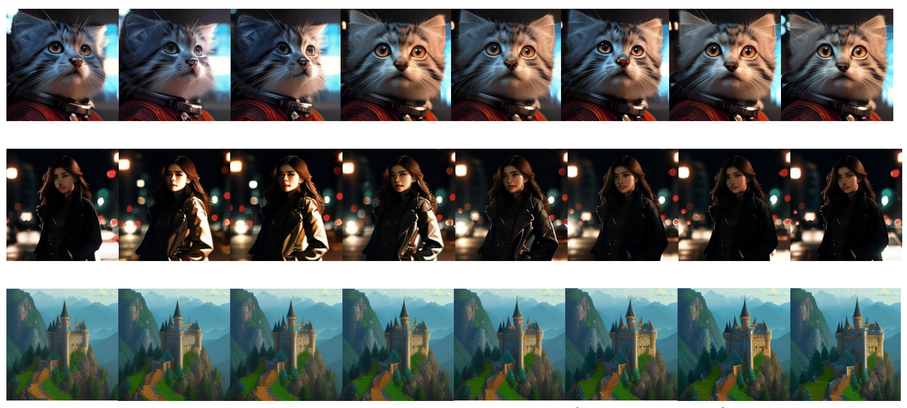
AnimateDiff をはじめとした動画生成システムに対する定量的評価指標の検討

村上研究室 29番 f19135 本田涼大

背景・目的

昨今、動画像生成分野は大きな発展を見せている。

AnimateDiff[1]：
Stable Diffusionに動きの情報を学習させ、連続した画像を出力できるようにした生成モデル。



生成した動画像の例（先頭8f）

しかし、どの論文上でも定性的な評価のみを掲載しており、定量的な評価がなされていない。

- 単に生成された動画像に対する**定量的評価指標が存在していない**

動画像に対する定量的評価指標を作成する

関連研究

EvalCrafter[2]

- 画像に対する評価指標などを画の品質や時間的整合性など様々な観点からピックアップし、ジャンルごとにスコアを出力する。

動画品質評価
Doverを用いた映像品質評価
インセプションスコア

テキストと動画の整合性
CLIP-Score BLIP-BLEW Count-Score
SD-Score Detection-Score Color-Score
OCR-Score Celebrity ID Score

動きの自然さ
Action-Score Motion AC-Score
Flow-Score

時間的整合性
ワープエラー CLIP-Temp 顔の一貫性

しかし…

1. **ソースコードが非公開**であり、信頼性に欠ける
2. **人間の評価との相関がそこまで高くない**

提案手法

画像に対する評価指標などを参考に、以下のようなアプローチをとる。

- 画像に含まれるノイズの量を測定する
- 骨格や顔面などの自然さを数値化する
- 動画からプロンプトを逆算し、元のプロンプトとのずれを見る
- 動画内の色の一貫性を見る
- フレームごとのセマンティックセグメンテーションの結果を比較する
- 隣接するフレーム同士のピクセル毎の変化量を見る

…etc.

上記のような評価項目について、人間にも、生成した画像に対してこちらが設定した評価項目ごとに10段階などで評価をするよう協力を依頼する予定である。最終的に、Ground truth（人間による評価）との高い相関（スピアマン相関・ピアソン相関）を示す定量的な評価指標を提案する。

現在の課題

- 動画像に対する評価すべき点を包括的に扱えているか不明
  ⇒今後他のジャンルの評価指標を参考に考察を深めていく

- 評価する人間を十分な数招集できるか
  ⇒5J や村上研究室、部活動の人間などに頼む
  不足分は X や Misskey.io などで募る

- 人間の評価との相関が認められなかった場合どうするのか
  ⇒EvalCrafter の定義する大項目別に人間の評価との相関を比較し、上回るものがあれば充分であるとする

今後の予定

- 評価方法の考察（～12上旬）
- サンプル動画像生成と人的評価（～12月中旬）
- 評価の実装（～12月中旬）
- 相関の計算及び比較（12月中旬～）
- 論文執筆（～1月末）

[1]:Yuwei Guo,Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. *arXiv preprint arXiv:2307.04725v1*, 2023
[2]:Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. EvalCrafter: Benchmarking and Evaluating Large Video Generation Models. *arXiv preprint arXiv:2310.11440v2*, 2023.