進化する動画生成ＡＩ（前編）

相次ぐ新モデル・サービスの登場

上級研究員　内田真穂

動画生成 AI（人工知能）が急速に進化し、実写さながらのリアルな映像を生成できる動画生成 AI サービスが次々と登場している。現状ではまだ不完全な部分もあるが、誰でも高品質な映像を作成できる時代が確実に近づいている。まず前編で動画生成 AI の進化の歴史と主要サービスの現在地を確認し、後編で動画生成 AI の将来の展望と課題について考察する。

１．はじめに

近年、動画生成 AI が目覚ましい進化を遂げている。その結果、これまで専門的技術や高性能な機材が必要だった動画制作が、誰にでも手が届くものへと変わりつつある。本稿では、動画生成 AI の進化の歴史と代表的な動画生成 AI サービスの現在地を確認する。

２．動画生成 AI の進化の歴史

動画生成 AI は画像生成 AI と共に進化し、深層学習技術の発展やコンピュータの性能向上により急速に進化してきた。

（１）初期の動画生成技術（2010 年代前半）

動画生成 AI の始まりは画像を生成する技術からスタートした。2014 年に登場した GAN（敵対的生成ネットワーク）は、二つのネットワーク（生成ネットワークと識別ネットワーク）が互いに競い合いながら学習する技術で、これにより「本物のような画像」を作り出すことに成功した。この技術をもとに、初期の動画生成 AI は、静止画を連続してつなぎ合わせるアプローチで開発が進められた。

（２）進化の第一段階（2016 年～2019 年）

2016 年以降、ディープラーニング技術が発展し、動画生成 AI はさらに自然な動画を生成できるようになった。2018 年頃からは、既存の動画を学習し、短い動画を自動的に生成する AI が登場した。この時期には、単に静止画を繋げるだけでなく、「動画としてのつながりや流れ」を意識した生成が行われるようになった。

（３）進化の第二段階（2020 年～2023 年）

2020 年代に入ると、計算能力の向上と新たな深層学習モデルの導入により、さらに長尺で高品質な映像を用いた動画が作れるようになった。これには Transformer の導入によって動画内時間の流れの理解に基づく一貫性のある映像作りが可能となったことが大きく貢献している。Transformer は、もともと自然言語処理（NLP）の分野で開発され、ChatGPT などの対話型 AI の基盤となった技術でもある。さらに、拡散モデル（Diffusion model）の導入により、生成する映像の解像度が著しく向上した結果、精細な動画の生成が可能となった。拡散モデルは、画像データにノイズを加えた後、そのノイズを段階的に除去して綺麗な画像を

生成する手法であり、テキストから画像を生成する AI である DALL-E 2[1]や Stable Diffusion[2]で採用されている。この技術はフレーム間の補完にも応用され、前後のフレームをより自然に接続し、滑らかさを大幅に向上させた。

2022 年に登場した DALL-E 2 や Stable Diffusion などの画像生成 AI は、動画生成 AI の発展の基盤となった。2023 年には米 Runway の「Gen-1」「Gen-2」や米 Google DeepMind の「Imagen Video」などが登場し、動画の品質と生成スピードの向上が加速した。

（４）最近の進化（2024 年～）

2024 年に入ると、これまでの技術革新が一気に集まって、テキストによる指示だけで高品質な動画を生成することが現実になった。特に米 OpenAI の「Sora」は、それ以前の動画生成 AI を凌駕する高精度かつリアルな動画を生成する画期的なツールとして注目を集めた。Sora はシーン全体の流れやストーリー性を考慮した映像生成を実現した。こうした近年の飛躍的な進化の背景には、自然言語処理（NLP）技術の発展に加え、大規模言語モデル（LLM）の進化[3]とデータセットの巨大化がある。さらに、テキスト、画像、音声、動画を統合するマルチモーダル AI の技術も進化を続けている。つまり、映像動画そのものだけでなく、音声や字幕を含めた一貫性のある動画コンテンツを総合的に生成できる時代が到来しつつある。

