AI 特許紹介(94)
AI 特許を学ぶ！究める！
～Video-Adapter 特許～

2026 年 11 月 10 日
河野特許事務所
所長弁理士　河野英仁

「AI 特許紹介」シリーズは、注目すべき AI 特許のポイントを紹介します。熾烈な競争となっている第４次産業革命下では AI 技術がキーとなり、この AI 技術・ソリューションを特許として適切に権利化しておくことが重要であることは言うまでもありません。

AI 技術は Google, Microsoft, Amazon を始めとした IT プラットフォーマ、研究機関及び大学から毎週のように新たな手法が提案されており、また AI 技術を活用した新たなソリューションも次々とリリースされています。

本稿では米国先進 IT 企業を中心に、これらの企業から出願された AI 特許に記載された AI テクノロジー・ソリューションのポイントをわかりやすく解説致します。

1.概要
特許出願人　Google
出願日　2024 年 5 月 17 日
公開日　2024 年 11 月 21 日
公開番号 US20240386529
発明の名称　拡散モデルを用いたドメイン固有の動画の生成

529 特許はテキストプロンプトを大規模なノイズ除去拡散モデルに入力することにより中間表現を生成し、並行して特定領域に特化した小規模のノイズ除去拡散モデルにテキストプロンプトを入力して中間表現を生成し、双方の中間表現を組み合わせることで動画を生成する Video-Adapter 技術に関する。

2.特許内容の説明
下記図は、動画生成システム 100 を示す。

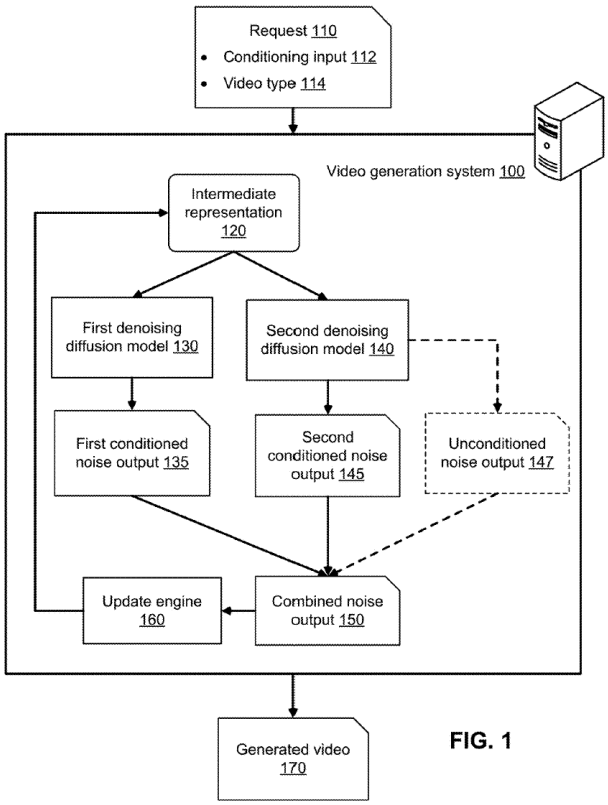


FIG. 1

動画生成システム 100 は、出力動画 170 を生成するためのリクエスト 110 を受け付ける。リクエスト 110 は、特定の所望の出力ビデオ 170 に向けてビデオ生成プロセスを誘導するガイダンスを提供するコンディショニング入力 112 を含む。コンディショニング入力 112 は、出力ビデオ 170 に何を含めるべきかを表すテキスト記述である。例として、コンディショニング入力 112 は、「家の中で猫がネズミを追いかけるビデオを生成せよ」といったテキストプロンプトである。

リクエスト 110 は、出力ビデオ 170 の特定のビデオタイプ 114 を指定することができる。ビデオタイプ 114 は、出力ビデオ 170 の特定のスタイル、特定の用途、特定のコンテンツ形式、あるいは特定の主題や分野を指定する。動画 170 を生成するために、動画生成システム 100 は、複数の反復の各々において中間表現 120 を反復的に更新する。各反復における中間表現 120 は、ノイズが付加された出力動画 170 である。

システム 100 は、各反復において、その反復に対応するノイズの推定値を除去することにより、中間表現を更新する。すなわち、システム 100 は、各反復においてノイズの推定値を決定し、その推定値に従って現在の中間表現を更新することで、中間表現 120 を精緻化する。システム 100 は、出力ビデオ 170 を出力するまで、反復を降順で行う。

