外部知識参照型拡散モデルによるテキストからのビデオ生成

王　　墻† 宮森　恒†

† 京都産業大学大学院 先端情報学研究科　〒 603–8555 京都府京都市北区上賀茂本山
E-mail: †{i2286031,miya}@cc.kyoto-su.ac.jp

あらまし　本稿では、外部知識を参照できるテキストからビデオを生成するモデルを提案する。拡散モデルでは、元データにノイズを徐々に付与する過程とノイズを除去して元データを復元する過程を通し、データとノイズ間の関係を学習することで、最先端の合成結果を得ることができる。しかし、テキストの意図とは異なる別オブジェクトが生成されたり、一部のフレームで適切にオブジェクトが生成されないといった問題がある。本稿では、テキストに関する外部知識を参照し、テキスト情報を増強した条件ベクトルを与えることで、より忠実性の高いビデオを生成する手法を提案する。実験では、提案手法によりどの程度忠実性の高いオブジェクトを含むビデオを生成できるかについて評価し、その有効性を検証する。

キーワード　外部知識，拡散モデル，画像合成，映像合成，マルチモーダル

1 はじめに

計算基盤と機械学習技術の急速な進展により、人間によるものと見分けのつかない計算機による創作が可能となってきている。例えば、曲の長さやジャンル等から音楽を生成 [12] したり、テキストから高品質の絵画画像を生成 [6], [15] することが可能となっている。これら技術を用いて人間の創作活動を支援することにより、これまで携わることが難しかった幅広い人々が種々の創作活動へ参加することが促進されると期待できる。

本研究では、テキストからビデオを生成するタスクに焦点を当てる。画像生成タスクでは、従来、GAN が主に利用されていたが、近年、拡散モデルが提案され研究が活発に進められている。拡散モデルは、GAN の訓練における不安定性と生成結果の多様性が不足する問題を改善し、よりリアリティの高い創作結果を出力することができる。さらに、拡散モデルを拡張し、ビデオを生成する研究 [7] も登場している。

しかし、テキストからビデオを生成する従来モデルは、訓練時に観測したオブジェクトのみを生成でき、新たなオブジェクトを生成させたい場合は再学習が必要である。例えば、特定の人物を想定して生成したい場合、特定人物に関する情報がなければ、訓練時に観測した範囲から推測される範囲での人物として生成されることとなり、想定通りの生成を行うことは困難である。また、拡散モデルでビデオ生成する場合、オブジェクトの形状が崩れたり別のオブジェクトと融合してしまう場合があり、オブジェクトの時間的連続性を維持することも課題である。

そこで本稿では、特定オブジェクトに関する情報を保存した外部知識を参照することで、訓練時には網羅できていないオブジェクトについても忠実に反映することができるビデオ生成手法を提案する。本手法では、まずテキストを外部知識と関連付け、補強するネットワークを訓練する。次に、外部知識で補強されたテキスト埋め込みからビデオフレームの隠れ情報を生成する拡散モデルを訓練する。最後に、隠れ情報から画像デコーダでビデオフレームを生成する。

テキストと外部知識を関連付けて補強するネットワークを訓練するため、既存の画像データセットから、オブジェクト、オブジェクトの説明文の各トークン、両者の関連度から構成される 3 つ組のデータセットを構築する。オブジェクトの説明文については必要に応じて生成することとした。

本稿の構成は以下の通りである。2 章では、テキストから画像およびビデオを生成する関連研究を紹介する。3 章では、提案手法とデータセットについて詳しく説明する。4 章では、実験内容を説明し、実験結果を示す。5 章では、まとめと今後の課題について述べる。

2 関連研究

2.1 拡散モデルによる画像生成

画像生成においては、従来 GAN [4] を用いた手法が多く提案されてきたが、近年、拡散モデル [6] が提案され研究が活発に進められている。

拡散モデルは、生成モデルの一種であり、元の情報に段階的にノイズを加え、最終的にランダムノイズになる過程からその逆過程を学習することで、生成時は、ランダムノイズから逆過程を通して画像を生成する。拡散モデルは、GAN の訓練における不安定性と生成結果の多様性が不足する問題を改善し、よりリアリティの高い画像を出力することができる。そのため、画像生成だけでなく、超解像や画像修復、画像編集、画像分割、画像分類等、様々なタスクへの適用が進んでいる。

一方、拡散モデルは生成時に時間がかかるという問題に対して、潜在拡散モデル (Latent Diffusion Models; LDM) [15] が提案されている。これは、画像上で逆拡散過程を学習するのではなく、変分オートエンコーダ (VAE) のエンコーダで低次元化された潜在空間において逆拡散過程を学習させる手法である。これにより、生成時の計算量が大幅に削減され、交差注意 (cross attention) による条件付けも容易になっている。

