拡散モデルを用いた動画生成における制御手法の開発への取り組み

理学専攻・情報科学コース
2340680
Hang Yuxin

1　はじめに

拡散モデルを用いたテキストから画像や動画を生成する技術は，近年急速に発展している. しかし，生成された動画の品質や評価方法には依然として課題が残されている. 評価方法については，美的品質や時間的一貫性を評価する手法が人間の直感と一致しない場合が多いことが指摘されている [1].

このことから，動画生成モデルに対する包括的な評価ベンチマークである VBench [1] の美学スコアを組み込んだ新たな損失関数を提案する. この手法により，生成動画の高解像度化とフレームの品質向上の両立を目指す.

2　関連研究

動画生成に関する研究は，主に GAN [2] や拡散モデルをベースに展開されている. 特に，Control-A-Video [3] は，テキストに加えて深度マップなどの制御信号を用いることで，一貫性のある動画を生成することを可能にしている. 一方で，生成された動画の品質評価には VBench [1] などが提案されており，時間的一貫性，視覚的品質，条件一致性など，複数の観点から総合的な評価を実現している．また，高解像度化を目的とした研究として StableVSR [4] がある. これは拡散モデルを活用し，従来の高解像度化手法と比べて，より詳細なテクスチャの生成や映像の時間的一貫性の向上に寄与している.

これらの先行研究を踏まえ，本研究では，VBench を活用した美学スコアを動画生成を学習する際の損失関数に組み込むことで，生成動画の品質をさらに向上させる新たなアプローチを提案する.

3　提案手法

本研究では，拡散モデルを動画生成の基盤モデルとし，VBench の美学スコア aesthetic quality を動画品質の指標として損失関数を設計し，動画品質向上モデルを構築する. 具体的には提案する手法は以下の 2 つの要素を組み合わせている.

3.1　テキストベースの動画生成

本研究では，まず初めに事前学習済みモデル Control-A-Video[3] を使ってプロンプトにより動画を生成する. このモデルの control map として depth map, canny map と hed map の 3 つの動画に反映させる map を採用し，それぞれに適した動画を生成できる.

3.2　美学スコアを組み込んだ動画品質の向上

生成した動画をもとに，VBench の美学評価器の結果を損失関数に組み込むことで，生成動画の美的品質向上モデルを開発する. 美学評価器が出力する美学スコアを aesthetic quality score とし，以下の損失関数を設計する.

$$\begin{aligned}&\text{aesthetic quality loss}\\&= \min(\max(1.0 - \text{aesthetic quality score}, 0.0), 1.0),\\&\text{total loss}\\&= (\text{MSE loss} + \lambda \cdot \text{aesthetic quality loss}) \cdot \frac{1}{1+\lambda}.\end{aligned}$$

ここで，aesthetic quality score は VBench に基づく美学評価器から出力されるスコアであり，λ は美学スコアと MSE 損失のバランスを調整するハイパーパラメータである. 損失関数の前半では，MSE（平均二乗誤差）損失によって生成フレームのピクセル単位での再現性を確保し，後半では (1 - aesthetic quality score) によって美学的な品質の向上を促す. これら 2 つの損失を正規化して加算することで，バランス良く学習を進める. 提案する損失関数により，動画生成時に時間的一貫性を維持しつつ，美的品質を高めるフレームが生成可能となる.

4　実験

図 1 に動画生成の全体構成図を示す.

4.1　実験設定

モデル訓練用のデータセットには，REDS4[1][5] を使用した.300 件の動画からなり，それぞれ 100 枚のフレームが含まれている. 実験では，データを訓練 8: バリデーション 1: 評価 1 に分割して使用している.

訓練ステージでは，3.2 節で示した美学スコアを考慮した損失関数を導入することで，生成動画の美的品質向上を目指している．図 1 では，REDS4 データセットの動画の N 個のフレームをエンコーダーに入力し，ノイズを入れて徐々に拡散させ，U-Net を使ってノイズを予測し，MSE loss を計算する．同時に，美学評価器を使って aesthetic loss を計算し，モデルを訓練する.

ハイパーパラメーターである学習率 $\alpha = 10^{-5}$, $\lambda = 0.1$，訓練ステップ数を 40,000 まで変化させて，validation において損失関数 total loss が最も小さいモデルを採用した.

評価ステージでは，テキストと入力動画を入れて，学習済みの Control-A-Video を使って，低解像度（LR）の中間動画を生成する. さらに，フレームの間に拡散確率モデル（DDPM）サンプリング [6] をし，動き推定とモーション補償を行うことで，各フレームを推定する. 最後に，美学評価器で美学スコアを評価し，閾値に達しなかった場合，最大 3 回までやり返し，高解像度（HR）のフレームを生成する.

評価する際に，REDS4 のテスト画像から生成画像を作り，ground truth の画像と比較して表 1 のスコアを算出する.