第 1 のノイズ除去拡散モデル 130 は、現在の反復における中間表現 120 を処理し、条件付け入力 112 に基づいて条件付けされた第 1 のノイズ出力 135 を生成する。第 2 のノイズ除去拡散モデル 140 は、現在の中間表現を処理し、条件付け入力 112 に基づいて条件付けされた第 2 のノイズ出力 145 を生成する。

現在の反復において第 1 の条件付きノイズ出力 135 および第 2 の条件付きノイズ出力 145 が生成された後、システム 100 は、少なくとも(i)第 1 の条件付きノイズ出力 135 および(ii)第 2 の条件付きノイズ出力 145 を組み合わせることによって、合成ノイズ出力 150 を生成する。システム 100 はさらに、第 2 のノイズ除去拡散モデル 140 を用いて、現在の中間表現 120 を含む第 3 の入力を処理することにより、非条件付きノイズ出力 147 を生成する。非条件付きノイズ出力 147 は、条件付け入力 112 に基づく条件付けがなされていない。

次いで、システム 100 は、(i)第 1 の調整済みノイズ出力 135、(ii)第 2 の調整済みノイズ出力 145、および(iii)未調整ノイズ出力 147 を組み合わせる（例えば、それらの加重和を算出する）ことによって、合成ノイズ出力 150 を生成する。

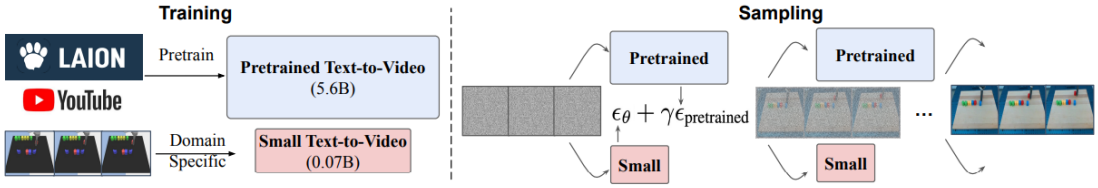
第 1 のノイズ除去拡散モデル 130 は、第 2 のノイズ除去拡散モデル 140 と比較して、より大きなモデルである。すなわち、第 1 のノイズ除去拡散モデル 130 は、第 2 のノイズ除去拡散モデル 140 よりも多くのモデルパラメータを有している。

大規模なデータセット（例えば、インターネット規模のデータ）で学習された大規模かつ汎用的な動画生成モデル（例えば、第 1 のノイズ除去拡散モデル 130）は、高忠実度の動画を生成できることが示されているものの、それらによって生成された動画は、特定の動画タイプ 114 には適合していない。一方、ドメイン特化型の動画生成モデル（例えば、第 2 のノイズ除去拡散モデル 140）は、ドメイン特化型の学習例を用いて学習させることが可能であるものの、柔軟性に欠ける上、ドメイン特化型の学習例や計算リソースの利用可能性が限られているといった制約から、通常はより小型のモデルとなる。

システム 100 は、事前学習済みモデル 130 とドメイン特化型モデル 140 によって生成されたノイズ出力を組み合わせることで、大規模なドメイン特化型生成モデルの学習、または、追加のドメイン特化型学習データを用いた大規模な事前学習済みモデルのファインチューニングを行うことなく、特定の動画タイプ 114 の高品質な出力動画 170 を生成できる。大規模な動画生成モデルの再学習またはファインチューニングには、計算上の大きな課題が伴う。特に、事前学習済みの大規模モデルをファインチューニングす

る場合、多大な計算リソースを必要とするだけでなく、モデルの過学習に起因する性能低下を招く恐れがある。本技術では、大規模なドメイン特化型モデルの学習、または、ドメイン特化型学習データを用いた大規模モデルのファインチューニングといった工程が不要となるため、レイテンシの改善及び計算リソース要件の低減を実現できる。これは、拡散プロセスにおいて、事前学習済みモデルとドメイン特化型モデルによって生成されたノイズ出力を組み合わせることで達成される。

下記図は動画生成システムのワークフロー例を示す。



事前学習済みのテキスト・動画変換モデル(Pretrained）は、インターネット上の動画例など、多種多様な多数の動画サンプルを用いて学習されている。小規模なテキスト・動画変換モデル（Small）は、ドメイン固有の動画（すなわち、特定のタイプに属する動画）からなる小規模なデータセットを用いて学習されている。

