一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
MVE2025-11(2025-08)

Stable Diffusionを用いた絵伝のアニメーション化に関する検討

荒田　涼太†　道満　恵介†,†† 井手　一郎††　目加田慶人†

† 中京大学 工学部　〒470–0393 愛知県豊田市貝津町床立 101
†† 名古屋大学 大学院情報学研究科　〒464–8601 愛知県名古屋市千種区不老町
E-mail: †arata.r@md.sist.chukyo-u.ac.jp, ††ide@i.nagoya-u.ac.jp, †{kdoman,y-mekada}@sist.chukyo-u.ac.jp

あらまし　絵伝とは，伝記を絵と詞書で連続的に描画したものであり，そこに描かれたストーリを宗教的世界観と共に解説することを絵解きという．我々は，画像生成モデルを用いることで，絵伝の画像に描かれた各場面を自動的にアニメーション化する手法を検討している．提案手法では，Stable Diffusion に基づく画像生成モデルをファインチューニングし，image-to-image やアニメーション生成に対応させる機能を利用することで，絵伝の画風に即した自然で滑らかなアニメーション化を目指す．本稿では，杭全神社蔵「聖徳太子絵伝」の各場面に本手法を適用し，その有用性を評価した結果について報告する．

キーワード　聖徳太子絵伝，アニメーション化，Stable Diffusion，ファインチューニング

A Study on Animating an Illustrated Biography using Stable Diffusion

Ryota ARATA†, Keisuke DOMAN†,††, Ichiro IDE††, and Yoshito MEKADA†

† School of Engineering, Chukyo University
101 Tokodachi, Kaizu-cho, Toyota, Aichi, 470–0393 Japan
†† Graduate School of Informatics, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, Aichi, 464–8601 Japan
E-mail: †arata.r@md.sist.chukyo-u.ac.jp, ††ide@i.nagoya-u.ac.jp, †{kdoman,y-mekada}@sist.chukyo-u.ac.jp

Abstract　An illustrated biography (*Eden*) is a painting in which multiple scenes of a biography are drawn continuously with some descriptions, and *Etoki* is story telling or preachment referring to religious perspective. We are studying a method for automatically animating each scene in an illustrated biography, using an image generation model. The proposed method aims to consistently animate the illustrations while keeping its style by fine-tuning a Stable Diffusion-based image generation model and utilize extensions for text-to-image and animation generation. We applied the method to each scene of the "Illustrated biographies of Prince Shotoku" in the collection of Kumata Shrine, and report its evaluation results for the usefulness of the proposed method.

Key words　Illustrated biography of Prince Shotoku, animating, Stable Diffusion, fine-tuning

1. はじめに

絵伝とは，伝記を絵と詞書で連続的に描画したものであり，そこに描かれたストーリを宗教的世界観と共に解説することを絵解きという．絵伝の例として，大阪市平野区の杭全神社に伝わり，15 世紀に描かれたとされる「聖徳太子絵伝」を図 1 に，絵解きの様子を図 2 にそれぞれ示す．この聖徳太子絵伝では，聖徳太子の入胎・誕生から薨去・その後までが全 10 幅にわたって描かれており，各幅では複数の場面が 1 枚の中で連続的に描かれている．通常，絵解きは，同図 (a) や (b) のように寺院や芸能会館で絵伝を展示し，絵解き者が必要に応じて指し棒を使う等して絵伝中の各場面を指し示しながら解説する形式で行なわれる．近年では，同図 (c) のようにディジタル端末を用いた「ディジタル絵解き」が実施されることもあり，これにより従来の静的な絵解きだけでなく，より動的でより魅力的な絵解きが可能と期待されている．しかし現状では，解説場面の拡大やシーン遷移のエフェクト等の簡単なアニメーションを盛り込むのに留まっており，ディジタル端末の活用は限定的である．これに対して，例えば登場人物や背景の木々等に動きをもたせてアニメーション化できれば，絵伝中の各場面の雰囲気や臨場感を聴衆に視覚的に説明でき，更に魅力的な絵解きを実現できると考えられる．しかし，そのためには，知識や技術をもった専門家が必要であったり，支援ツールを利用できるとしても事前の作り込みが必要であるため，いくつもの場面に対してそのよ

This article is a technical report without peer review, and its polished and/or extended version may be published elsewhere.
Copyright ©2025 by IEICE

第一幅　第二幅　第三幅　第四幅　第五幅　第六幅　第七幅　第八幅　第九幅　第十幅

図 1　杭全神社蔵「聖徳太子絵伝」（全 10 幅）.

