Vlogger論文和訳

タイトル: VLOGGER: Multimodal Diffusion for Embodied Avatar Synthesis

女性の写真のコラージュ

中程度の精度で自動的に生成された説明

図1 VLOGGERは、オーディオから人物を合成するための革新的なフレームワークです。第1列に示されるような単一の入力画像とサンプルオーディオ入力を与えると、私たちの方法はその人物の話す様子や生き生きとした動きを持つ、フォトリアルで時間的に整合したビデオを生成します。右列の合成された画像からも分かるように、私たちの方法では頭部の動き、視線、まばたき、唇の動きに加えて、従来の手法とは異なり上半身や手のジェスチャーも生成します。これにより、オーディオ駆動の合成技術をさらに一歩進めています。

はじめに

我々は、1枚の人物画像から音声駆動による人物ビデオ生成を行う手法「VLOGGER」を提案します。この手法は、最近の生成拡散モデルの成功に基づいています。我々の手法は、1) 確率的な人間-3Dモーション拡散モデル、および 2) テキストから画像へのモデルに空間的および時間的コントロールを加える新しい拡散ベースのアーキテクチャから構成されています。これにより、人間の顔や体の高レベルな表現を通じて容易にコントロール可能な高品質な可変長のビデオ生成がサポートされます。これまでの研究と対照的に、我々の手法は、各人物ごとにトレーニングを必要とせず、顔検出やクロッピングに依存せず、完全な画像（顔や唇だけでなく）を生成し、コミュニケーションする人間を正確に合成するために重要な幅広いシナリオ（例えば、見える胴体や多様な被写体のアイデンティティ）を考慮します。我々はまた、3Dポーズと表情のアノテーションが付いた新しく多様なデータセット「MENTOR」を収集し、以前のものより1桁大きい（80万のアイデンティティ）動的ジェスチャーを含むデータセットを収集し、これに基づいて主要な技術的貢献をトレーニングおよびアブレーションします。VLOGGERは、画像品質、アイデンティティ保持、時間的一貫性を考慮しながら、上半身のジェスチャーも生成し、3つの公開ベンチマークで最先端の手法を上回る性能を示します。VLOGGERのパフォーマンスを多様性指標に関して分析し、我々のアーキテクチャの選択とMENTORの使用がスケールにおいて公平で偏りのないモデルのトレーニングに役立つことを示します。最後に、ビデオ編集やパーソナライズの応用を示します。

1.説明

私たちは、テキストや音声に基づいて話したり動いたりする人物のビデオを、たった1枚のその人物の画像から自動生成する手法「VLOGGER」を紹介します。コンテンツ制作、エンターテイメント、ゲームなどの業界では人間の合成に対する高い需要がありますが、現実的な人間のビデオの作成は依然として複雑で、多くのアーティファクトが発生します。現実的な結果を得るためには多大な手動介入が必要です。しかし、完全な自動化は創造的なプロセスを容易にするだけでなく、例えば強化されたオンラインコミュニケーション、教育、パーソナライズされたバーチャルアシスタントなど、まったく新しいユースケースを可能にします。これは、最近のチャットエージェントの成功を考えると特に関連性があります。研究によると、そのようなソリューションは自然に感じられず、共感を生むには不十分とされています。また、いくつかの研究者は、擬人化や行動の現実感（視線、表情、全身の動きなど）が社会的存在感を創り出し、ユーザーからの共感的な反応を引き出すために重要であると主張しています。このような機能は、カスタマーサービス、遠隔医療、教育、人間とロボットの相互作用などの分野でエージェントの広範な採用につながるでしょう。本研究では、まさに自動化と行動の現実感を目指しています。VLOGGERは、音声とアニメーションによるビジュアル表現を備え、複雑な表情や増大する身体の動きを特徴とする、人体化された対話型エージェントのマルチモーダルインターフェースです。VLOGGERは、プレゼンテーション、教育、ナレーション、低帯域幅のオンラインコミュニケーション、テキストのみのHCIのインターフェースとして単独で使用することができます。本論文では、さらにビデオ編集タスクにおけるその可能性を示します。

マルチモーダルでフォトリアリスティックな人間の合成は、データの取得、自然な表情の表現、音声との同期、遮蔽、全身の動きの表現（特に1枚の入力画像から）などの課題があるため、複雑です。多くの試みは、ドライビングビデオの口の領域を編集することで、リップシンクに専念しています。最近では、顔の動きを予測して音声から話す頭のビデオを生成するために、顔の再現の大規模な進歩に依存しています。時間的一貫性は通常、顔のキーポイントから滑らかなガイド動作に依存するフレームごとの画像生成ネットワークで達成されますが、これはぼやけを引き起こす可能性があり、顔から離れた領域の時間的一貫性を確保するものではありません。その結果、ほとんどの手法は、体の大部分が見える場合に頭部を検出してクロップする必要があります。本論文では、コミュニケーションは「単なる」音声と唇や顔の動きの組み合わせ以上のものであると主張します。人間はジェスチャー、視線、まばたき、姿勢を通じて体を使ってコミュニケーションします。最近、MODAは顔と体のアニメーションを探索し始めましたが、限られたシナリオであり、新しいアイデンティティへの一般化はされていません。それとは対照的に、私たちは、頭部と手のジェスチャーを含む動きのリアリズムと多様性に焦点を当てた、一般的で人物に依存しない合成ソリューションを目指しています。私たちの目的は、アイデンティティやポーズを制御できない動的なビデオを生成できる最近のビデオ合成の取り組みと、制御可能な画像生成手法とのギャップを埋めることです。

