

令和5年度 卒業論文 (準学士) 論文番号 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

Study and Evaluation of Quantitative Evaluation Metrics for Videos Generated by Text-to-Video Model.

2024 年 2 月 15 日 一関工業高等専門学校

未来創造工学科 情報・ソフトウェア系

村上研究室

f19135 本田涼大

Suzuto Honda

主査 村上力

副査 千田栄幸

副査 小保方幸次

要旨

近年動画画像生成分野が目覚ましい発展を見せている．オープンソースかつ個人のコンピュータでも十分に利用できるほど軽量な，新たなモデルや LoRA [1] が次々に登場している．しかし，動画画像生成モデルに対する評価指標に関する研究は殆ど行われてきていない．特に，Text-to-Video モデルにおけるプロンプトと生成動画画像の関連性に対する評価指標においては，人間の行う評価との高い相関があることが示された手法が存在していない．既存手法である CLIPScore [2] においても高い相関が確認されているわけではない．

そこで，プロンプトと生成動画画像の関連性に対する評価指標として，動画画像キャプション生成モデルを用いて生成動画画像に対するキャプションを取得し，文書埋込みモデルを用いてプロンプトとキャプションを比較した結果をプロンプトに対する動画画像の追従度とする手法を提案する．本論文では，その評価指標としての有効性を検証する．

複数のキャプション生成モデル及び文書埋込みモデルの組み合わせについて検証した結果，提案手法と人間の評価との間に高い相関は確認されず，CLIPScore と比較しても大きな差がないことが分かった．

Abstract

Nowadays, the development of Video Generation has been remarkable. New models and LoRAs appear one after another that are open source and light enough to be used on personal computers. However, there has been little research on evaluation indices for the video generation models. In particular, no method has been shown to correlate highly with human evaluations of the relationship between prompts and generated videos in the Text-to-Video model. The CLIPScore, an existing method, does not show a high correlation.

We propose a method to evaluate the relationship between a prompt and a generated video image by acquiring a caption for the generated video image using a video caption generation model and comparing the result of the comparison between the prompt and the caption using a text embedding model as a measure of the video image following to the prompt. This paper examines its effectiveness as an evaluation indicator.

The validation of several combinations of caption generation and language embedding models showed no high correlation between the proposed method and human evaluation, and there was no significant difference compared to CLIPScore.

目次

第 1 章	はじめに	4
1.1	研究の背景	4
1.2	研究の目的	4
1.3	本論文の構成	4
第 2 章	関連研究	6
2.1	EvalCrafter	6
2.2	CLIPScore	6
2.3	T2V モデル	6
2.4	キャプション生成モデル	6
2.5	文書埋込みモデル	7
第 3 章	提案手法	8
第 4 章	性能評価	9
4.1	概要	9
4.2	アンケート	10
4.3	実験環境	10
4.4	結果	10
4.5	考察	13
第 5 章	結論	14
5.1	まとめ	14
5.2	今後の予定	14
	参考文献	14

図目次

2.1	CLIPScore を用いたプロンプトに対する画像の追従度算出	7
3.1	提案手法による追従度の評価	8
4.1	TimeSformer-GPT2 Image Captioning と BERT を用いたスコアと Ground-truth の散布図	11
4.2	ViT-GPT2 Image Captioning(画像入力, 最小値採用) と E5 を用いたスコアと Ground-truth の散布図	11

表目次

4.1	提案手法のスコアと人間の評価の相関.	11
4.2	各モデルのキャプションとスコア（抜粋）.	12

第1章 はじめに

1.1 研究の背景

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。画像生成モデルである StableDiffusion [3] などが様々な分野で活躍している。特に動画像生成分野においては、StableDiffusion をもとに開発された Text-to-Video (以下 T2V) モデルである AnimateDiff [4] や、イラストや写真内に写る人物に指定した動きをさせることができる Animate Anyone [5] 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。数少ない T2V モデルに対する評価指標の 1 つである EvalCrafter [6] では動画像生成モデルに対する評価を「映像の品質」「プロンプトと動画の関連性」「動き方の自然さ」「隣接フレーム間の整合性」の 4 つの観点から行っているが、いずれも人間の評価との相関は非常に高いとはいえず、十分な性能ではない。特に「プロンプトと動画の関係性」に関しては、現在 CLIPScore [2] 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 [7] 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。

