日本語入力にネイティブ対応したテキストからの動画生成のフルスクラッチ開発と公開

尾崎 安範 (責任著者)[1,a)]　石原 昌文[1,b)]　富平 準喜[1,c)]

概要

本技術報告では、日本語を入力できるテキストからの動画生成をフルスクラッチで開発し、公開したことを報告する。我が国のコンテンツ産業は半導体産業に匹敵する輸出額に匹敵しており、その支援が急務となっている。そこで、米中の知見を活かし、動画生成のフレームワークを利用しながら、日本語圏向けの新しい動画生成を提案する。提案した動画生成は日本語に対し、他のモデルを FVD、アラインメントを上回る結果を得られた。今後は計算量を拡充することで映像品質の向上が必要であることを示した。

1. はじめに

我が国は『新しい資本主義のグランドデザイン及び実行計画 2024 年改訂版』に示されるように国際競争を念頭に経済成長を目指している。例えば、我が国のコンテンツ産業の輸出額は半導体産業に匹敵すると本書に書かれており、生成 AI を活用したコンテンツ制作業務の支援が重要であると内閣府が述べている。この生成 AI を活用したコンテンツ制作業務の支援に期待できる技術として、テキストからの動画を生成する AI（以下、動画生成）がある。動画生成の開発は世界的に競われており、特にアメリカや中国で活発化している。以上の背景から、動画生成に対するアメリカや中国の高度な知見を活かしながら、日本語に対応したコンテンツ制作業務支援のための高度な動画生成を開発することを本研究の目的である。

2. 関連研究

動画生成は Brooks らの Sora[1] を皮切りに米中で競争が激化している。Sora を始めとした動画生成は図 1 のとおりのアーキテクチャにまとめられる。図 1 のように一般的な動画生成は、入力プロンプトを埋め込みに変換するテキストエンコーダー、入力された埋め込みに応じて、入力されたノイズを潜在空間における動画に変換していくノイズ除去器、潜在空間における動画をピクセル空間における動画に変換する VAE の 3 つのパーツで構成される。例えば、このフレームワークを Yang らの CogVideoX[7] に落とし込むと、テキストエンコーダーに T5-XXL を、ノイズ除去器に拡散トランスフォーマー (Diffusion Transformer[5]) 、VAE に独自の VAE を使用していることになる。このようにシンプルな構成になるが、日本語にネイティブ対応した動画生成は我々が見る限りでは見つからなかった。そこで、本研究では、日本語にネイティブ対応した動画生成を作ることを目的とする。

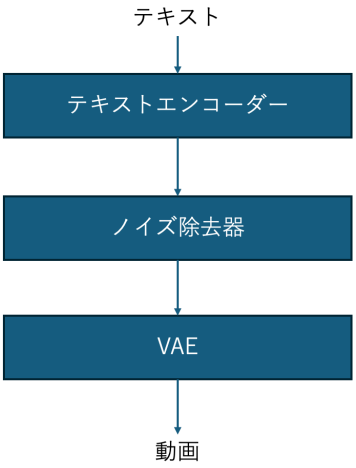

図 1: 動画生成のフレームワーク

3. 提案手法

フレームワークには図 1 と同じ構造を使う。

関連研究を見るとテキストエンコーダーに T5-XXL が多くは使われており [7]、テキストエンコーダーが日本語にネイティブ対応していない。そこで動画生成を作るためにテキストエンコーダーとして日本語に対応している大規模言語モデルを用いる。

また、ノイズ除去機は Rectied Flow Transformer を用いる。Rectied Flow Transformer [2] とは、Rectified Flow を拡散トランスフォーマーの枠組みでとけるようにしたもの

[1] AIdealab 株式会社
[a)] ozaki.yasunori@outlook.com
[b)] maty0505@gmail.com
[c)] tomihira@aidealab.com