その目標に向けて、まず入力音声信号に従って体の動きと表情を予測する生成拡散ベースのネットワークを提案します。この確率的アプローチは、音声とポーズ、視線、表情との間の微妙な（多対一の）マッピングをモデル化するために必要です。次に、最近の画像拡散モデルに基づいた新しいアーキテクチャを提案し、これをアブレーションします。このアーキテクチャは時間的および空間的な領域での制御を提供します。事前トレーニング中に取得された生成的な人間の事前知識を利用することで、画像拡散モデルが一貫した人間の画像（例えば、目）を生成するのに苦労する問題をどのように改善できるかを示します。

VLOGGERは、基本モデルと高解像度のビデオを取得するための超解像度拡散モデルから構成されています。顔の表情だけでなく、体や手を含む全身を表現する2Dコントロールに基づいてビデオ生成プロセスを条件付けます。任意の長さのビデオを生成するために、時間的アウトペインティングアプローチを採用し、前のフレームに基づいて新しいビデオクリップを条件付けします。最後に、VLOGGERの柔軟性により、入力ビデオの特定の部分（例えば、唇や顔の領域）を編集することができます。

　堅牢性と一般化のために、私たちはスキントーン、体のポーズ、視点、発話、体の可視性など、これまでのデータよりもはるかに多様性のある大規模なデータセットを収集しました。これまでの試みとは異なり、このデータセットには動的な手のジェスチャーを含むビデオも含まれており、人間のコミュニケーションの複雑さを学ぶ上で重要です。VLOGGERは、さまざまな多様性指標においてこれまでの研究を上回り、以前のHDTF [97] および TalkingHead-1KH [79] データセットにおいて、最先端の画像品質と多様性の結果を得ています。さらに、私たちの手法は、頭部と上半身の動きの高解像度ビデオを生成し、非常に多様な表情とジェスチャーを特徴とすることにより、ベースラインよりも広範なシナリオを考慮しています。最後に、実験セクションでは、VLOGGERの柔軟性と異なるシナリオに適応する能力を示すために、下流アプリケーションを探求します。例えば、VLOGGERはインペインティングによるビデオ編集に使用できます。

　要約すると、主な貢献は次のとおりです：

① VLOGGERは、音声入力から話し、動く人間を生成する初のアプローチです。

② トレーニングとテストのために、既存のデータセットより1桁大きい多様なキュレートされたデータセット「MENTOR」を活用しています。

③ 提案手法を制御されたビデオ生成で検証し、既存の拡散ベースのソリューションと比較して、提案された2Dボディコントロールの利点を示す大規模なアブレーションスタディを行いました。

④ VLOGGERは、3つの公開ベンチマークでの大規模な定量比較で、既存の最先端技術を上回っています。

⑤VLOGGERが低バイアスを示し、異なる人間属性の認知においてベースラインを上回る多様性分析を行いました。

⑥VLOGGERのビデオ編集への応用と、その確率性の分析を行いました。

表1 関連研究と比較したVLOGGERの主な特性は次のとおりです。顔の再現（Face Reenactment）【9,19,29,49,69,87,96】は、一般的に音声やテキストを駆動として考慮していません。音声からモーションへの研究【14,18,57,65,68,84,90】は、音声を3D顔の動きにエンコードすることで構成要素を共有していますが、フォトリアリズムに欠けています。リップシンク【21,54】は異なる被写体の入力ビデオを考慮しますが、口の動きだけをモデル化しています。一般化能力を考えると、SadTalker【95】とStyletalk【42】が最も近いですが、顔のクロップされた画像を必要とし、体のコントロールがなく、ビデオ編集もできません。

グラフ

低い精度で自動的に生成された説明

2.関連研究

音声駆動の話す顔の生成には、多くの研究があり、駆動入力、中間表現、出力形式によって分類できます。表1に我々の研究と比較した概要を示します。音声セグメントに基づいて3Dモーフ可能な顔【14,18,57,65,68,84】や全身【90】のモデルをアニメーション化する研究があります。これらの取り組みは、さまざまな統計的頭部または身体モデルの時間的一貫性のあるポーズおよび形状パラメータ の形式で、多様な3Dトーキングヘッドを生成できます。我々も生成された動きをガイドするために同様のネットワークを検討しますが、本論文ではターゲットの人物の画像と一致する表情や頭部または身体の動きの多様性を持つフォトリアリスティックな話す人間を生成することを目指します。時間的一貫性、被写体の多様性、髪型、視線、出力ビデオの詳細などの課題を考慮します。