EvalCrafter において最も人間の評価との相関が低かった分野である「プロンプトと動画の関係性」を評価する指標を新たに考案することにより、T2V モデルのさらなる発展を望むことができると考える。

1.2 研究の目的

本研究では、T2V モデルにおけるプロンプトに対する動画像の追従度の新たな評価指標を提案し、提案手法がどれほどの有効性を持つのかを検証する。

1.3 本論文の構成

本論文の構成は以下の通りである。

第 2 章では, EvalCrafter や CLIPScore などの本研究に関連した研究を紹介する. 第 3 章では提案手法の概要や求めるものについて説明を行う. 第 4 章では提案手法の出したスコアと人間の評価の相関を計算し, 既存手法と比較する. 第 5 章では論文全体のまとめと今後の展望を記述する.

第2章 関連研究

2.1 EvalCrafter

本研究の先行研究には EvalCrafter [6] がある。EvalCrafter は、動画像生成モデルを多角的に評価するために様々な指標を組み合わせたものである。第1章に記述したように、動画像生成モデルに対する評価指標を4つの分野に分けており、それぞれについて人間の行った評価との相関係数の値が記述してある。「映像の品質」「隣接フレーム間の整合性」については中程度の相関がみられたが、「プロンプトと動画の関連性」「動き方の自然さ」については高い相関が得られなかった。

2.2 CLIPScore

CLIPScore [2] は、CLIP [7] という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である。図 2.1 のように、プロンプト及び動画像をモデルに入力として与えることでプロンプトに対する動画像の追従度が算出される。画像 i とそのプロンプト p に対して、 $w = 2.5$ として、埋込みを $\text{emb}(\cdot)$ としたとき、CLIPScore は次のように計算する。

$$\text{CLIP-S}(p, i) = w \cdot \max(\cos(\text{emb}(p), \text{emb}(i)), 0) \quad (2.1)$$

生成された動画を V ，その各フレームを v_t 動画のフレーム数を N とする。動画像に対して CLIPScore を計算するとき、最終的なスコアはすべてのフレームの個々のスコアの平均によって導かれる。

2.3 T2V モデル

本稿では T2V モデルとして AnimateDiff [4] を実験に使用する。AnimateDiff は Text-to-Image (以下 T2I) モデルである StableDiffusion [3] を動画像出力のために拡張したモデルである。特徴的な点として、大規模な動画データセットを用いて汎化性能の高いモーションモデリングモジュールに時間的な連続性を別途学習させる点がある。このモーションモデリングモジュールを StableDiffusion に適用することで、生成されるフレーム間の動きの滑らかさと一貫性を与え、事前に学習していた重みを保持しつつ動画像出力に適応させることを可能にした。

2.4 キャプション生成モデル

キャプション生成モデルは一般に動画像を処理するエンコーダ部とキャプションへと変換するデコーダ部からなる。今回使用したモデルはエンコーダ部に Vision Transformer [8] を用いてい

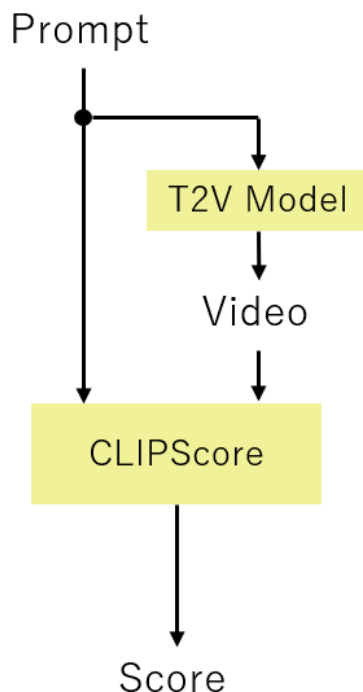


図 2.1: CLIPScore を用いたプロンプトに対する画像の追従度算出

るもの [9] と TimeSformer [10] を用いているもの [11] の 2 つである。デコーダ部は GPT2 [12] で共通している。いずれも Huggingface¹から学習済みのパラメータを含んだチェックポイントが取得可能である。

Vision Transformer は、Transformer [13] のエンコーダ部分を用いて画像パッチを単語のように読み込むことで非常に高性能な画像のエンコードが可能なモデルである。TimeSformer は、動画を時間軸アテンション及び空間軸アテンションの 2 種のパッチを適用させ、画像と同様に Transformer に認識させる機能を持つ。