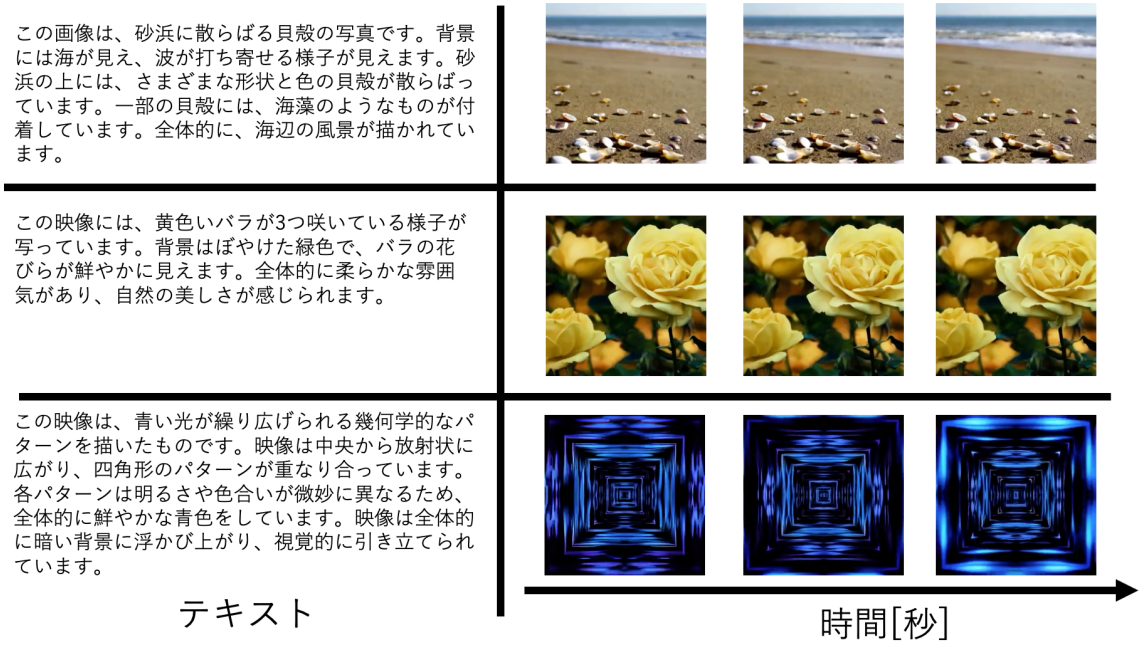

図 2: テキストとそれに対応する提案手法第 1 ステージで生成された動画

である。Rectified Flow[4] とはノイズとデータを最短距離で結ぶ流れ（ベクトル場）を直接学習するシンプルな生成モデルである。Rectified Flow では、データ分布 π_0 からある目標分布 π_1 へと流すベクトル場（vector field）を直接学習する。従来の拡散モデルでは、ノイズからデータへと復元する過程を逆拡散プロセスとして近似していたが、Rectified Flow はその近似をバイパスし、初めから「真っ直ぐに」（rectified）データ分布に向かう経路を学習する。

具体的には、以下のような連続時間の流れを考える。

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{v}_\theta(\mathbf{x}(t), t), \tag{1}$$

ここで $\mathbf{v}_\theta$ は学習するベクトル場、$\mathbf{x}(t)$ は時間 $t \in [0,1]$ におけるデータ点である。

Rectified Flow の学習目標は、「ノイズとデータの間を直線的に結ぶ流れ」を模倣するようにベクトル場を設計することである。このため、以下の形式の損失関数が提案されている。

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x}_0\sim\pi_0,\mathbf{x}_1\sim\pi_1,t\sim\mathcal{U}(0,1)}\left[\|\mathbf{v}_\theta(\mathbf{x}_t,t) - (\mathbf{x}_1 - \mathbf{x}_0)\|^2\right], \tag{2}$$

ここで、$\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$ は線形補間によって定義される中間点であり、$\mathcal{U}(0,1)$ は一様分布である。

つまり、各中間点 $\mathbf{x}_t$ において、「出発点から到達点へのベクトル」（すなわち $\mathbf{x}_1 - \mathbf{x}_0$）を指し示すようにベクトル場 $\mathbf{v}_\theta$ を学習する。

学習が完了した後、サンプリングは単純な常微分方程式（ODE）を解くことで行うことができる。すなわち、初期点 $\mathbf{x}(0) \sim \pi_0$ を取り、以下の ODE を数値積分する。