　画像領域では、口の編集に焦点を当てた初期の研究が行われています（[11, 13, 31, 54, 73, 97]）。これらの研究は、入力音声に同期したリップモーションのみを予測することに特化しています。その後の研究では、頭部の動き、視線、まばたきなどの拡張された特徴を加えることが行われています（[32, 41, 56, 67, 98, 102]）。これらは、中間的な2D、3Dのランドマークやフローベースの表現を使用しています。フォトリアリズムのレベルを向上させるために、多くの研究では損失の一部として識別器を広範に使用してきました（[8, 9, 17, 55, 80, 92]）。また、最近のいくつかの手法では拡散モデルの使用を提案しています（[65, 66, 93]）。しかし、GANの潜在空間や一般的な拡散モデルで操作する際に、体の動き、頭部の動き、まばたき、視線、および顔の表情の適切な分離を確保することは難しいです。私たちの手法では、カスタムの知覚損失、視線保存損失、身元保存損失、リップシンク損失を必要としません。データの欠如とビデオの一貫性の難しさのため、これまでに体の動きやジェスチャーは考慮されていませんでした。私たちは大規模なデータセットをキュレーションし、この問題に向けた完全なパイプラインを提案しています。VLOGGERは、さまざまな表情、頭部および体の動き、視線、まばたき、正確な口の動きを持つ、一貫した顔と上半身の動きを生成できます。

**顔の再現（Face Reenactment）**は、ソースビデオの動きをターゲットの人物に転送することを目的としたビデオベースの話す顔の生成で、過去に広く探究されてきました（[9, 23, 28, 29, 49, 69, 81, 87, 96, 99, 101]）。ほとんどの手法は、疎なまたは密なランドマーク、セマンティックマスク、3D密な表現、または変形した特徴などの中間表現に依存しています。3D領域では、いくつかの研究がNeRF [4, 44]ベースの解決策を活用しています（[22, 39, 88, 89]）。しかし、これにはターゲットの人物が話す多数のフレームが必要であり、それらを再トレーニングしてアニメーション化する必要があります。このタスクは私たちの目標と密接に関連しており、以前のいくつかの研究ではこれらの中間表現を音声を入力とする際に適応しています。しかし、私たちの場合、顔のみのビデオから進み、体や髪の動きなど、より多様な入力サンプルを考慮することを目指しています。

**ビデオ生成**

私たちの研究に関連するテーマとして、ビデオ生成があります。このタスクはコミュニティで広く探究されており、ここでは最も関連性の高い方向性に焦点を当てます。テキストから画像への拡散モデルの成功に伴い（[16]）、多くの研究がビデオ領域への拡張を探究しています（[2, 6, 24, 26, 35, 36, 64, 72, 83]）。しかし、これらのほとんどは数秒または解像度に制限があります。さらに、これまでの研究の多くは利用可能なデータの量にもかかわらず、人間を明示的に扱っていません。

私たちの場合、現在の最先端の画像拡散モデルを時間領域に拡張し、時空間制御を追加することで、可変長のビデオを生成するための反復アウトペインティング手法を提案します。同時進行の研究は、より一般的なシナリオのために類似したネットワークアーキテクチャを探究していますが（[2, 64]）、私たちの目標は、話す人間をアニメーション化することであり、各フレームを1）ポーズを取った3Dボディモデルの密なレンダリングと2）変形された参照画像でパラメータ化します。これらの制御により、生成プロセスは実験セクションで説明するように、より安定します。

3.手法

私たちの目標は、ターゲットとなる人物が話しているフォトリアリスティックなビデオ V を生成することであり、現実的な頭部の動きやジェスチャーを含む可変長のビデオを合成することです。私たちのフレームワーク「VLOGGER」は図2に示されています。VLOGGERは、音声からビデオへの一対多のマッピングを表現するための確率的拡散モデルに基づいた二段階のパイプラインです。

最初のネットワークは、サンプルレート の音声波形 を入力として受け取り、ターゲットビデオの長さ Nにわたる視線、表情、および3Dポーズを担当する中間的な体の動き制御 Cを生成します。第二のネットワークは、大規模な画像拡散モデルを拡張した時間的な画像から画像への翻訳モデルであり、予測された体の動きを制御して対応するフレームを生成します。このプロセスを特定のアイデンティティに条件付けするために、ネットワークは人物の参照画像も使用します。VLOGGERは、新たに導入したMENTORデータセット（§3.3）でトレーニングされます。次に、両方のネットワークについて説明します。

グラフィカル ユーザー インターフェイス

中程度の精度で自動的に生成された説明

図2 高レベルの概要

VLOGGERは、統計的な3Dボディモデルを使用してビデオ生成プロセスを条件付けします。入力画像（左）を与えられると、予測された形状パラメータはターゲットアイデンティティの幾何学的特性をエンコードします。まず、ネットワーク M は入力音声のメルスペクトログラム a を受け取り、Nフレームの3D顔の表情

