# 12 Text-to-Videoモデルを用いて生成された動画に対する定量的評価指標の検討とその評価

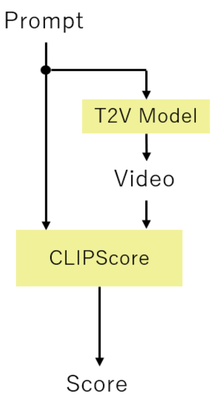
村上研究室 f19135 本田涼大

## 背景・目的

近年，俗に生成 AI と呼ばれる深層学習を用いたテキスト生成モデルや画像生成モデルが大きな発展を見せている．特に動画像生成分野においては，Text-to-Video (以下 T2V) モデルである AnimateDiff などの高性能なモデルが次々に登場している．
しかしながら，動画像生成分野においては評価指標の研究はそれほど盛んではない．T2V モデルに関するどの論文も定性的な評価のみを掲載し，定量的な評価については言及しないことが殆どである．
T2Vモデルに対する評価指標のうち，最も発展の遅い「プロンプトと動画の関係性」について，新たな評価手法を提案することにより，T2Vモデルの更なる発展を望むことができると考える．

## 関連研究

**CLIPScore**：CLIP（テキスト及び画像データを同じ埋め込み空間に置くモデル）を用いて，テキストと画像の類似度を算出するモデル

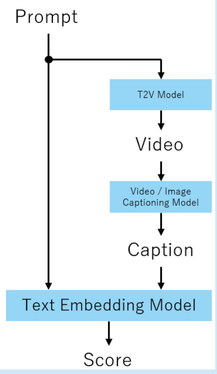



「プロンプトと動画の関係性」に関しては，現在 CLIPScore が主に使われているが，他の分野と比較して人間の評価との相関が最も低く，十分な性能があるとはいえない．要因として以下のことが考えられる．

- CLIP は一般的な画像に対して有効であるため，類似したスタイルの画像に対して正確に評価することができない可能性がある．
- 昨今の動画像生成モデルは品質の高い動画像を生成できることが多く，品質が高いというだけでプロンプトからの埋込み空間内での距離が大きく動かず，スコアにプロンプトの追従度が反映されにくくなっている可能性がある．
- 上記とは逆に，意味的に近いにも拘らず埋込み空間内での距離が近くならない可能性があり，そうした場合に意図せず追従度が低く算出されてしまう．

## 提案手法

図のように，動画に対し生成したキャプションがプロンプトと意味的にどれだけ近いのかを文書埋め込みモデルを用いて比較評価することでプロンプトへの動画の追従度を評価する手法を提案する．




パラメータ$\theta$を持つ文書埋込みモデルを$\mathrm{emb}_\Theta(\cdot)$，プロンプトを$p$，キャプションを$c$とするとき，生成された動画のプロンプトへの追従度 を以下のように定める．

$$\cos(p, c; \Theta) = \frac{\mathrm{emb}_\Theta(p) \cdot \mathrm{emb}_\Theta(c)}{||\mathrm{emb}_\Theta(p)||||\mathrm{emb}_\Theta(c)||}$$

## 性能評価

アンケートにより収集した人間の評価と提案手法により得られたスコアの相関を見ると，以下のようになった．

ただし，T2Vモデルには**AnimateDiff**，キャプション生成モデルには**TimeSformer-GPT2 Video Captioning**および**ViT-GPT2 Image Captioning**，文書埋め込みモデルに**E5**および**BERT**を採用した．

| | モデル名 | 入力 | 文書埋込みモデル | Spearman's $\rho$ | Kendall's $\phi$ |
|---|---|---|---|---|---|
| 既存手法 | CLIPScore | 動画 | - | -0.1881 | -0.1306 |
| 提案手法 | TimeSformer-GPT2 Video Captioning | 動画 | E5 | 0.1356 | 0.1088 |
| | | | BERT | **0.2176** | **0.1611** |
| | ViT-GPT2 Image Captioning | 動画 | E5 | -0.2717 | -0.2046 |
| | | | BERT | -0.2556 | -0.1828 |
| | | 画像 (平均) | E5 | -0.2557 | -0.1915 |
| | | | BERT | -0.2981 | -0.1959 |
| | | 画像 (最高) | E5 | -0.1425 | -0.1132 |
| | | | BERT | -0.1786 | -0.1349 |
| | | 画像 (最低) | E5 | **-0.3819** | **-0.2742** |
| | | | BERT | -0.3569 | -0.2655 |

| | | | | | |
|---|---|---|---|---|---|
| 動画像のサムネイル | |  |  |  |  |
| プロンプト | | In 3d game style, Salvador Dali with a robotic half face | unicorn sliding on a rainbow | A big Ferris wheel is rotating in an amusement park. | With the style of oil painting, a close-up of a bloody mary cocktail |
| ViT+GPT2 画像（最低） & E5 | キャプション | a woman wearing a clown mask and holding a cell phone | a horse statue on the side of a road | a red and white water fountain with a red and white clock on it | a woman with red hair and red eyes |
| | スコア | 0.7296 | 0.7225 | 0.7212 | 0.7754 |
| TSF+GPT2 & BERT | キャプション | A person is drawing a character in a cartoon character | A person wearing a helmet is skating on a snowy surface. | A group of people are in a gym and one of them swings a ball up and down. | A person is coloring in a picture of a cartoon character. |
| | スコア | 0.3452 | 0.1422 | 0.2535 | 0.3546 |
| CLIPScore | | 0.2673 | 0.2388 | 0.1998 | 0.2120 |
| 人間の評価 | | 0.3333 | 0.5119 | 0.7738 | 0.2619 |

## 結論

動画像キャプショニングモデルと文書埋め込みモデルを用いたプロンプトに対する生成動画像の追従度と人間の評価の相関を測定した結果，高い相関を得ることができず，既存手法と比較しても大きな差がなかった．

画風等の情報を生成できるようにキャプション生成モデルをファインチューニングするなど，工夫する余地はあるので，今後それを試していきたい．

また，キャプション生成モデルが充分な性能であるかを確認するため，追加でアンケートを実施予定である．