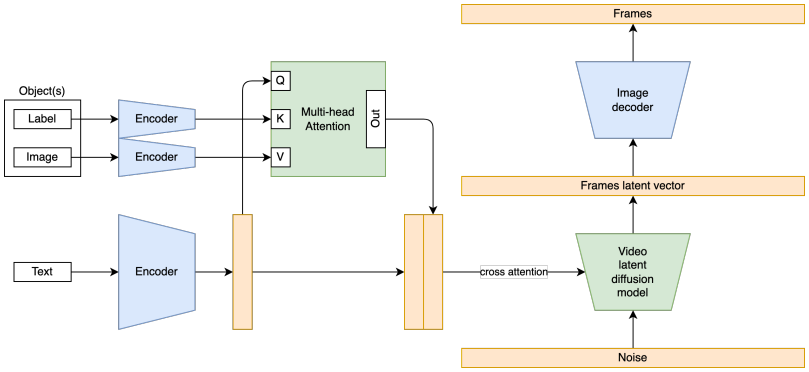

図 1: 提案手法の概要．青はバックボーンの事前学習済みモデル、緑は学習対象を表す

さらに、ビデオを生成する拡散モデルとして、Video Diffusion Model [7] が提案されている。この手法では、固定長のビデオを生成し、より長いビデオについては時間方向の内挿か自己回帰的な外挿により拡張していく方法となっている。提案手法は、テキストからビデオを生成する潜在拡散モデルに基づく手法であり、外部知識を利用する点が異なっている。

2.2 外部知識による事前学習モデルの補強

外部知識を検索することで事前学習のみでは網羅できない世界知識を補強する手法が提案されている。REALM [5] および RAG [9] は、自然言語処理における質問応答等の種々のタスクにおいて、Wikipedia の密ベクトル索引をニューラル検索で取得し、回答を生成する手法である。外部知識で世界知識を補強することにより、回答の正解率が大きく改善することが確認されている。同様の効果は、画像キャプション生成 [13] においても確認されている。

Re-Imagen [3] および RDM [1] は、画像生成タスクにおいて外部知識による世界知識の補強を導入した手法である。外部知識を用いない従来手法では、珍しいエンティティなどの画像を適切に生成できない課題があったが、外部知識で補強することにより、言及されたエンティティに対する視覚的な詳細が強化され、より適切な外観の画像を生成することができるとされている。提案手法は、ビデオ生成タスクにおいて外部知識を利用する点が異なっている。

3 提 案 手 法

3.1 モデルアーキテクチャ

提案手法は、外部知識埋め込み部とフレーム生成部で構成される。提案手法の概要を図 1 に示す。外部知識埋め込み部は、エンコードしたテキストと関連があるトークンに外部知識としてのインスタンス情報を付与し、フレーム生成部は、エンコードしたテキストとインスタンス情報を参照し、フレームを生成する。

3.2 外部知識埋め込み部

外部知識埋め込み部は、テキストに関連した外部知識を参照し、テキスト埋め込みを補強する役割を担う。これにより、事前学習済みモデルでは十分に網羅されていない珍しいエンティティに関する情報が強化され、より適切な外観の画像を生成することにつながると期待される。

外部知識埋め込み部への入力は、生成したい画像を表現したテキストと、インスタンス画像とその説明文テキストのペアの集合となる (図 1)。インスタンス画像とその説明文テキストのペアの集合を外部知識として利用する。生成したい画像を表現したテキスト中のトークンごとに、外部知識のインスタンスとの関連に応じた補足情報の埋め込み結果を出力する。

外部知識埋め込み部の訓練段階の構造を図 2 に示す。マスクされたテキストの埋め込みが、関連知識の交差注意を介してトランスフォーマデコーダで再構築され、外部知識が各トークンに埋め込まれることを意図している。

提案手法では、トランスフォーマデコーダにおいて十分再現できない珍しいインスタンスを発見し、それに応じて外部知識を参照することでテキストの埋め込みを強化する。インスタンスに関する情報は、インスタンス分割された画像を用いる。

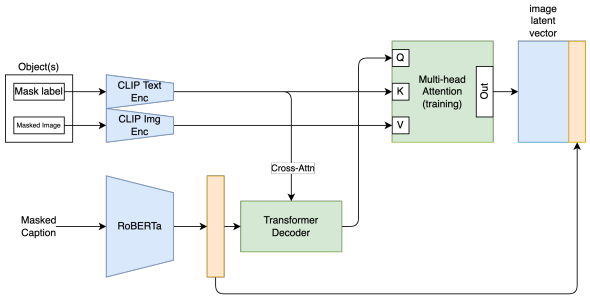

図 2: 外部知識埋め込み部

外部知識とテキストを関連するため、Trasnformer に紹介された Cross Attention を利用して、エンコードされた情報をオブジェクトの情報を入れ替える。このアプローチによって、各タームに外部知識が埋め込みができると想定している。