デコーダに用いられている GPT2 も基本的には Transformer と同様の構造を持つ。特定のタスクに特化した学習は行わず、代わりに大きな言語モデルを事前学習させることにより、zero-shot での高い性能を誇る。

2.5 文書埋込みモデル

本稿では文書埋込みモデルとして BERT [14] と E5 [15] を採用した。

BERT (Bidirectional Encoder Representations from Transformers, Transformer による双方向エンコーダ) は Google から発表された文書埋込みモデルである。文章を文頭と文末から双方向に学習するように設計されており、それにより学習していない単語の意味を類推できる。E5 (Embeddings from bidirectional Encoder representation, 双方向エンコーダ表現からの埋込み) は reddit や wikipedia から収集した大規模なテキストと画像のペアのデータセットを用いて対照学習を行った後に高品質なデータセットで学習を行っている。

¹Hugging Face – The AI community building the future. <https://huggingface.co/>

第3章 提案手法

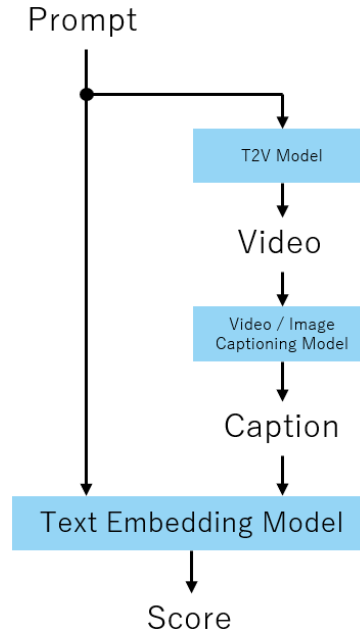


図 3.1: 提案手法による追従度の評価

T2V モデルはプロンプトと呼ばれる入力指示文からに従い動画を生成する．動画キャプションモデルは動画を元に説明文を作成する．ここにはいくつか評価指標があり，高精度でキャプションが作成できることがわかっている．従って，図 3.1 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ことができるかと仮定する．一度テキストに変換してから比較を行うため，画像中の物体などの画風でない情報が強調され，予想される既存手法の問題点である「類似したスタイルの画像に対して正確な評価ができない」という点や「美しい画像同士では距離が大きく動かないためにスコアが変化しにくい」という点に対して改善が見込める．

EvalCrafter [6] ではプロンプトに対する動画の追従度に関する指標として CLIPScore [2] 以外にもいくつか提案されている．しかし，そのいずれも CLIPScore よりも性能が低いか，特定の画像に対してのみ有効である．一般的な画像に対して使用可能である手法の中で最も性能が高い CLIPScore を提案手法の比較対象とする．

パラメータ θ を持つ文書埋込みモデルを $\text{emb}_{\theta}(\cdot)$ ，プロンプトを p ，キャプションを c とするとき，例えばプロンプトの埋込みは $\text{emb}_{\theta}(p)$ で表される．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．

$$\cos(p, c; \Theta) = \frac{\text{emb}_{\theta}(p) \cdot \text{emb}_{\theta}(c)}{\|\text{emb}_{\theta}(p)\| \|\text{emb}_{\theta}(c)\|} \quad (3.1)$$

第4章 性能評価

4.1 概要

提案手法の有効性を評価するために、AnimateDiff を用いて生成した動画像 31 本に対し、提案手法により出力したスコアとアンケートの結果から得た評価の相関を見る。また、従来手法として CLIPScore [2] を用いた場合の人間の評価との相関も算出する。

実験に使用するプロンプトと動画像のペアを作成する。プロンプトは EvalCrafter [6] のリポジトリに置かれている 700 のプロンプトの中から無作為に抽出した。動画像生成には AnimateDiff [4] を使用した。このとき、モデルのファインチューニングに用いる LoRA (Low-Rank Adaption, 低ランク適用) [1] は AnimateDiff の GitHub リポジトリ¹内に存在する bash ファイルからダウンロードできる ToonYou², Lyriel³, Rcnz Cartoon 3d⁴, majicMIX realistic⁵, Realistic Vision V6.0 B1⁶, Tusun^{7,8}, FilmVelvia^{9,10}, GhibliBackground^{11,12} の 8 つを使用した。生成された動画はすべて秒間 8 フレーム、全体 12 フレームの 2 秒の動画で、解像度は 512×512 である。