と体のポーズのシーケンスを生成します。動く3Dボディの密な表現をレンダリングして、ビデオ生成段階での2D制御 として機能させます（制御の例はSup. Mat.にあります）。これらと対象の参照画像を合わせて、時間的な拡散モデルと超解像モジュールに入力され、ターゲットアイデンティティのフォトリアリスティックな再現のシーケンスを生成するようにトレーニングされます。実装の詳細はSup. Mat.にあります。

3.1 音声駆動のモーション生成

**アーキテクチャ**

私たちのパイプラインの最初のネットワーク M は、入力された音声に基づいて駆動モーションを予測するように設計されています。テキスト入力も考慮し、テキストを音声に変換するモデルを通じて入力を波形に変換し（[70]）、生成された音声を標準的なメルスペクトログラムとして表現します。Mはトランスフォーマーアーキテクチャ（[71]）に基づいており、時間次元に沿った4つのマルチヘッドattention層を持ちます。フレーム数と拡散ステップの位置エンコーディング、入力音声と拡散ステップの埋め込みMLPを含みます。各フレームでモデルが前のフレームのみに注目するように因果マスクを使用します。モデルは、可変長のビデオを使用してトレーニングされ、非常に長いシーケンスを生成できるようにしています。例えば、TalkingHead-1KHデータセット（[79]）における例（§4参照）のようにです。

私たちは、統計的かつ表現力豊かな3Dボディモデルの推定パラメータ（[33, 51, 63, 85]）に依存して、生成されたビデオの中間制御表現を生成します。これらのモデルは顔の表情と体の動きを考慮しており、より表現力豊かで動的なジェスチャーを持つ人間の合成を可能にします。モーション生成ネットワークに、フレーム i の入力音声 に基づいて顔と体のパラメータ を予測させます。特に、モデルは表情と体のポーズの残差を生成します。変位、つまりを予測することで、モデルはターゲットの参照ポーズ を持つ入力画像を使用し、人物をフレーム の間

と相対的にアニメーション化します。幾何学的領域における人物のアイデンティティは、体の形状コードによってモデル化されます。

トレーニングとテストの両方で、入力画像に対してパラメトリックボディモデルを適用して得られた推定3D形状パラメータを使用します。CNNベースのアーキテクチャを使用して2D/3D予測を活用するために、予測された表情とポーズパラメータを使用してモデルをポーズさせ、ポーズされた体のテンプレート頂点位置を密な表現としてラスタライズし、密なマスクを得ます。また、体の意味領域

を異なる意味クラスに対してラスタライズします。

さらに、従来の顔の再現作業では、ワープ画像に依存することがよくありますが（[19, 76, 95, 99]）、これらは人間のアニメーションのための拡散ベースのアーキテクチャでは見過ごされていました（[10, 30, 78]）。私たちはこれらの2つの表現のギャップを埋め、生成プロセスをガイドするためにワープ画像を使用することを提案します。これにより、ネットワークのタスクが容易になり、被写体のアイデンティティの保持に役立つことがわかりました（表3を参照）。参照画像で見える各体の頂点にピクセルカラーを割り当て、新しい各フレームで体をレンダリングし、部分的なワープ

を得ます。すべてのレンダリングについて、ラスタライズプロセスはフルパースペクティブカメラを想定し、視野の対角線はトレーニングビデオまたは参照画像から推測されます。図2を参照してください。次のセクションと追加資料（Sup. Mat.）で時間的画像拡散モデルを説明します。また、実験セクションでは、密な表現とワープ画像の使用についても検討します。

**損失関数**

このモデルは、diffusionフレームワークに従い、条件付き音声入力 a とともに、真のサンプルに対してガウスノイズを段階的に加えます。目標は、ノイズの入った入力 から追加されたノイズを予測するデノイジングネットワーク

をトレーニングすることで、真の頭と体の動作分布をモデル化することです。ここで、は任意のdiffusionステップです。私たちの場合、真の分布を直接予測することでより良い性能を得ました。

(1)省略

また、連続するフレーム間の予測差をペナルティ化するために追加の時間的損失を含めます。これを任意のフレームに対して適用し、両方の損失の線形結合を使用して完全なモデルをトレーニングします。

実際には、表情と体のポーズに対して異なる時間的損失の重みを使用し、頭と手の動きがより滑らかになるようにしながら、表情に対してはより大きな動的変化を許容します。

3.2 フォトリアリスティックな話す動く人間の生成

**アーキテクチャ**

次の目標は、人物の入力画像 を、前述の体と顔の動きに従ってアニメーション化することです。これらの動きはセマンティック、スパース、およびデンスマスク C で表されます。これらの画像ベースのコントロールに基づいて、最新のdiffusionモデル [60] の時間認識拡張を提案します。ControlNet [94] にインスパイアされ、最初に訓練されたモデルをフリーズし、そのencoding層のゼロ初期化されたトレーニング可能なコピーを作成します。このコピーは入力時間コントロールC を受け取ります。各ダウンサンプリングブロックの最初の層の後、そして2番目のGroupNormアクティベーションの前に、時間領域で1d畳み込み層をインターリーブします。ネットワークは連続するNフレームとコントロールを取り、入力コントロールに従ってアニメーション化された参照人物の短いクリップを生成するように訓練されます。