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{v}_\theta(\mathbf{x}(t), t), \tag{3}$$

$t = 1$ に到達した時点の $\mathbf{x}(1)$ がサンプルとなる。

一般的な ODE ソルバを用いることができ、非常に少ないステップ数で高品質なサンプルを得ることが可能である。Rectified Flow の特長は以下の通りである。

- **高速サンプリング**：従来の数百ステップを要する拡散モデルに対し、数ステップでサンプリング可能。
- **学習が安定**：単純な 2 乗損失により、トレーニングが容易かつ安定する。
- **シンプルな実装**：ノイズスケジューリングや特殊なトリックを必要とせず、標準的な ODE フレームワークに適合する。

4. 実験

実験方法は、機械的評価と人的評価の 2 つに分かれる。

機械的評価は次の手順で行う。

(1) 実験用に用意した動画を学習用動画と評価用動画とをランダムに分ける。

(2) 評価用に用意した動画から視覚言語モデルで日本語のキャプションをつける。

(3) 先程得た日本語のキャプションから比較手法で動画を生成する。

(4) 先程得た日本語のキャプションから提案手法で動画を生成する。

(5) 提案手法を比較手法にして上記と同じように日本語の FVD を計算する。

人的評価は次の手順で行う。

(1) 実験用に用意した動画を学習用動画と評価用動画とをランダムに分ける。

表 1: 各モデルの FVD。提案手法第 1 ステージが最も優れている。

モデルの種類	FVD ↓
CogVideoX-2B	4540
提案手法第 1 ステージ	276
提案手法第 2 ステージ	804

(2) 評価用に用意した動画を視覚言語モデルで日本語でキャプション付けする。
(3) 先程得たキャプションと提案手法から動画を生成する。
(4) 先程得たキャプションと比較手法から動画を生成する。
(5) 提案手法の動画と比較手法の動画の動画品質とテキストアラインメントの観点からランダム化比較試験を行う。

5. 実験結果

ノイズ除去機として CogVideoX-2B と同じアーキテクチャを用いた。このため、モデルサイズは 2B である。動画生成を作るためにテキストエンコーダーとして、LLM-jp-3-1.8B[*1]を用いた。学習するデータセットは Hugging Face 上にある動画とそれを視覚言語モデルでキャプションづけした約 600 万ペアを用いた。具体的には動画は pixabay からのデータセット[*2]、finevideo[3] を用い、視覚言語モデルに Qwen2VL[6] を用いた。学習には DGX H100 (H100x8) を 2 ノード、半年間使用した。学習時間の合計は 3 万 GPU 時間 (30 k [GPU hours]) となった。学習はマルチステージ学習を行った。第 1 ステージは $256[px]\times256[px]\times48[frame]$ の 400 万ペアのデータセットで 2 万 GPU 時間学習した。第 2 ステージは $848[px]\times480[px]\times48[frame]$ の 40 万ペアのデータセットで 1 万 GPU 時間学習した。比較手法として、同じアーキテクチャの CogVideoX-2B と比較する。

機械的評価は次の手順で行った。
(1) pixabay からの動画を学習用動画と評価用動画とをランダムに分けた。
(2) 評価用に用意した動画を Qwen2VL で日本語でキャプション付けした。
(3) 先程得た日本語のキャプションと CogVideoX-2B から 1000 個動画を生成した。
(4) 先程得た日本語のキャプションと提案手法から 1000 個動画を生成した。
(5) 生成した動画と評価用動画と FVD を計算した。

計算した FVD を表 1 に示す。実際のテキストやプロンプトの一部を図 2 に示す。計算の詳細や評価に用いた動画は Hugging Face のサイトに載せた[*3]。

[*1] https://huggingface.co/llm-jp/llm-jp-3-1.8b
[*2] https://huggingface.co/datasets/LanguageBind/Open-Sora-Plan-v1.0.0
[*3] https://huggingface.co/datasets/aidealab/aidealab-videojp-eval

