンスのシュミレーション結果を視聴できることは楽しく、初心者からプロダンサーまで習熟や振付の支援への利用可能性が高いため，日常的に使用したいという評価結果であった．

一方，「生成ダンスの品質」については，5～2 点と評価値が割れる結果となった．5 名全員が，AI がダンスを生成した振付の品質の高さに非常に驚きを持っており，実際のモーションキャプチャと見分けがつかないものが多いという意見があった．一方で，大まかなダンス動作は十分に生成できているものの，細かな手足，腰といった部位の機微な動作や，音楽の特徴的な個所を動作に反映させる“音はめ”の生成が少ないという意見があった．現時点の生成品質でも，習熟や振付の支援に有用であるが，指摘のあったより細かなダンス動作の表現力も高い生成技術が構築できれば，より有用であることが示唆された．このような品質の改善は今後の課題である。

7.　まとめ

本研究では，ユーザが音楽と任意のダンススタイルを条件指定できるダンス生成技術 SDCGM を提案した．SDCDM を実装し，実験を実施した結果，SDCGM は，先行研究における最先端のダンス生成技術である EDGE よりも，より自然なダンスの生成を可能にし，指定されたダンススタイルを適切に反映したダンス動作を生成できることが示唆された．また，ダンス生成を自由に実施可能な GUI アプリケーションを開発し，ユーザ評価の初期検討を通じて，SDCGM が提供するダンス生成機能は，振付をおこなうプロダンサーにとって，ダンスの習熟や振付の支援に有益である可能性が示唆された．SDCGM を用いることで，誰もが自由に任意のダンススタイルと音楽を指定して様々なダンスを生成することで，ダンスの振付の創作活動を実施できるようになると期待される．

今後は，より大規模なダンスモーションキャプチャデー

タを収集し，より多様で自然さや表現力の品質の高い指定のダンスを生成できるモデルを構築する．また，アプリケーションのさらなる開発も進めていく予定である．

参考文献

[1] Nakano, T., Murofushi, S., Goto, M. and Morishima, S.: DanceReProducer: An Automatic Mashup Music Video Generation System by Reusing Dance Video Clips on the Web, *Proceedings of the 8th Sound and Music Computing Conference (SMC)*, pp. 183–189 (2011).

[2] Fan, R., Xu, S. and Geng, W.: Example-based automatic music-driven conventional dance motion synthesis, *IEEE transactions on visualization and computer graphics*, Vol. 18, No. 3, pp. 501–515 (2011).

[3] Lee, M., Lee, K. and Park, J.: Music similarity-based approach to generating dance motion sequence, *Multimedia tools and applications*, Vol. 62, pp. 895–912 (2013).

[4] Ofli, F., Erzin, E., Yemez, Y. and Tekalp, A. M.: Learn2dance: Learning statistical music-to-dance mappings for choreography synthesis, *IEEE Transactions on Multimedia*, Vol. 14, No. 3, pp. 747–759 (2011).

[5] Au, H. Y., Chen, J., Jiang, J. and Guo, Y.: ChoreoGraph: Music-conditioned Automatic Dance Choreography over a Style and Tempo Consistent Dynamic Graph, *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, New York, NY, USA, Association for Computing Machinery, p. 3917 – 3925 (online), DOI: 10.1145/3503161.3547797 (2022).

[6] Tseng, J., Castellon, R. and Liu, C. K.: EDGE: Editable Dance Generation From Music, *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 448–458 (online), DOI: 10.1109/CVPR52729.2023.00051 (2023).

[7] Fan, D., Wan, L., Xu, W. and Wang, S.: A bi-directional attention guided cross-modal network for music based dance generation, *Computers and Electrical Engineering*, Vol. 103, p. 108310 (online), DOI: https://doi.org/10.1016/j.compeleceng.2022.108310 (2022).