[1]REDS4: https://paperswithcode.com/dataset/reds

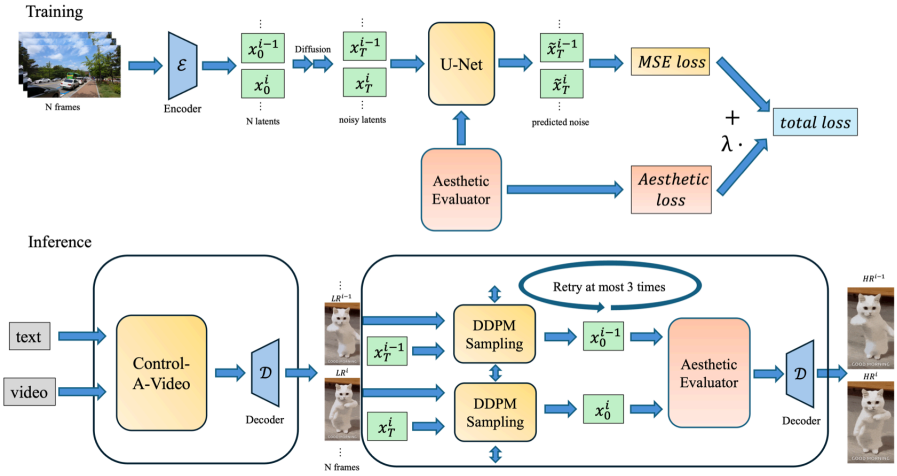

図 1: 動画生成の全体構成図

4.2 結果と考察

図 2 では，元画像（左）と生成画像（中央）のピクセルの違いを右の画像に示し，青は違いが小さく，赤は違いが大きいことを表す. 提案モデルにより生成されたフレームは特にエッジの表現力が強い. エッジのピクセルの比較（下）では，元画像のフレームのエッジ部分のピクセルは均一であるが，提案モデルの生成フレームはピクセルの値が大きく変化しており，輪郭をより描き出せている.

表 1: VSR 技術の指標比較 (一部データは [4] より引用)

VSR method	PSNR↑	SSIM↑	tLP↓	tOF↓	LPIPS↓	DISTS↓
Bicubic	26.13	0.729	22.72	4.04	0.453	0.186
EDVR	31.02	0.879	9.18	2.85	0.178	0.082
BasicVSR	31.39	0.891	9.21	2.87	0.165	0.081
BasicVSR++	32.38	0.907	9.02	2.95	0.131	0.066
RVRT	32.74	0.911	6.44	2.74	0.134	0.060
RealBasicVSR	27.07	0.778	6.27	2.61	0.130	0.054
StableVSR	27.97	0.800	5.57	2.68	0.097	0.045
Ours	**36.79**	**0.942**	14.96	7.11	**0.061**	0.046

生成画像 (K) と ground truth の画像 (I) の比較では，I と K がどれだけ一致しているフレーム再構築の指標として，ピクセルごとの類似度 PSNR と構造的類似性 SSIM を使用している.

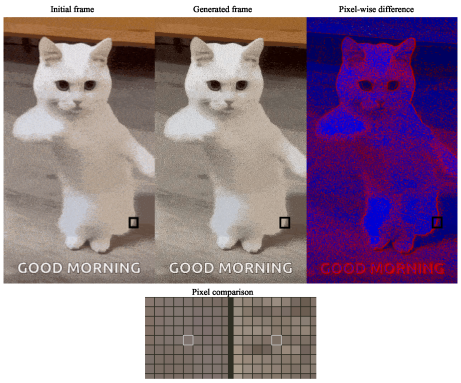

図 2: 元の猫画像と生成した猫画像との違い

時間的一貫性の指標として，局所的なピクセル値の差から運動が少ない時間平滑性 tLP と動的シーンでの運動連続性 tOF を使用している.

知覚的品質の指標として，畳み込みニューラルネットワークを用いた高次元特徴を測る LPIPS と構造的類似性を考慮した DISTS を使用している.

表 1 では，矢印↑ (↓) は値が高い (低い) ほうが品質が良いと示唆する．表 1 に示す結果から我々の提案手法は，PSNR と SSIM というフレーム再構築品質を評価する指標において，他の手法を大幅に上回る結果を示した. これは，提案手法がフレーム間の詳細なテクスチャの再現と高い忠実度の再構築を可能にしていることを示唆している．また，視覚的品質 LPIPS では他の全ての手法を上回り，視覚的品質の向上が定量的に確認された.

5 まとめ

本研究では，拡散モデルを基盤とし，VBench の美学スコアを用いた損失関数を新たに設計することで，動画生成の品質向上を目指した. 具体的には，提案手法によりフレーム再構築品質と視覚的品質において，他の最先端手法を上回る成果を得ることができた. 本研究の成果は，動画生成モデルの品質向上における新たな可能性を示しており，拡散モデルを活用した高度な生成技術の発展に寄与するものと考える.

今後の課題としては，時間的一貫性をさらに向上させるための制約条件を損失関数に組み込むことや，異なるデータセットや評価指標を用いた汎化性能の確認が挙げられる.

参考文献

[1] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[2] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[3] Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models, 2023.

[4] Claudio Rota, Marco Buzzelli, and Joost van de Weijer. Enhancing perceptual quality in video super-resolution through temporally-consistent detail synthesis using diffusion models. *arXiv preprint arXiv:2311.15908*, 2023.

[5] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.