表 2: 各モデルの勝った数。提案手法第 2 ステージは CogVideoX-2B に対し、アラインメントでは 96% 勝利していることがわかる。

モデルの種類	映像品質	アラインメント
CogVideoX-2B	372	18
提案手法第 2 ステージ	69	423

人的評価は次の手順で行った。
(1) pixabay からの動画を学習用動画と評価用動画とをランダムに分けた。
(2) 評価用に用意した動画を Qwen2VL で日本語でキャプション付けした。
(3) 先程得たキャプションと提案手法第 2 ステージから動画を 9 個生成する。
(4) 先程得たキャプションと CogVideoX-2B から動画を 9 個生成する。
(5) 提案手法第 2 ステージの動画と CogVideoX-2B の動画の動画品質とテキストアラインメントの観点から一対比較試験を行った。

一対比較実験は Google フォームで行った。被験者は日本語圏の人間を対象として、49 人に対して行われた。結果を表 2 に示す。また、実験に使われた映像の様子を図 3 に示す。

6. 考察

表 1 や表 2 を見ると、日本語性能は他のモデルと比べて十分に上回る性能であった。これはテキストエンコーダーの変更を行い、フルスクラッチで開発したためと言えよう。提案手法に対して、CogVideoX-2B はアーキテクチャは同じものの、全く日本語に対応できていない。これはテキストエンコーダーである T5-XXL が日本語を対象外としているためである。ただし、提案手法第 2 ステージはあまり映像品質が芳しくなかった。理由として、計算量の不足がある。

7. モデルの公開

本実験で作られた提案手法第 1 ステージの重みをインターネット上に公開した[*4]。また、動画を生成するデモも公開した[*5]。デモの様子を図 4 に示す。本デモを訪れた人は 1 万人を超えている。

8. まとめ

本技術報告では、日本語を入力できるテキストからの動画生成を開発し、公開したことを報告した。日本語における評価は他のモデルと比べて十分に上回る性能であった。

[*4] https://huggingface.co/aidealab/AIdeaLab-VideoJP
[*5] https://huggingface.co/spaces/aidealab/AIdeaLab-VideoJP

(a) 提案手法第 2 ステージによって生成された結果の例

(b) CogVideoX-2B によって生成された結果の例

図 3: 人的評価に使った動画の例。これらの動画は次のテキストを入力して得られたものである。「この画像は、都市部の空撮写真です。中央に位置するのは、緑色のドームを持つ教会で、その周囲には多くの建物が見えます。教会の周辺には緑豊かな公園が広がり、その背景には川や橋が見えます。都市部全体は、様々な色の建物で構成されており、街並みが広がっています。」

図 4: 公開したデモンストレーションの様子

今後の課題として、計算量の拡充が挙げられる。

謝辞

本報告は、経済産業省および NEDO が支援するプロジェクト「GENIAC[*6] (Generative AI Accelerator Challenge)」において著者らが実施した研究開発の成果に基づいて作成されました。本研究には NEDO からの助成金が用いられていますが、本稿は商用目的ではなく、著者らの判断で公表するものです。

参考文献

[1] Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R. and Ramesh, A.: Video generation models as world simulators (2024).

[2] Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z. and Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis, *Proceedings of the 41st International Conference on Machine Learning*, ICML'24, JMLR.org (2024).

[3] Farré, M., Marafioti, A., Tunstall, L., Von Werra, L. and Wolf, T.: FineVideo, `https://huggingface.co/datasets/HuggingFaceFV/finevideo` (2024).

[4] Liu, X., Gong, C. and qiang liu: Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow, *The Eleventh International Conference on Learning Representations* (2023).

[5] Peebles, W. and Xie, S.: Scalable Diffusion Models with Transformers, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205 (2023).

[6] Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J. and Lin, J.: Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution, *arXiv preprint arXiv:2409.12191* (2024).

[7] Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., Yin, D., Yuxuan.Zhang, Wang, W., Cheng, Y., Xu, B., Gu, X., Dong, Y. and Tang, J.: CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer, *The Thirteenth International Conference on Learning Representations* (2025).

[*6] `https://www.meti.go.jp/policy/mono_info_service/geniac/index.html`