**学習**

私たちの方法は、ユニークな人間被験者のフルレングスビデオで構成されるMENTORデータセットで訓練されます。トレーニング中、ネットワークは連続するフレームのシーケンスとその人物の任意の参照画像 を受け取ります。理論的には、任意のビデオフレームを参照として割り当てることができます。実際には、ターゲットクリップから時間的に離れた参照をサンプリングします。近い例はトレーニングをトリビアル化し、一般化の可能性を提供しないためです。ネットワークはまず単一フレーム上で新しいコントロールレイヤー [94] を学習し、その後時間コンポーネントを追加してビデオでトレーニングします。この2段階のトレーニングにより、最初の段階で大きなバッチサイズを使用し、頭の再現タスクをより早く学習することができます。画像モデルは学習率5e-5で、バッチサイズ128で両段階で400kステップトレーニングします。このトレーニングスケジュールの効果を表3で実験し、トレーニング手順の詳細は追加資料に提供されています。

**損失関数**

前述のセクションや式（1）で説明されているように、私たちは拡散プロセスに従ってノイズ ϵI を真の画像 I に追加します。私たちの作業は内部データソースで訓練された Imagen [60] のバージョンに基づいています。このノイズ ϵI を予測します。

前述の手法は解像度に依存しないが、私たちは基本のビデオを128×128解像度で生成し、2つの超解像バリアントで256×256または512×512のより高品質なビデオを生成するためにカスケード拡散アプローチを使用します。生成された画像は {Gi}1≤i≤N と表記されます。高解像度の例は図1と図4に示されています。

**推論中の時間的アウトペイント**

提案された時間的拡散モデルは、固定数のフレームNのみを生成するように訓練されているため、可変長のビデオにどのように拡張するかは明らかではありません。以前の拡散ベースのビデオ生成方法は、通常短いクリップに制限されています[27, 35, 83]または滑らかに生成された中間トークン表現に依存していますが、ピクセルドメインでの滑らかな変化が保証されていません[72]。ここでは、時間的アウトペイントのアイデアを探ります。まず、Nフレームを生成し、その後、前のN-N'フレームに基づいて反復的にN' < Nフレームをアウトペイントします。連続する2つのクリップ間のオーバーラップ量、すなわちN - N'は、品質と実行時間のトレードオフとして選択されます。各ビデオクリップの生成にはDDPMを使用し、このようなアプローチが数千フレームにスケーリングできることを示します。詳細については、Tab. 2の検証を参照してください。

MENTORデータセットは、内部ビデオの大規模なリポジトリから収集されたもので、ほぼ正面向きの、カメラに対して上半身を中心に話す単一のスピーカーが含まれています。ビデオは英語で主にコミュニケーションを行い、1秒あたり24フレーム（10秒のクリップ）で構成されており、オーディオは16 kHzで収録されています。

私たちの目標は、全身を含むコミュニケーションする人間のモデリングであり、連続するフレーム間の投影エラーや時間的な差を最小化することで、3Dボディジョイントと手を推定し、統計的な関節付き3Dボディモデルを適合させています。背景が意味を持って変化するビデオ、顔や体が部分的にしか検出されていないか、その推定が不安定なビデオ、手が完全に検出されていない（例えば物をつかんだり操作したりする場合）、またはオーディオの品質が低いビデオはフィルタリングしています。このプロセスにより、訓練セットは800万秒（2.2千時間）以上と80万の個体数、テストセットは約120時間と約4千の個体数になりました。これにより、個体数や長さの観点で今日までで最大のデータセットとなり、より高解像度での使用が可能です。さらに、MENTORデータセットには多様な被写体（肌の色、年齢など）、視点、および体の可視性が含まれています。詳細な統計データや既存のデータセットとの比較は、補足資料で提供されています。私たちは、このカリキュレートされたビデオID、顔の適合、推定されたボディポーズを広範な研究コミュニティに公開することを目指しています。

4. 実験

データとトレーニングについて。我々はセクション3.3で述べた通り、MENTORデータセットを使用してVLOGGERをトレーニングします。ベースの解像度は128 × 128で、カスケード解像度は256 × 256および512 × 512です。評価はHDTF [97]、TalkingHead-1KH [79]、およびMENTORのテストセットで行われます。また、MENTORデータセットでのさまざまなシナリオでの我々の手法のパフォーマンスを分析し、年齢、知覚される性別、肌の色などの多様性メトリクスに基づいて、ベースラインとの比較を報告します。

**Baselines**

我々はいくつかの最先端の手法と比較します。具体的には、[42, 76, 77, 95, 104]です。なお、我々の手法と異なり、すべてのベースラインは顔領域をクロップする必要があります。これは彼らが顔を検出してアニメーション化できるが、頭部のみです。

