スポーツ動作に対応した MTVCrafter のファインチューニング戦略

ER22051 野田悠太　　指導教授：藤吉弘亘

1. はじめに

スポーツにおいて, 映像フィードバックは選手のパフォーマンス向上に有効である．フィードバックの際は, 選手の動作を任意の条件で再現可能な映像生成が求められる．しかしながら, スポーツ特有の急激な動作に対応した映像生成の実現には, 対象とする動作に特化した学習データが必要となる．拡散モデルを活用した生成モデルである MTVCrafter は，web 上から収集した日常生活の映像を用いて学習されており，スポーツ特有の動作の生成能力は不十分である．そこで本研究では，スポーツ領域での映像生成の品質向上を目指して，MTVCrafter のファインチューニング戦略を検討する．

2. MTVCrafter

MTVCrafter[1] は, 参照画像と 3 次元関節座標の系列データから参照画像に映る人物の動作映像を生成する．MTVCrafter のモデル構造を図 1 に示す．MTVCrafter は, VAE を用いて画像を潜在空間へ写像し，潜在特徴を獲得する．そして，潜在特徴と動作情報を Transformer に入力して動画を生成する．その際，Cross-Attention を通じて動作情報を条件として取り込む．この時，学習していない動作情報は，適切な条件とならないため，破綻した映像が生成されることがある．

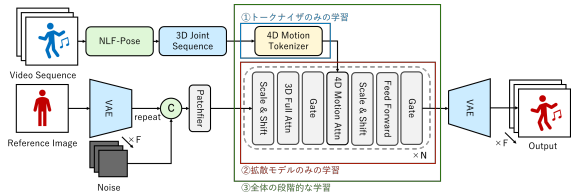


図 1: MTVCrafter のモデル構造

3. 提案手法

本研究の目的は，MTVCrafter のスポーツ領域での映像生成の品質向上である．そのため, モデルをファインチューニングするための学習戦略を検討する．具体的には, スポーツ特有の急激な動作としてフィギュアスケートのジャンプに着目し，後述の独自データセット FSJD を用いたファインチューニングを行う．

3.1. 学習戦略

MTVCrafter の学習には，①トークナイザのみの学習，②拡散モデルのみの学習，③全体の学習，の 3 つの過程がある．各学習過程が生成映像に与える影響を比較検証する．

3.2. 動画-3D ポーズ系列ペアのデータセット作成

本研究では，Figure Skating Jump Dataset (FSJD) を作成した．まず，収集した 272 本のフィギュアスケートの競技映像からジャンプシーンを切り出した．次に，各シーンに対して 3 次元人体形状・姿勢推定モデル NLF-Pose [2] を用いて，シーンの各フレームに 3 次元関節座標を付与した．得られた FSJD のデータ数は 1,072 組で，各データは解像度 512 × 512 画素，フレーム数 49 に統一されている．

4. 評価実験

本実験では，以下の学習戦略の異なる 3 つのモデルの生成性能を比較する．

モデル 1 ：トークナイザのみの学習

モデル 2 ：拡散モデルのみの学習

モデル 3 ：トークナイザを学習・凍結後, 拡散モデルを学習

本実験では, FSJD のうち学習用に 972 組, 評価用に 100 組を使用した．

4.1. 各学習戦略におけるトークナイザの比較

トークナイザの実験結果を表 1 に示す．トークナイザを学習した場合，しない場合と比較して MPJPE の低下からトークン化の再現性が確認できる．また，FID の大幅な低下から FSJD 特有の動作分布を適切に学習したといえる．

表 1: 各学習戦略の比較

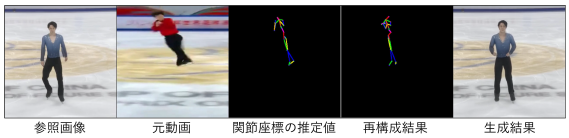
評価指標	モデル 1	モデル 2	モデル 3
MPJPE	287.28	302.47	287.28
FID	186.17	2798.47	186.17

4.2. 各学習戦略における定性的比較

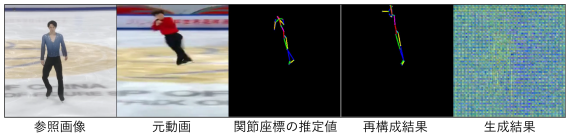
各学習戦略における推論例の比較を図 2 に示す．図 2(a) より，トークナイザのみを学習した場合，入力された動作条件を無視し，学習済みの一般的な動作が生成された．これは，拡散モデルの Cross-Attention 層が未知ドメインのトークンに対応しておらず，条件情報が意味を持たないノイズとして処理されたためと考えられる．

図 2(b) より，拡散モデルのみを学習した場合，不定形の青いノイズが生成された．これは，未学習の急激な動作に対して，トークナイザが適切に情報を保ってトークン化できず，正解の映像と矛盾した条件情報が入力されたことで，拡散モデルの最適化が阻害されたためと考えられる．

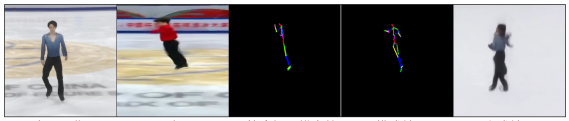
図 2(c) より，全体を段階的に学習した場合，既存モデルでは生成に失敗しやすかった空中での急激な回転時に，外観や動作が失敗なく生成されている．そのため，全体の学習がドメイン特有の動作再現において有効といえる．



(a) トークナイザのみを学習した際の推論例



(b) 拡散モデルのみを学習した際の推論例



(c) 全体を学習した際の推論例

図 2: 各学習戦略における推論例の比較

4.3. 全体学習の定量的評価

全体を段階的に学習した結果を表 2 に示す．特に FVD 及び FID-VID の改善から，ドメイン特化の学習により，ジャンプ特有の動作に対応した生成が可能となった．

表 2: 各評価指標における結果比較

比較対象	FVD	FID-VID	PSNR	SSIM	LPIPS	FID
学習前	479.15	39.11	13.10	0.457	0.525	35.67
学習後	303.39	35.30	13.58	0.577	0.530	35.73

5. おわりに

本研究では，MTVCrafter の学習データにない別ドメインでの映像生成の品質向上を試みた．具体的には，FSJD を構築し，トークナイザ及び拡散モデルをファインチューニングした．結果，全体を段階的に学習した場合に，定性・定量評価において，既存モデルでは難しかったジャンプ特有の動作の再現性の向上といった映像生成の品質向上を確認した．今後は，さらにテキスト情報を条件として加えることで，生成における能動的なジャンプ種別の指定を図る．

参考文献

[1] Yanbo Ding, *et al.*, “MTVCrafter: 4D Motion Tokenization for Open-World Human Image Animation”, arXiv:2505.10238, 2025.

[2] István Sárándi and Gerard Pons-Moll, “Neural Localizer Fields for Continuous 3D Human Pose and Shape Estimation”, arXiv:2407.07532, 2024.