次に、生成された動画像に対してキャプションを生成する。動画キャプションモデルの性能比較のため、キャプション生成は ViT+GPT2 [9] と TSF+GPT2 [11] の 2 つのモデルを用いて行った。

最後に各動画像のプロンプトと生成されたキャプションを文書埋込みモデルを用いてプロンプトへの動画像の追従度を算出し、人間の評価との相関を CLIPScore を用いたときの相関と比較する。

¹guoyww/AnimateDiff: Official implementation of AnimateDiff. <https://github.com/guoyww/AnimateDiff>

²ToonYou - Beta 6 ★ | Stable Diffusion Checkpoint | Civitai <https://civitai.com/models/30240/toonyou>

³Lyriel - v1.6 | Stable Diffusion Checkpoint | Civitai <https://civitai.com/models/22922?modelVersionId=72396>

⁴RCNZ Cartoon 3d - v2.0 | Stable Diffusion Checkpoint | Civitai <https://civitai.com/models/66347/rcnz-cartoon-3d>

⁵majicMIX realistic 麦橘写真 - v5 preview | Stable Diffusion Checkpoint | Civitai <https://civitai.com/models/43331?modelVersionId=79068>

⁶Realistic Vision V6.0 B1 - V5.1 (VAE) | Stable Diffusion Checkpoint | Civitai <https://civitai.com/models/4201?modelVersionId=130072>

⁷LEOSAM's TuSun/Pallas's cat/manul/マヌルネコ LoRA - v4.0 | Stable Diffusion LoRA | Civitai <https://civitai.com/models/33194?modelVersionId=97261>

⁸LEOSAM's HelloWorld XL - Reality2.0 | Stable Diffusion Checkpoint | Civitai <https://civitai.com/models/43977?modelVersionId=50705>

⁹LEOSAM's FilmGirl 膠片机 Film Grain LoRA & LoHA - VELVIA 2.0 LoRA | Stable Diffusion LoRA | Civitai <https://civitai.com/models/33208?modelVersionId=90115>

¹⁰majicMIX realistic 麦橘写真 - v5 preview | Stable Diffusion Checkpoint | Civitai <https://civitai.com/models/43331?modelVersionId=55911>

¹¹(Pyramid).lora.Ghibli_Background - v1.0 | Stable Diffusion LoRA | Civitai <https://civitai.com/models/64610/pyramidloraghiblibackground>

¹²Counterfeit-V3.0 - v3.0 | Stable Diffusion Checkpoint | Civitai <https://civitai.com/models/4468?modelVersionId=57618>

4.2 アンケート

アンケートでは、生成した動画像とそのプロンプトに対して人間による評価を行った。31 組の動画像とプロンプトのペアに対し、被験者 21 人に「1. 動画に対する簡単な説明」「2. 動画像がプロンプトに従っているかの 5 段階評価」「3. 動画像の総合的な品質の 5 段階評価」を質問した。ただし、被験者が回答について相談などすることがないように、対面にてアンケートを行った。アンケートは Google Forms により作成した。質問 1 に回答する前に質問 2 に記述されているプロンプトが目に入らないように間に改ページを挟み、すべての回答欄を必須フィールドにしている。

4.3 実験環境

実験環境 1(AnimateDiff による動画像生成)

- OS: Ubuntu 22.04.3 LTS
- CPU: 12th Gen Intel(R) Core(TM) i5-12600K
- RAM: DDR4-2666 64GB (16GBx4)
- GPU: NVIDIA GeForce RTX 3060 12GB
- Python 3.10.13, Anaconda 23.11.0, PyTorch 1.13.1

実験環境 2(上記以外)

- OS: Windows 11 Home 64bit 22H2 (OS Build 22621.3007)
- CPU: 11th Gen Intel(R) Core(TM) i7-11800H
- RAM: DDR4-3200 16GB (8GBx2)
- GPU: NVIDIA GeForce RTX 3060 Laptop GPU
- Python 3.10.11, PyTorch 1.13.1