提案手法はまずデコーダで間違えた情報を見つかって、最後正しい外部情報を対応する場所で置いておくのデザインになる。

そのために、新しい、インスタンスを分離するモデルも提案する。

3.3 フレーム生成部

フレーム生成部は、エンコードしたテキストとインスタン

ス情報を参照し、フレームを生成する (図 4)。入力をトランスフォーマのエンコーダデコーダで変換し、拡散モデルの参照データとする。拡散モデルは参照データを用いてフレームを生成する。フーレム生成部の訓練には、埋め込み情報と参照フレーム、生成フレームの位置情報を入力とし、目標フレームを出力とする学習データを用いる。

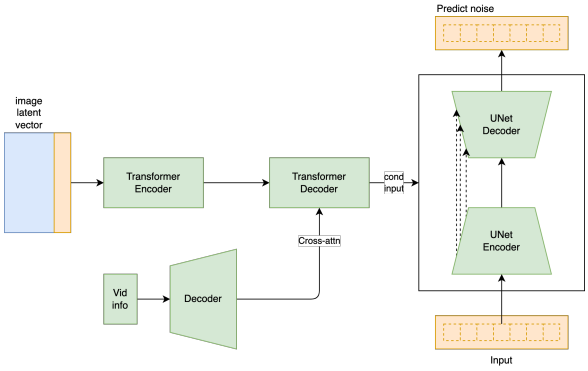

図 3: フレーム生成部

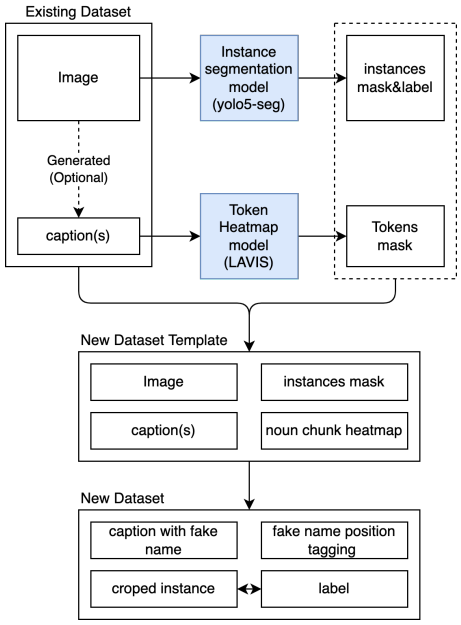

図 4: データセットの構築方法

4 実　　験

4.1 データセットの構築

既存のデータセットから提案手法に適したデータセットを構築する。ベースのデータセットとして MSVD [2] を用いた。このデータセットには、1750 件の動画と対応するキャプションが含まれている。外部知識埋め込み部のデータセットは、ランダムにフレームを抽出し、その時点の画像とキャプションで構成する。生成フレーム生成部のデータセットは、埋め込み部と同じ動画の別のフレームで構成する。

データセットの構築手順は以下の通りである (図 4)。画像とテキストからそれぞれインスタンスと名詞を抽出し、名詞とインスタンスのマッチングヒートマップを計算する。次に、インスタンスとの一致度がある閾値より高い場合、その名詞と関連付ける。生成したい画像を表現したテキストの名詞を、ランダムな名詞に置き換える。これにより、外部知識中のより具体的な名前を提起することを期待する。名前とインスタンスを外部知識として扱い、データセットを構築する。

4.2 実 験 設 定

テキストエンコーダとして RoBERTa [11] を、インスタンスエンコーダとして CLIP [14] をそれぞれ採用した。また、インスタンス分割に YOLOv5 [8]、画像ヒートマップの出力に LAVIS [10] を用いた。

また、モデルを複雑にしないため、トークンごとに対応づけできるインスタンス数の上限を 4 としている。

4.3 実験 1: 外部知識の埋め込み

4.3.1 実 験 目 的

外部知識の埋め込みが正しく機能しているかを検証する。

4.3.2 実 験 方 法

外部知識参照の有無により、検証データ誤差にどのような違いが見られるかを確認する。特に、珍しいトークンが含まれるテキストの場合について調査する。また、他のモデルと比較した場合の違いについても同様に確認する。

4.3.3 実 験 結 果

外部知識埋め込み部の訓練結果を図 5,6 に示す。エポックが進むにつれて検証データ誤差が減少し、訓練が進んでいることを確認した。ただ、誤差を十分に抑えることができていないことも確認された。