３．主要プレイヤーとその動画生成 AI サービス

Sora の登場以降、市場では動画生成 AI サービスの開発競争が激しさを増している。特に 2024 年末から 2025 年初頭にかけて、新たなモデルのリリースや機能のアップデートが相次いだ。以下では、「Sora」、「Veo2」、「Dream Machine」、「Runway」、「Kling」について、実際に使ってみた感想を交えながら特徴を紹介する。

（１）Open AI「Sora」― 高画質のリアルな動画を実現

OpenAI が 2024 年 2 月に発表した「Sora」は、テキストを入力するだけで高画質のリアルな動画を生成し、動画生成 AI が一躍注目されるきっかけとなったツールである。複数の被写体や特定の動き、背景との位置関係を崩すことなく複雑なシーンを生成できるのが特徴で、それ以前の動画生成 AI との大きな違いでもある。Sora は発表から約 10 か月後、2024 年 12 月 9 日に正式にリリースされた。

≪図表１≫Sora で作成した動画

プロンプトで動きは特に指示していなかったが、首を左右に振る動画が生成された。
（出典）Sora で著者作成

試しに「お内裏様とお雛様に扮する犬と猫」というプロンプトを入力すると、着物を羽織った犬と猫が和室でお行儀よく座り、首を左右に振る動画が 10 秒足らずで生成された（図表１）。わずか 5 秒という短い動画ながらも、その動きは驚くほど自然だった。

Sora には画像から動画を生成する機能もある。試しに二羽の鳥の写真をアップロードして「birds are

[1] OpenAI が 2022 年 4 月に発表した画像生成 AI ツール
[2] 英スタートアップ Stability AI が開発した画像生成 AI サービス。2022 年 8 月にオープンソース形式で公開され、画像生成 AI ブームの火付け役となった。
[3] 例えば、OpenAI の最先端の AI モデル GPT-4o は数百億から数千億の範囲にあると推測されている。正確なパラメータ数が公開されていないが、GPT-3.5 のパラメータ数が約 1750 億であり、GPT-4o はそれを大きく上回ると考えられている。Google Gemini のフルスペックモデルである Ultra は 1.6 兆パラメータである。

eating food」と入力すると、鳥が餌をついばみ、池の水を飲むという動きが加わった（図表２）。自分が撮影した鳥が写真の枠外に移動して池の中に入ったことは、著者の想像を超えたことであり、驚愕したことを記しておきたい。

≪図表２≫Sora で静止画（左）から動画（右）を作成

つがいの鳥
（画像：著者撮影）

奥にいる一羽が動き出し、移動した先に別の一羽がいた。
（動画：sora で著者作成）

Sora の利用には ChatGPT の有料プランへの加入が必要だ。2025 年 3 月 20 日現在、720p・5 秒間の動画を月 50 本まで生成できる月額 20 ドルのプラン（ChatGPT Plus）と、1080p・10 秒または 720p・20 秒の動画を月 500 本まで生成できる月額 200 ドルのプラン（ChatGPT Pro）がある。

（２）Google DeepMind「Veo2」― ４K 解像度でシネマティックな映像表現

Sora の一般公開から 1 週間後の 2024 年 12 月 16 日、米 Google DeepMind が「Veo2」を発表した。Veo2 の特徴は 4K 解像度の高画質と、細部までリアルな映像表現にある。カメラアングルを指定できるため、さまざまな視点から被写体の動きを捉えることができ、まるで映画のような映像を作成できる。その性能は他社を凌駕し、Meta が提供するベンチマークテスト「MovieGenBench」では、プロンプトの忠実度と映像の質の両面でトップ評価を得た[4]。

YouTube の公式チャンネルでは多数のデモ動画が公開されおり、人物の表情や動物の動き、水の描写、物体の質感などが極めてリアルに表現されている（図表３）。2025 年 2 月下旬より、日本でも一部の開発者向けプラットフォームで提供が開始された。早速試してみたところ、人物の表情や動きのリアルさには目を見張るものがあった。ただし、不自然なモーションや描写の破綻が見受けられる場合も多かった。また、数分間の長尺動画を技術的には生成可能と説明されているが、現時点では一般ユーザーの利用は最大 5 秒までの動画に制限されている。