(a) 寺院での絵解き　(b) 芸術会館での絵解き　(c) ディジタル絵解き

図 2　絵解きの様子.

うな作業をするためには膨大な手間がかかる.

以上のような背景から，本研究では，絵伝画像と絵解き時に使用される絵解きシナリオテキストから，絵伝中の場面をアニメーション化する手法を提案する．これにより，ディジタル端末を活用したより動的で魅力的な絵解きを，手軽に実現できるようになることが期待される．提案手法では，text-to-image な画像生成モデルである Stable Diffusion およびその周辺機能を利用して，1 枚の静止画像とテキストプロンプトの組から当該場面をアニメーション化することを考える．このとき，事前学習済みモデルを絵伝にファインチューニングすることで，対象とする絵伝の画風により即したアニメーション化を目指す．なお，本研究では，保存状態の良さやスキャンデータの解像度の観点から杭全神社蔵「聖徳太子絵伝」を対象とするが，提案手法は他の絵伝を対象とすることも可能である．以降，関連技術を概説したのち，手法を詳述し，評価実験の方法および結果について報告する.

2. 関 連 研 究

本節では，本研究で利用する画像生成モデルである Stable Diffusion およびその周辺機能を概説する.

2.1 画像生成モデル：Stable Diffusion

Stable Diffusion [1] は，Stability AI を中心に，CompVis，Runway，LAION などの研究機関により共同で研究・開発され，2022 年 8 月に Stable Diffusion v1 として一般公開された画像生成モデルである．text-to-image タスクのための機能 txt2img を基本として，後述の拡張機能を追加することで様々なタスクに対応できる．本モデルは，主に Variational Auto Encoder（VAE）[2]，Text Encoder [3]，Diffusion Model [4] の 3 つの要素から構成される.

VAE は，高次元の画像を低次元の潜在空間に圧縮するためのモジュールであり，エンコーダは画像を潜在表現へ変換し，デコーダは潜在表現を元の画像空間へ再構成する．エンコーダとデコーダの構造には U-Net [5] が用いられている.

Text Encoder は，テキストプロンプトをベクトルに変換し，画像生成の条件情報として用いる役割を果たす．Stable Diffusion では，画像とテキストを共通の意味空間に埋め込むために，事前学習済みの CLIPTextModel を使用している.

一般に拡散（Diffusion）モデルは，画像生成の過程において段階的にノイズを加える Forward process と，そのノイズを除去して画像を復元する Reverse process の 2 つの過程から構成される．学習時には両プロセス間の誤差を最小化することで最適化が行われ，推論時には Reverse process のみを実行して画像を生成する．この際，Text Encoder から得られた潜在ベクトルは，U-Net による雑音除去処理における条件として用いられ，入力テキストに対応した画像が生成される.

2.2 機能：img2img

Stable Diffusion には，テキストの情報のみで画像生成を行なう text-to-image（txt2img）機能に加えて，既存の画像を入力として画像から画像生成を行なう image-to-image（img2img）機能が実装されている．この機能は，入力画像に雑音を加え，テキストプロンプトを条件としてその雑音を拡散的に再構成することで，既存の画像の構図や形状を保ちながら，新たなスタイルや要素を付与する．これにより，既存の画像に対して自然な変化やスタイルの転写を行なうことが可能である.

2.3 拡張機能：AnimateDiff

AnimateDiff [6] は，拡散モデルにおける時間的一貫性（Temporal consistency）を確保しつつ，連続したフレームを生成することで，静止画像を対象とした拡散モデルを動画像生成に拡張する手法である．これにより，アニメーション生成が可能となる.

2.4 拡張機能：LoRA

LoRA（Low-Rank Adaptation）[7] は，大規模なニューラルネットワークを効率的にファインチューニングするための手法であり，元のモデルの重みを固定したまま少数の低ランク行列を追加・学習することで，スタイルやタスクに特化した適応が可能となる技術である.