複数の反復のそれぞれにおいて、事前学習済みモデルは、中間表現を処理して第 1 のノイズ出力 $\epsilon_{\text{pretrained}}$ を生成する。小型モデル 240 は、中間表現を処理して第 2 のノイズ出力 ϵ_θ を生成する。システムは、第 1 および第 2 のノイズ出力を組み合わせて、合成ノイズ出力 $\epsilon_\theta+\gamma\epsilon_{\text{pretrained}}$ を生成する。システムは、その合成ノイズ出力を用いて、次の反復のための中間表現を更新する。最終的な反復の後、システムは、更新された中間表現を生成ビデオ出力として出力する。

3.クレーム

529 特許のクレーム 1 は以下の通りである。

1. 入力に基づいて出力ビデオを生成するための、コンピュータによって実施される方法において、

条件付け入力を受け取り、

出力ビデオの現在の中間表現を初期化し、

複数の反復の各々において、現在の中間表現を更新し、

前記更新は以下を含み、

前記条件付け入力によって条件付けられた第 1 のノイズ除去拡散モデルを用いて、前記現在の中間表現を含む第 1 の入力を処理することにより、第 1 のノイズ出力を生成

し、前記第１のノイズ除去拡散モデルは、第１の複数の学習ビデオを含む第１の学習データを用いて学習されており、

前記条件付け入力によって条件付けられた第２のノイズ除去拡散モデルを用いて、前記現在の中間表現を含む第２の入力を処理することにより、前記条件付け入力に基づいて条件付けられた第２のノイズ出力を生成し、前記第２のノイズ除去拡散モデルは、(i)前記第１のノイズ除去拡散モデルとは異なり、かつ(ii)第２の複数の学習ビデオを含む第２の学習データを用いて学習されており、

少なくとも(i)前記第１のノイズ出力および(ii)前記第２のノイズ出力を組み合わせることにより、合成ノイズ出力を生成し、

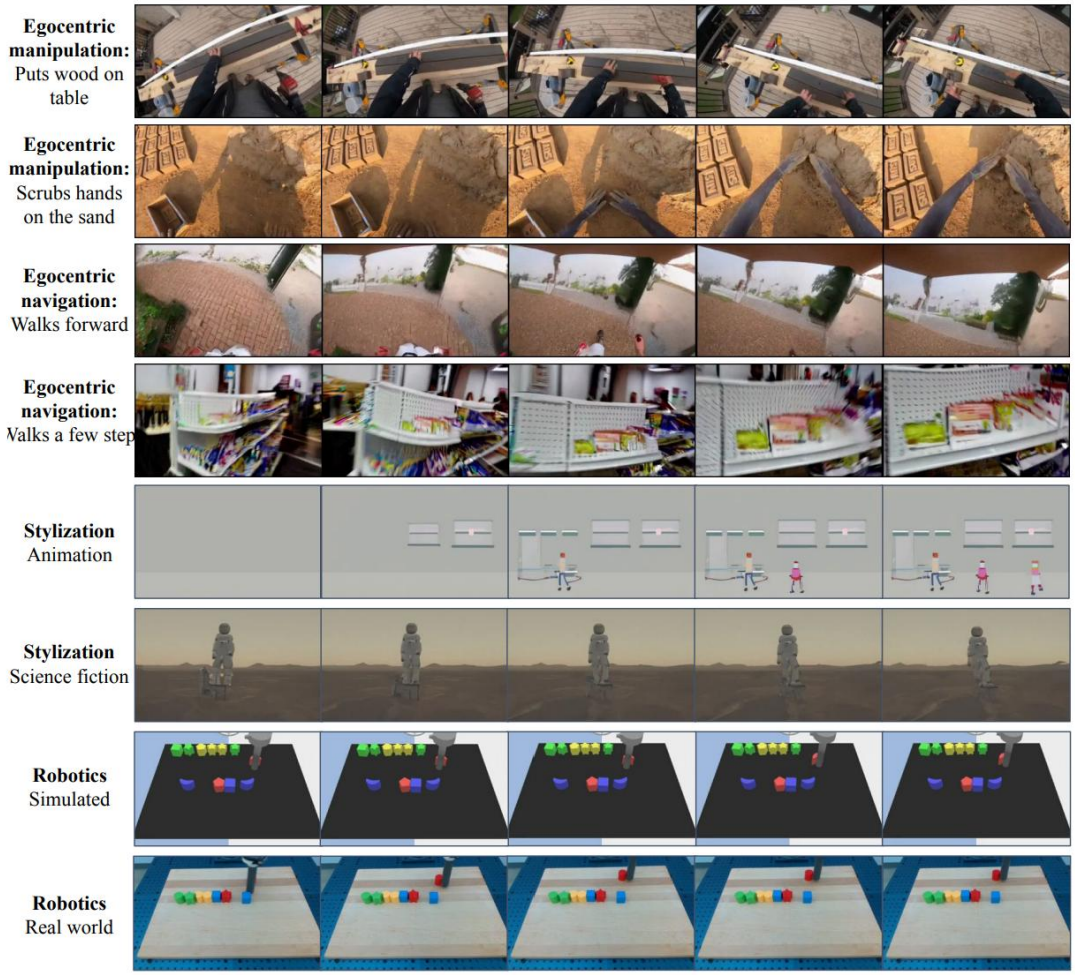
当該反復について生成された前記合成ノイズ出力を用いて、前記現在の中間表現を更新し、