4.4 結果

実験結果を表 4.1 及び表 4.2 に示す。ただし、表 4.1 中において、「入力」列にはそのモデルに動画形式で入力したか画像形式で入力したかを記述している。画像形式で入力した場合、各画像のスコアの平均・最高値・最低値を記録した場合でそれぞれ場合分けしている。また、 ρ はスピアマン相関、 ϕ はケンドール相関を示す。

ほとんどの組み合わせにおいて無相関または負の弱い相関を記録した。スピアマン相関・ケンドール相関共に、TimeSformer-GPT2 Video Captioning [11] と BERT [14] の組み合わせが最も Ground-truth との相関が高く（図 4.1）、1 フレームずつ入力して最低値を採用したときの ViT-GPT2 Image Captioning [9] と E5 [15] の組み合わせが Ground-truth との相関係数の絶対値が最も高い結果となった（図 4.2）。

出力されたスコアに着目すると、ほとんどの動画像に対して CLIPScore [2] 及び TimeSformer-GPT2 Video Captioning は 0.20 ～ 0.40 の値を記録し、ViT-GPT2 Image Captioning では 0.70 ～ 0.80 の値を記録した。

プロンプトに着目すると、TimeSformer-GPT2 Video Captioning により生成されたキャプションは「A person」等の人間を指す言葉から始まっている。これは抜粋した部分に限らず、すべてのキャプションに対して見られた。

表 4.1: 提案手法のスコアと人間の評価の相関.

	モデル名	入力	文書埋込みモデル	ρ	ϕ
既存手法	CLIPScore [2]	動画	-	-0.1881	-0.1306
提案手法	TimeSformer-GPT2 Video Captioning [11]	動画	E5 [15]	0.1356	0.1088
			BERT [14]	0.2176	0.1611
	ViT-GPT2 Image Captioning [9]	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

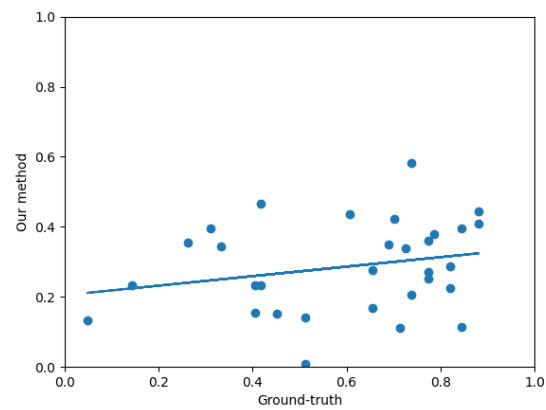


図 4.1: TimeSformer-GPT2 Image Captioning と BERT を用いたスコアと Ground-truth の散布図

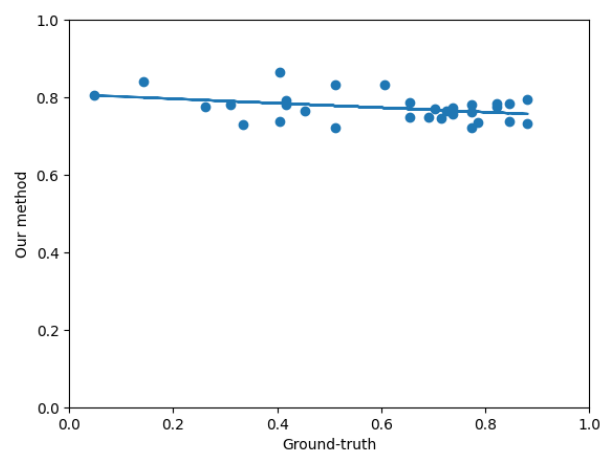


図 4.2: ViT-GPT2 Image Captioning(画像入力, 最小値採用) と E5 を用いたスコアと Ground-truth の散布図

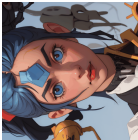
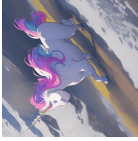

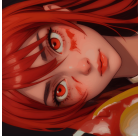
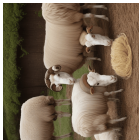