[8] Huang, Y., Zhang, J., Liu, S., Bao, Q., Zeng, D., Chen, Z. and Liu, W.: Genre-Conditioned Long-Term 3D Dance Generation Driven by Music, *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4858–4862 (online), DOI: 10.1109/ICASSP43922.2022.9747838 (2022).

[9] Kim, J., Kim, J. and Choi, S.: FLAME: Free-Form Language-Based Motion Synthesis amp; Editing, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, No. 7, pp. 8255–8263 (online), DOI: 10.1609/aaai.v37i7.25996 (2023).

[10] Li, J., Yin, Y., Chu, H., Zhou, Y., Wang, T., Fidler, S. and Li, H.: Learning to generate diverse dance motions with transformer (2020). arXiv:2008.08171.

[11] Petrovich, M., Black, M. J. and Varol, G.: TEMOS: Generating diverse human motions from textual descriptions, *European Conference on Computer Vision*, Springer, pp. 480–497 (2022).

[12] Sharif Razavian, A., Azizpour, H., Sullivan, J. and Carlsson, S.: CNN features off-the-shelf: an astounding baseline for recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 806–813 (2014).

[13] Tang, X., Wang, H., Hu, B., Gong, X., Yi, R., Kou, Q. and Jin, X.: Real-time controllable motion transition for characters, *ACM Transactions on Graphics (TOG)*, Vol. 41, No. 4, pp. 1–10 (2022).

[14] Chen, K., Tan, Z., Lei, J., Zhang, S.-H., Guo, Y.-C., Zhang, W. and Hu, S.-M.: Choreomaster: choreography-oriented music-driven dance synthesis, *ACM Transactions on Graphics (TOG)*, Vol. 40, No. 4, pp. 1–13 (2021).

[15] Siyao, L., Yu, W., Gu, T., Lin, C., Wang, Q., Qian, C., Loy, C. C. and Liu, Z.: Bailando: 3d dance generation by actor-critic gpt with choreographic memory, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11050–11059 (2022).

[16] Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A. and Sutskever, I.: Jukebox: A generative model for music (2020). arXiv:2005.00341.

[17] Castellon, R., Donahue, C. and Liang, P.: Codified audio language modeling learns useful representations for music information retrieval (2021).

[18] Donahue, C. and Liang, P.: Sheet sage: Lead sheets from music audio (2021). Proc. ISMIR Late-Breaking and Demo.

[19] Ho, J., Jain, A. and Abbeel, P.: Denoising diffusion probabilistic models, *Advances in neural information processing systems*, Vol. 33, pp. 6840–6851 (2020).

[20] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. and Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics, *International conference on machine learning*, PMLR, pp. 2256–2265 (2015).

[21] Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J. et al.: Imagen video: High definition video generation with diffusion models (2022). arXiv:2210.02303.

[22] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T. et al.: Photorealistic text-to-image diffusion models with deep language understanding, *Advances in neural information processing systems*, Vol. 35, pp. 36479–36494 (2022).

[23] Perez, E., Strub, F., De Vries, H., Dumoulin, V. and Courville, A.: Film: Visual reasoning with a general conditioning layer, *Proceedings of the AAAI conference on artificial intelligence*, No. 483, pp. 3942–3951 (2018).

[24] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G. and Black, M. J.: SMPL: A skinned multi-person linear model, *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 851–866 (2023).

[25] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B.: High-resolution image synthesis with latent diffusion models, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695 (2022).

[26] Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D. and Bermano, A. H.: Human motion diffusion model (2022). arXiv:2209.14916.

[27] Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L. and Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 46, No. 9, pp. 4115–4128 (2024).

[28] Tsuchida, S., Fukayama, S., Hamasaki, M. and Goto, M.: AIST Dance Video Database: Multi-genre, Multi-dancer, and Multi-camera Database for Dance Information Processing, *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, Delft, Netherlands (2019).

[29] Li, R., Yang, S., Ross, D. A. and Kanazawa, A.: AI Choreographer: Music Conditioned 3D Dance Generation with AIST++, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13401–13412 (2021).