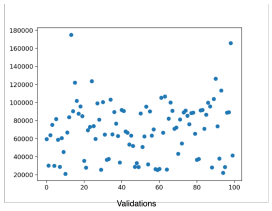

図 5: epoch0

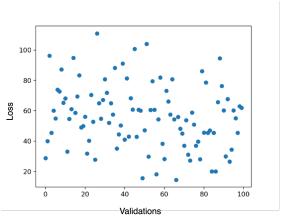

図 6: epoch5

4.4 実験 2: フレーム生成結果

4.4.1 実 験 目 的

外部知識を参照することでフレーム生成結果にどのような影響があるかを明らかにする。

4.4.2 実 験 方 法

外部知識の埋め込みの有無により、フレーム生成結果にどのような違いが見られるかを比較する。

4.4.3 実 験 結 果

連続 20 枚のフレーム生成結果を図 5,6 に示す。現状では意味のある内容の画像が生成されていない。外部知識の埋め込み自体のデザインが不十分である可能性があり、それを参照する

フレーム生成結果も意味のある内容とはならなかったと考えられる。

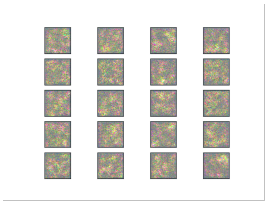

図 7: epoch0

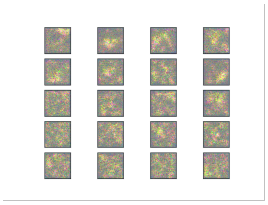

図 8: epoch5

5 考　　察

外部知識埋め込み部については、モデルの学習が進んでいることは確認できたものの、検証誤差を十分に抑えることができていないことを確認した。モデル自体のデザインが不十分であると考えている。今後は、インスタンス情報を個々に付与するのではなく、固定長ベクトルにまとめて付与するなどの方法、また、フレーム生成モデルと同時に学習を進める方法について検討を進める予定である。

フレーム生成部に関して、バックボーンには事前学習済みモデルを用いたが、生成した結果はノイズのみとなり、意味のある画像は生成できなかった。外部知識の埋め込み自体に問題がある可能性があり、それを参照するフレーム生成結果も意味のある内容とはならなかったと考えられる。また、フレーム情報を感知する部分が単純すぎる可能性があると考えている。今後は、事前学習済みモデルの微調整 [8] や、前後のフレーム情報を参照する機能を追加する方法について検討を進める予定である。また、画像全体を生成するのではなく、インスタンスごとに段階的に生成する方法について検討する予定である。

データセットについても、Flickr30K Entities Dataset など、インスタンス追跡とそのテキストとが関連づけられたビデオデータセットを活用し改善を図る予定である。

6 ま と め

本稿では、特定オブジェクトに関する情報を保存した外部知識を参照することで、訓練時には網羅できていないオブジェクトについても忠実に反映することができるビデオ生成手法を提案した。これにより、事前学習済みモデルでは十分に網羅されていない珍しいエンティティに関する情報が強化され、より適切な外観の画像を生成することにつながると期待される。

実験の結果、モデルの学習が進んでいることは確認できたものの、検証誤差を十分に抑えることができず、意味のある画像を生成するには至らなかった。

今後は、外部知識の補強方法、フレーム生成方法のそれぞれについて改良を進める予定である。

謝　　辞

本研究の一部は科研費 18K11557 の助成を受けたものである.

文　　献

[1] Andreas Blattmann, Robin Rombach, Kaan Oktay, and Björn Ommer. Retrieval-augmented diffusion models, 2022.

[2] David L. Chen and William B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011)*, Portland, OR, June 2011.

[3] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[5] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR, 13–18 Jul 2020.

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[7] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

[8] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, Kalen Michael, TaoXie, Jiacong Fang, imyhxy, Lorna, 曾逸夫 (Zeng Yifu), Colin Wong, Abhiram V, Diego Montes, Zhiqiang Wang, Cristi Fati, Jebastin Nadar, Laughing, UnglvKitDe, Victor Sonck, tkianai, yxNONG, Piotr Skalski, Adam Hogan, Dhruv Nair, Max Strobel, and Mrinal Jain. ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation, November 2022.

[9] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[10] Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C. H. Hoi. Lavis: A library for language-vision intelligence, 2022.

[11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[12] Gautam Mittal, Jesse Engel, Curtis Hawthorne, and Ian Simon. Symbolic music generation with diffusion models. *arXiv preprint arXiv:2103.16091*, 2021.

[13] Sofia Nikiforova, Tejaswini Deoskar, Denis Paperno, and Yoad Winter. Generating image captions with external encyclopedic knowledge. 2022.

[14] 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*, pages 8748–8763. PMLR, 2021.

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.