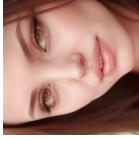
動画像のサムネイル					
ViT+GPT2 画像 (最低) & E5	プロンプト	In 3d game style, Salvador Dali with a robotic half face	unicorn sliding on a rainbow	A big Ferris wheel is rotating in an amusement park.	With the style of oil painting, a close-up of a bloody mary cocktail
	キャプション スコア	a woman wearing a clown mask and holding a cell phone 0.7296	a horse statue on the side of a road 0.7225	a red and white water fountain with a red and white clock on it 0.7212	a woman with red hair and red eyes 0.7754
TSF+GPT2 & BERT	キャプション スコア	A person is drawing a character in a cartoon character 0.3452	A person wearing a helmet is skating on a snowy surface. 0.1422	A group of people are in a gym and one of them swings a ball up and down. 0.2535	A person is coloring in a picture of a cartoon character. 0.3546
	CLIPScore	0.2673	0.2388	0.1998	0.2120
人間の評価		0.3333	0.5119	0.7738	0.2619
動画像のサムネイル					
ViT+GPT2 画像 (最低) & E5	プロンプト	3 sheep enjoying spaghetti together	Two white swans gracefully swam in the serene lake	a moose with the style of Hokusai	Angelina Jolie's full lips curve into a smile, her gaze intense and captivating.
	キャプション スコア	a herd of sheep standing on top of a dirt field 0.7722	three swans are swimming in the water near a body of water 0.8653	a statue of a man and a horse on a beach 0.7566	a beautiful young woman in a pink dress posing for a picture 0.7356
TSF+GPT2 & BERT	キャプション スコア	A man is standing in front of a flock of sheep and he is feeding them. 0.5830	A woman is showing how to make a fish out of a piece of bread. 0.1549	A man and woman are standing in front of a painting of a horse. 0.2060	A woman is looking at the camera and talking to the camera. 0.3790
	CLIPScore	0.1932	0.1887	0.2342	0.1787
人間の評価		0.7381	0.4048	0.7381	0.7857

表 4.2: 各モデルのキャプションとスコア (抜粋)。

4.5 考察

実験の結果として、動画像ごとのスコアの上下があまり大きくならないこと、文書埋込みモデルによる相関係数の差はほとんど無視できることがわかった。また、TimeSformer-GPT2 Video Captioning の生成したキャプションはすべて人間を表す語句から始まっていた。このことから、このモデルのチェックポイントを公開している作者による説明にはそのような記述はなかったが、人物の写った動画に対して特化したモデルである可能性がある。

Ground-truth との高い相関が得られなかった理由として、以下の理由が推測される。

1. プロンプトに含まれることの多い画風や画角などの情報 (*e.g.* style of Hokusai, close-up) や固有名詞 (*e.g.* Angelina Jolie, Darth Vader) をキャプション生成の時点で復元できず、比較の際にスコアが下がっている。
2. キャプションモデル若しくは文書埋込みモデルの性能が不十分である。
3. アンケートの回答者の年齢層及び所属が偏っていることが Ground-truth に何らかの偏りを生み出している。

第5章 結論

5.1 まとめ

動画像キャプション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。

5.2 今後の予定

今後の研究では上記の考察を検証するために、以下のことを実施し、プロンプトと生成動画像の関連性の評価に関する検討を続けていきたい。

1. プロンプトから画風や画角に関する記述、および固有名詞を排除した状態で実験を行う。
2. 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。
3. キャプション及び文書埋込みモデルについて、Ground-truth との比較により性能の評価を行う。
4. 幅広い年齢層及び所属の人間からアンケートをとり、今回の実験の結果と比較する。

参考文献

- [1] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.
- [2] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718, 2021.
- [3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10684–10695, 2022.
- [4] 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:2307.04725, 2023.
- [5] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. arXiv preprint arXiv:2311.17117, 2023.
- [6] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tiejong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. arXiv preprint arXiv:2310.11440, 2023.
- [7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pp. 8748–8763. PMLR, 2021.
- [8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
- [9] NLP Connect. vit-gpt2-image-captioning, 2023. <https://huggingface.co/nlpconnect/vit-gpt2-image-captioning>.
- [10] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In ICML, Vol. 2, p. 4, 2021.

-
- [11] Caelen Wang. Timesformer-gpt2 video captioning, 2022. <https://huggingface.co/Neleac/timesformer-gpt2-video-captioning>.
- [12] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, Vol. 1, No. 8, p. 9, 2019.
- [13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, Vol. 30, , 2017.
- [14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [15] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. arXiv preprint arXiv:2212.03533, 2022.