≪図表３≫Veo2 のサンプル動画

Veo2 demo | Swimming dog
（出典）Google DeepMind 公式動画

なお、現在はテキストによる指示から動画を生成する「Text to Video」機能のみに対応しているが、今後は画像から動画を生成する「Image to Video」機能の追加が予定されている。

（３）Runway「Gen-3 Alpha」「Gen-3 Alpha Turbo」― 多彩な編集機能を搭載

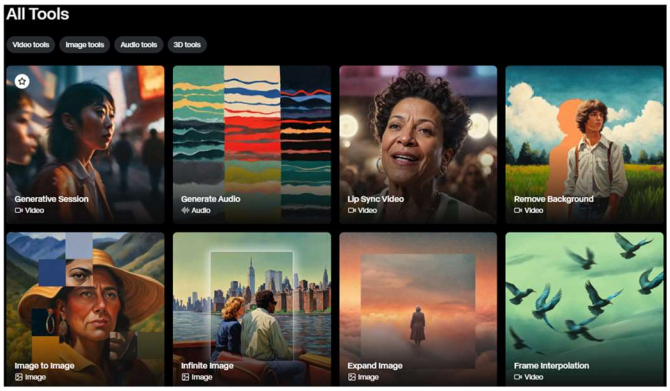
Runway は 2018 年創業のニューヨークを拠点とするスタートアップ企業で、動画生成 AI ツール「Runway」を提供している。2024 年 7 月から 8 月にかけてリリースされた「Gen-3 Alpha」「Gen-3 Alpha Turbo」は

[4] https://deepmind.google/technologies/veo/veo-2/

テキストと画像から 5 秒または 10 秒の動画の生成が可能である[5]。生成時間は 10 秒超～30 秒程度と早く、他の動画生成 AI と同様にカメラワークをプロンプトで指定できる。

Runway の特徴は多彩な編集機能にある（図表 4）。例えば、人物の口元の動きと音声の言葉を一致させる「リップシンク」、「背景の除去」や「オブジェクト削除」などの高度な編集機能があるほか、他のユーザーとの共同編集機能も備えている。スマートフォンで撮影した人物の表情を、アニメーションキャラクターに転換させる「Act-One」と呼ばれるユニークな機能もある。動画生成自体は直感的に行えるので、初心者から上級者まであらゆるレベルのユーザーが楽しめる。無料のお試しプランからビジネスユース向けのエンタープライズプランまで 5 種類のプランがある。

≪図表 4 ≫多彩な機能が特徴の Runway



Runway で利用できる機能が一覧化された「AI Tools」
（出典）Runway のプラットフォーム

（4）Luma AI「Ray2」― 感情表現と動作の一貫性に優れる

Luma AI は 2021 年に設立されたサンフランシスコを拠点とするスタートアップ企業で、動画生成プラットフォーム「Dream Machine」を提供している。2025 年 1 月にリリースされた最新モデル「Ray2」は、1080p の高解像度での出力に対応し、10 秒までの動画生成が可能である。

Ray2 は、登場人物の感情表現や、自然で一貫性のある動きの表現に優れており、実際に使用してもそのクオリティの高さを実感できた（後述の BOX 参照）。リリース当初はテキストによる指示から動画を生成する「Text to Video」機能に特化していたが、現在では「Image to Video」機能が追加され、さらに生成した動画にオリジナルの音楽を追加する「Audio」機能も実装されている。

公式ウェブサイトには、静止画から作成したサンプル動画が多数掲載されており、例えばスケボーでキックフリップする（板を縦に一回転させる）僧侶や、火を噴くドラゴンなど、自然で違和感のない動きが表現されている（図表 5）。なお、無料プランは生成速度が遅いため、有料プランの利用が推奨される。