最終的な反復の後に、前記現在の中間表現から前記出力ビデオを生成する。

4. 本特許に関連する論文

本特許に関する論文 “Probabilistic Adaptation of Text-to-Video Models” [1]が、Google Deep Mind の Mengjiao Yang 氏らにより公表されている。下記図は Video Adapter による動画生成例を示す。

[1] Arunkumar Byravan, et al. “TOWARDS A UNIFIED AGENT WITH FOUNDATION MODELS” rXiv:2307.09668v1 [cs.RO] 18 Jul 2023



図に示すように、Video Adapter は、操作またはナビゲーションに基づく一人称視点の動きを含む動画、アニメ及び SF といった独自のスタイルを反映した動画、さらにはシミュレーション及び実機によるロボットの動画など、多様なスタイルを持つ動画を柔軟に生成することができる。

下記表にビデオモデリングにおける定量的性能を示す。

	Bridge			Ego4D		
Model	FVD ↓	FID ↓	Param (B)↓	FVD ↓	IS ↑	Param (B) ↓
Small (S)	186.8	38.8	0.07	228.3	2.28	0.07
Small (S) + Pretrained	**177.4**	**37.6**	0.07	156.3	2.82	0.07
Small (L)	152.5	30.1	0.14	65.1	3.31	2.8
Small (L) + Pretrained	**148.1**	**29.5**	0.14	**52.5**	**3.53**	2.8
Pretrained	350.1	42.6	5.6	91.7	3.12	5.6

異なる動画生成技術の性能を評価するために、2 つのドメイン固有のデータセットが使用される。Bridge データセットには、事前学習データの分布外にあるロボットのタスク固有の動画が含まれている。Ego4D データセットには、インターネット上では一

般的に見られない、主に一人称視点（エゴセントリック）の動画が含まれている。事前学習済みモデルとしては、インターネット上の汎用的な動画を用いて学習させた、56 億（5.6B）パラメータのベースモデルが使用される。ドメイン固有のモデルについては、事前学習済みモデルからパラメータ数を 80 分の 1 または 40 分の 1 に縮小した拡散モデルが用いられる。各技術について、生成された出力動画に対し、FVD(Frechet Video Distance)及び FD（Frechet Distance）を含む動画品質指標が算出される。

表に示すように、Bridge データセットにおいて、事前学習済みモデルの 1.25%のパラメータ数を持つ小規模モデルを学習させた場合（1 行目）、すでに事前学習済みモデルよりも良好な指標が達成されている。しかし、事前学習済みモデルを確率的プライア（事前分布）として組み込むことで、小規模モデルの指標はさらに向上する（2 行目）。Ego4D においては、一人称視点動画の複雑さゆえに、事前学習済み動画モデルの 1.25%の規模を持つ最小モデルでは、もはや事前学習済みモデルを上回る性能を達成できない（1 行目）。しかし、サンプリング時に事前学習済みモデルを組み込むことで、依然として性能は向上する（2 行目）。

Video Adapter は、下記サイトにて紹介されている。
https://video-adapter.github.io

著者紹介
河野英仁
河野特許事務所、所長弁理士。立命館大学情報システム学博士前期課程修了、米国フランクリンピアースローセンター知的財産権法修士修了、中国清華大学法学院知的財産夏季セミナー修了、MIT(マサチューセッツ工科大学)コンピュータ科学・AI 研究所 AI コース、生成 AI ビジネスコース修了。MIT スローン経営大学院 Agentic AI コース修了。
生成 AI・エージェント型 AI 特許コンサルティング、AI 特許コンサルティング、フィジカル AI 特許コンサルティング、医療 AI 特許コンサルティングの他、米国・中国特許の権利化・侵害訴訟を専門としている。著書に「世界のソフトウエア特許(共著)」、「FinTech 特許入門」、「AI/IoT 特許入門 4」、「ブロックチェーン 3.0(共著)」がある。