**メトリクス**

生成されたビデオの品質、リップシンク、時間的一貫性、および生成されたビデオのアイデンティティの保存を評価するために、複数のメトリクスの組み合わせを使用します。画像の品質については、FIDスコア[25]が生成された画像の分布と正解の画像の分布との間の距離を測定します。また、Cumulative Probability of Blur Detection (CPBD) [47, 48]やNatural Image Quality Evaluator (NIQE) [45]は生成された画像の品質を検証します。口の同期の品質を測定するために、口の頂点位置の差（LME）を報告します。さらに、LSE-Dスコア[12]も報告します。生成されたビデオの時間的な滑らかさを測定するために、jitter（またはjerk）エラーを報告します。生成されたビデオから予測された表情パラメータの標準偏差も提供し、発話からビデオへの変換が必ずしも一対一のマッピングでないことを考慮し、実際のビデオの分布を生成することが重要であることを評価します。また、VLOGGERはジェスチャーを考慮する最初のモデルであり、これを質的に評価します。

4.1 デザインの削減研究

Tables 2および3では、主要な設計選択肢を広範に削減しています。Table 2では、最終手法（最後の行）の代表的なメトリクスをまとめ、各行が1つの特徴を変更した場合の効果（例えば、モーション予測器の訓練時に一時的な損失を使用しない場合）を示しています。Table 3では、ビデオ生成に使用される2次元の制御の重要性を検証しています。以下で結果について議論します。

モーション生成について。Table 2の上部では、一時的な損失を使用しない場合やΔを予測しない場合の時間的一貫性の低下を示しています（Sec 3.1を参照）。ボディモーションに対する残差の予測が安定性と滑らかさを向上させ、全体的な画質を高めることが示されています。また、補足資料で議論されたクラシファイアフリーのガイダンスが、LMEやFID [25]に対する肯定的な影響を示しています。

ビデオ生成について。Table 2の下部は、時間的ビデオ生成モデルの設計選択肢を削減しています。まず、提案されたアウトペイント手法の有効性を検証し、可変長ビデオ生成をサポートするだけでなく、滑らかさと低いジッターを確保します。最終モデルでは生成フレームと与えられたフレームの50%の重複があり、より大きな値でもプラトーになりますが、より小さい重複（25%）やアウトペイントなしと比較して顕著な改善を達成します。また、ボディポーズ制御を使用することでモデルの性能も向上します。

**2次元制御がビデオ生成に与える影響**

最後に、Table 3で、ビデオ生成プロセスをガイドするために使用される異なる表現の重要性を削減します。具体的には、テストセットのサンプルをその正解モーションで再演し、画像再構成メトリクスを報告します。私たちは、2次元ランドマーク、密な表現、および提案された最終的な制御方法を探索します。この方法は、密なボディ表現と参照入力画像から変形した部分的なビューを組み合わせています。後者はネットワークのタスクを大幅に緩和し、最良の結果をもたらします。さらに、セクション3で説明されているトレーニングスケジュール（および補足資料）により、最初に単一の画像でトレーニングし、後でビデオの時間的層をファインチューニングすることで、パフォーマンスを追加で向上させます。

表2 VLOGGERの主要な設計選択肢の削減研究結果をMENTORデータセットで評価し、画像品質をFID [25]スコア、表現力とリップシンクの品質をランドマークエラー（LME）、および顔頂点のジッターに基づく時間的一貫性で検証するための代表的なメトリクスを報告します。最初の部分では、一時的損失とクラシファイアフリーのガイダンスが、画像品質とLMEで最高のパフォーマンスをもたらすことが示されています（比較のための最後の行に完全なモデルがあります）。二番目の部分では、時間的拡散モデルの設計選択肢に対する改善がまとめられています。最終的なパイプラインはボディ制御を採用し、提案された時間的アウトペイント（フルモデルでの50%の重複）が最も良い時間的一貫性をもたらします。より多くの重複ではモデルがプラトーに達することが観察されました。

グラフィカル ユーザー インターフェイス

中程度の精度で自動的に生成された説明

表3 MENTORデータセットにおけるビデオ生成における2次元制御の削減研究です。この実験では、同時期の研究で考慮されている異なる2次元制御（例えば、2次元スケルトンの駆動 [30, 78]、密なボディ表現 [86]、または私たちが提案する制御方法である密な表現と変形画像）を削減します。最初の画像を取り、元の動きに従って残りのビデオをアニメーション化し、平均画像類似度メトリクスと部位ごとの平均を報告します。すべてのバリエーションは同じデータで訓練されています。