≪図表 5 ≫Luma Ray2 のサンプル動画

LumaRay2 のウェブサイトには多数のサンプル動画が掲載されている。
（出典）https://lumalabs.ai/ray

（5）快手科技「Kling AI」― 高精度なプロンプト理解力

AI の開発競争では中国が米国を猛追している。中国の快手科技の動画生成 AI「Kling 」の最新バージョン Klingv1.6（2024 年 12 月リリース）が、Sora や Runway を超える高性能モデルとして注目されている。テキストおよび画像から動画生成でき、1080p の高画質で、動画の長さは 5 秒または 10 秒を選択できる。

[5] 「Gen-3 Alpha Turbo」は「Gen-3 Alpha」の高速版で Gen-3 Alpha の 7 倍のスピードで動画を生成し、かつ消費クレジットが少ない。ただし、完成度は Gen-3 Alpha のほうが高いという評価もあり、優劣はつけられない。

特に優れているのはプロンプトの理解力で、他の動画生成 AI と比べて破綻が少なく、精度が高いと感じた。例えば、「公園で犬の散歩をするヒューマノイドロボット」を生成した際、リードが尻尾に繋がるハルシネーションは見られたものの、ロボットや犬の歩き方に不自然さはなかった。「編み物をするウサギ」や「空を飛ぶ流線形の電車」など、架空の世界を描くプロンプトも忠実に表現された（図表 6）。

≪図表６≫Kling で生成した動画

（出典）Kling で筆者作成

Kling の特徴として「Creativity⇄Relevance」という独自の調整機能を搭載し、AI の創造性を重視するか、プロンプトに忠実な動画を生成するかを選択できる点が挙げられる。意図しない動画が生成されないよう、不要な要素や避けたい描写を排除するネガティブプロンプト機能も備えている。また、無料プランと有料プランで動画の品質や生成所要時間が変動するシステムを採用しており、筆者が利用したスタンダードプランでは、生成所要時間は 4 分から 8 分だった。なお、無料プランでは 30 分待っても生成されず、現状では有料プランを使うのが現実的と感じた。

４．実際に利用してみて

本稿では５つの動画生成 AI ツールを紹介したが、多くのユーザーが最初に直面する壁はプロンプト作成だろう。特に公開されているデモ動画のような高品質な作品を作るには、登場人物の動き、天候、背景、光の当たり方、カメラアングルなど、詳細な指示をする必要がある。プロンプトの作成には補助ツールを利用する手もある。例えば、ChatGPT の『GPTs』[6]に載っている、「Veo2Promter」や「Master Video Prompt Creation KLING AI 1.6」などがそれだ。

また、多くのサービスがフリーミアムモデル[7]を採用している。無料プランでは出力回数や解像度が制限されているのに対して、有料プランではこれらの上限が拡大され、商用利用を許可しているものもある。価格帯も幅広く、個人向けの数千円程度のプランから大企業向けの高額プランまで選択肢があり、ユーザーは自分の用途に応じて柔軟に選べるようになっている。

５．おわりに

現在の動画生成 AI は、実写と区別がつかないほどリアルな映像を生成できる段階に達している。高品質な動画を作るには一定のスキルが求められるが、今後、技術が進化し、より直感的な操作が可能になれば、多くの人が高品質な映像を作成できるようになるだろう。動画生成 AI は映像制作のあり方を根本から変える可能性があり、多様な分野での活用が期待される。次稿では、広告、企業 PR や SNS での活用事例を取り上げるとともに、今後の展望と課題について掘り下げる。