テーブル

自動的に生成された説明

4.2 定量的結果

**トーキングヘッド生成**

Table 4では、VLOGGERのパフォーマンスを従来の最先端手法と比較し、音声によるビデオ生成のタスクでの結果をまとめています。我々は、HDTFデータセット [97]とTalkingHead-1KHデータセット [79]での結果を報告しています。HDTFデータセットは大規模ですが、アイデンティティ数（300）が少なく、視点の多様性にはやや限界があります。トーキングヘッド生成は複数の望ましい特性を持つ難しいタスクであり、さまざまなメトリクスで評価されます。顕著なのは、画質、多様性、アイデンティティの保存の間にトレードオフがあることです。VLOGGERは実際のビデオに存在する表現の多様性に近づきながら、最高の画像品質とアイデンティティの保存を達成しています。また、StyleTalk [42]に次いで二番目に低い動きのジッターを示しており、非常に少ない顔の動きを導入しています（Fig. 4を参照）。時間的一貫性は、私たちの時間層とアウトペイント手法の効果を示しており、最先端の拡散モデルの高品質な画像生成能力を活用しています。すべての方法が類似したリップシンクスコアを達成し、両データセットで評価されたすべてのメトリクスで結果が一貫しています。

さらに、我々は各被験者ごとに生成されるサンプルの数（3、5、または8）を変化させて、それぞれの最良のビデオを選択することで、我々の手法の性能を評価しました。このアプローチはサンプル数が増加するにつれて性能が著しく向上し、VLOGGERが各被験者に対して多様な出力を生成する能力を示しています。

また、既存のベースラインは通常顔の画像のみに焦点を当てていますが、私たちの目標は手を含む可視部位のモデリングに及んでいます。ベースラインは身体の動きやジェスチャーを考慮していない一方で、私たちはこの点での設計選択の削減実験をTables 2および3で詳細に評価しました。

表4 HDTFおよびTalkingHead-1KHデータセットでの定量的評価。我々のモデルが複数のメトリクスでリアルな話す顔を生成する能力を測定します。VLOGGERは、いくつかのメトリクスで最高の視覚的品質と最高のアイデンティティの保存を達成し、表現の多様性と時間的一貫性もグラウンドトゥルースのビデオに近い結果を示しています。リップシンクの品質に関しては、すべての手法が比較可能なスコアを得ています。VLOGGERによって生成される多様性を示すために、3、5、または8のビデオを生成した場合の性能向上も報告しています（FIDを除き、これは画像分布内の類似性を測定します）。すべてのメトリクスにおける結果は両データセットで一貫しています。

テーブル

自動的に生成された説明

図3では、我々は公平性と汎化能力（トレーニングセットの規模と多様性による部分も含めて）を示し、他の方法との比較を通じて複数の知覚される属性に対して走査します。過去の研究では、異なる属性（例：肌の色、若い vs 年老いたなど）に対して明確なパフォーマンスの低下が見られ、また、肩や手が見えるビデオには汎化できないことがあります。これに対して、VLOGGERはすべての評価された軸でかなり低いバイアスを示しています。MENTORのリリースがコミュニティに重要な公平性の問題に取り組み、最先端技術のさらなる進展を促進することを期待しています。

グラフ

自動的に生成された説明

図3　当社のモデルと最も近い競合モデルを、肌の色、性別、年齢などの異なる知覚属性について、MENTORデータセットのテストセットで比較します。当社のモデルは、大規模な事前学習済み拡散モデルと当社の提案する大規模データセットからの事前情報を活用しています。そのため、他の方法とは異なり、すべてのカテゴリで一貫した性能を発揮し、ほとんどバイアスがないことを示しています。また、当社のモデルは、画像内の人物を多様な視点でアニメーション化する能力を示しており、顔の周りに狭いバウンディングボックスをクロップする代わりに、広範な視野で作業を行うことを示しています。

ポーズをとる男性グループの写真

中程度の精度で自動的に生成された説明

Fig. 4. 入力画像（左）と生成されたフレームを示す定性的比較。ベースラインは通常、シーケンス全体で表情を維持し、頭部をクロップする必要があります[42, 77, 95]。これに対して、VLOGGERは顔を考慮した場合（第三行）、さらには上半身が見える場合（第五行）でも可視領域で変化を生成します。この図はアニメーション化された顔を示していますが、ジェスチャーを伴う例はFig. 1および補足資料に示されています。

Cgで描かれた女性の写真のコラージュ

低い精度で自動的に生成された説明

Fig. 5. モデルの多様性を示す図。VLOGGERは確率的であり、同じ被験者に対してさまざまなビデオを生成することができます。被験者の画像と入力音声が与えられた場合、列2から列5はそれぞれ1秒から4秒後のピクセル色の偏差を示します。これらは24本の生成されたビデオから得られています。1秒後の時点で（第2列）、モデルは手のポーズと顔の表情に大きな多様性を示し、すべてのビデオが良好な視覚的品質を持っています。

4.3 定性的結果

Fig. 4にて、最近の最高性能なベースラインと比較して、画像の実世界での定性的結果を示します。ほとんどの過去の作品は生成能力に制限があり、参照画像で遮蔽された部分を生成することが難しいです（例えば、歯が口の内部を隠している場合、生成されたビデオでもそれが継続されます）。これに対して、当社のモデルはより多様な表情を生成し、移動する頭部の遮蔽された領域を正しくインペイントします。

**サンプルの多様性**

VLOGGERは確率的であるため、同じ入力音声/テキストに対して複数のモーションやビデオを生成することができます。Fig. 5で示されているように、最初の行では背景がほぼ静止している一方で、顔、髪、視線、および体の動きは時間的に展開するにつれて変化しています。

**ビデオ編集**

同様に、当社の拡散アプローチはビデオ編集の能力を示しています。Fig. 6は、入力ビデオ（上段）に対する編集例を示しており、口を閉じる（第二行）、目を閉じる（第三行）、または被写体の目を開いたままにする（例：まばたきしない、第三行）などを、時間的に整合した方法で行います。この場合、顔の表情を編集した後、地面の画像と異なるようにプロジェクトされるボディ座標に基づいてインペイントマスクを自動的に生成し、この時間的マスクを使用してピクセルを新しいターゲットコントロールに従って再生成します。このプロセスはビデオの長さ、カメラへの距離、または被写体のアイデンティティに依存せずに行われ、これらの結果が創造的なビデオ編集の新しい応用につながることを期待しています。補足資料にあるビデオも参照してください。

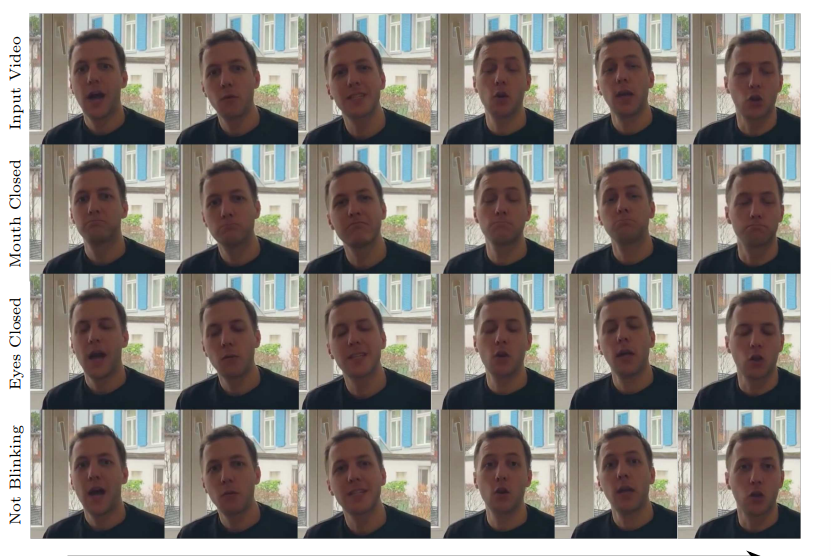


Fig. 6. ビデオ編集の結果を示す図。入力ビデオ（第1行）が与えられた場合、我々は新しい顔の表情を定義して口を変えます（第2行）、目を変えます（第3行）、またはビデオ全体で目を開いたままにします（第4行）。時間的インペイントマスクは自動的に体の変化する部分から定義されます。詳細は補足資料でご覧ください。

画面操作された男性の顔

低い精度で自動的に生成された説明

Fig. 7. モデルの個人化に関する定性的結果。ユーザーの単一のビデオで当社のモデルを微調整することは、幅広い表情にわたってより現実的な合成をサポートします。

最近の拡散モデルにおける個人化は、主題駆動の生成のために広く探求されています（[59]）。当社の場合、VLOGGERは単眼の入力画像のみを使用して合成を行います。それにより合理的な結果を生成できますが、被覆された部分にアクセスできず、その結果得られるビデオはその人物の微細な分析で正確ではないかもしれません。

図7では、主題の単眼ビデオを用いて当社の拡散モデルをさらにデータで微調整することで、VLOGGERが個々の特性をより正確に捉える能力を向上させることを示しています。たとえば、参照画像が目を閉じている場合などに、モデルがその個人の特徴をより良く反映したビデオを生成できるようになります。

結論

VLOGGERを紹介しました。この手法は、単一の入力画像に基づいて音声またはテキストで条件付けられた、顔と身体の両方を含む人間のビデオ合成手法です。VLOGGERは、制御ベースの拡散モデルの時間的拡張として構築されており、3D人間の頭部と身体のポーズ表現を基盤としています。これにより、可変長の高品質なアニメーションを生成します。私たちは、従来のデータセットよりも1桁大きな多様で大規模なデータセットを導入しました。MENTORデータセットでのVLOGGERの性能を検証し、複数の他のリポジトリでもその性能を示しました。これにより、話す顔生成のタスクで従来の最先端技術を上回り、私たちのアプローチが異なる多様性の軸でより堅牢であることを示しました。補足資料では制限と社会的影響について議論しています。

謝辞：Alonso Martinez、Anja Hauth、Sergi Caelles、Hernan Moraldo、Erik Frey、Krishna Somandepalli、Brendan Jouに感謝します。彼らは、MENTORをキュレーションするために大規模で多様なビデオリポジトリを慎重に収集し、分析してくれました。