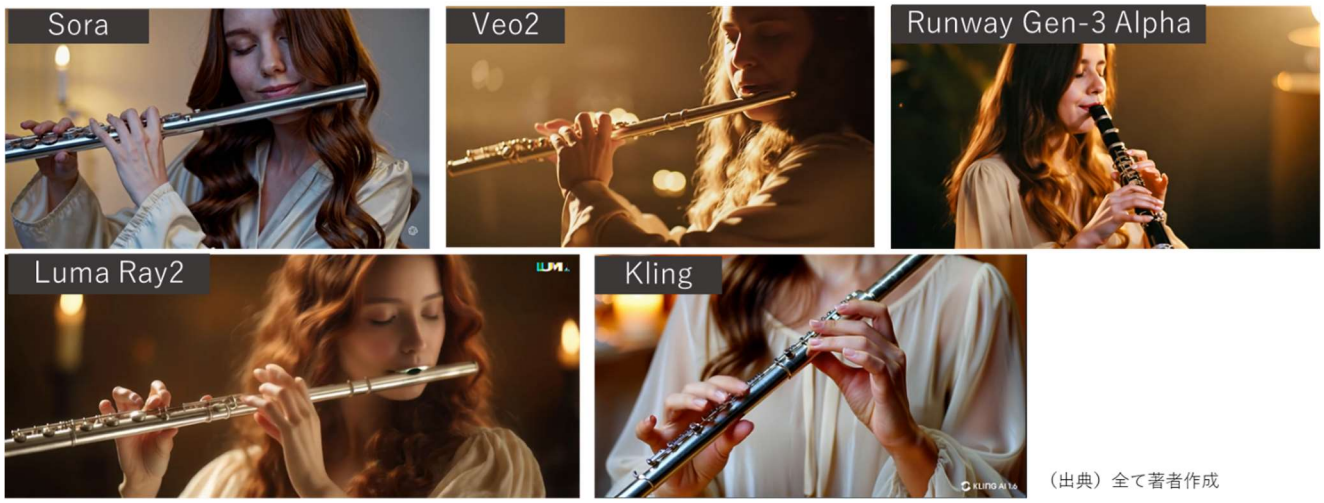
---

[6] GPTs とは、ChatGPT を特定の目的に合わせて自分の好みにカスタマイズできる機能であり、ユーザーが作成した便利なチャットボットが多数公開されている。

[7] 「フリー（無料）」と「プレミアム（割増料金）」の造語で、基本的なサービスや製品を無料で提供し、さらに高度なサービスや機能に関しては有料で行う事により収益を得るビジネスモデルのこと。

《BOX》５つの動画生成 AI ツール　同じプロンプトで比較してみた

本稿で取り上げた５つの動画生成 AI ツールに同じプロンプトを入力して動画を作成してみた。今回作成したのは「フルートを演奏する女性」で、下の写真はその動画をキャプチャしたものである。



（出典）全て著者作成

プロンプト：A cinematic close-up shot captures a young woman with long, wavy chestnut hair playing the flute, her fingers gliding effortlessly over the instrument's silver keys. She wears a flowing ivory blouse, the soft fabric catching the warm glow of diffused golden light. Her eyes are closed, lost in the melody, as her breath flows through the flute, creating a serene, ethereal atmosphere. The camera moves in a slow, steady tracking motion, emphasizing the elegance of her performance. The blurred background shimmers with hints of candlelight, adding to the intimate and mesmerizing scene.

Sora はフルートを構える姿勢と指の動きは表現できていたが、表情や動きに息づかいが感じられず、リアリティに欠けていた。Veo2 はフルートの持ち方が縦笛のようになっており、Runway に至ってはフルートではなく、クラリネットのような別の楽器になってしまっていた。Kling は楽器を持つ手の形こそ自然だが、楽器の形状が不自然で、「フルートを演奏している」という肝心な部分が表現されていなかった。

最も完成度が高かったのは Luma Ray2 だ。指の形状にやや違和感があるものの、表情や指の動きは本当に演奏しているように感じられた。アスペクト比の違いが完成度に影響した可能性もある。Luma Ray2 には音声合成機能も備わっており、「Audio」をクリックするとフルートの音色が合成された。

もちろん、この一例だけで、どのツールが最も優れているかを断定することはできない。ただし、このケースに限らずいくつかのパターンを試してみた結果、個人的には Kling と Luma Ray2 が、クオリティと使いやすさのバランスに優れていると感じた。また、シンプルなプロンプトでも、詳細なプロンプトより高品質な動画が生成されるケースもあり、それぞれのツールの得手・不得手も見えてきた。