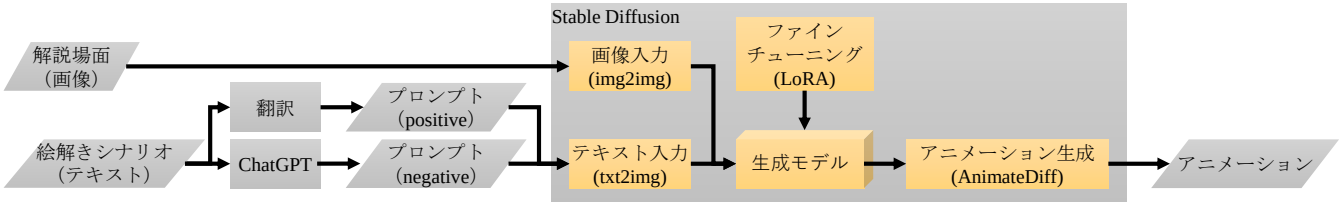


図 3　提案手法におけるアニメーション化手順.

3. 提 案 手 法

提案手法の処理手順を図 3 に示す．絵伝中の 1 つの場面に対応する部分画像と絵解きシナリオテキストの組を入力し，それらに基づいて当該場面をアニメーション化する．画像生成モデルとして Stable Diffusion（txt2img）を利用し，それに前節で紹介した img2img と AnimateDiff，LoRA の各機能を組み合わせることで，当該場面に対応する絵伝の画風に即してアニメーション化する．以降，提案手法における各手順について詳述する．

3.1　入力データ

提案手法では，絵伝中の 1 つの場面に対応する部分画像と絵解きシナリオテキストの組を入力データとする．シナリオテキストに加えて，アニメーション化したい場面の静止画像を指定することで，利用者である絵解き者の意図に沿った画風で登場人物や背景が動くアニメーションを生成する．以降，各データについて詳述する．

a） 絵伝中の場面に対応する部分画像：

絵解き者は，絵伝に描かれた場面を選択し，それに対応する部分領域を抽出する．ここで，テキストと画像の組を入力できるように，text-to-image タスクのための Stable Diffusion の機能の一つである img2img を利用する．なお，部分領域の抽出（場面を囲む矩形のアノテーション）作業は，絵伝の 1 つの幅について 1 回のみ必要な作業である．これまでのディジタル絵解きにおける準備作業に含まれている工程でもあるため，本手法により追加のコストが生じるものではない．

b） 絵解きシナリオテキスト：

絵解き者は，各解説場面について，登場人物やその動作，背景状況などを説明したテキストを絵解きシナリオとして事前に設定する．本手法ではこのシナリオテキストを基に，後段の生成モデルに対するプロンプト（ポジティブプロンプト）を生成する．ただし，通常，絵解きシナリオは日本語で記述されている一方で，本手法で利用する生成モデルは英語テキストを想定したものである．そこで，絵解き者が作成した日本語のシナリオテキストを英訳したものをポジティブプロンプトとして与える．例えば，高精度な意味変換が可能な機械翻訳サービスとして DeepL(注1)があり，これを利用することで原文の語調や意図を保持した翻訳を期待する．なお，プロンプトとして，否定の条件を指示するネガティブプロンプトも存在する．本手法では，ChatGPT(注2)を用いて，各場面に対応するシナリオテキストを入力し，Stable Diffusion 用のネガティブプロンプトを自動生成する．ここで，入力文は以下の形式で与える：

> 「Stable Diffusion に入力するプロンプトを下記条件に基づいて生成してください．◯◯のシーン．ネガティブプロンプトを生成してください．75 単語以内で生成してください.」

ここで「◯◯」には，各場面のタイトルおよびシナリオテキストを入力してから自動生成し，それをネガティブプロンプトとして指定する．

3.2　出力：絵伝アニメーション

Stable Diffusion に基づく画像生成モデルを用いて，絵伝中の 1 つの場面に対応する部分画像と絵解きシナリオテキストの組から，その場面をアニメーション化する．本手法では，Stable Diffusion の拡張機能 AnimateDiff を利用することで，出力を静止画像ではなく時系列的に連続した画像列とし，滑らかなアニメーション化を目指す．

また，前述のとおり，テキストプロンプトに加えて，img2img を利用することで画像を生成時の条件とすることができるが，出力結果を絵伝の画風に近づけるために，生成モデルをファインチューニングする．本手法では，少量の学習データでも効果的にモデルの特化を実現することを目的として，Stable Diffusion の拡張機能である LoRA を利用する．これにより，絵伝特有の色使いや筆致，登場人物の特徴を保持したアニメーション化を目指す．なお，本研究では，アニメーション化したい場面を含む絵伝を用いてファインチューニングする．一般に検出・認識タスクにおいては，事前学習またはファインチューニングのために使用した学習データに含まれない未知のデータでテストし，モデルの汎化性能を評価する．一方，本研究で対象とするアニメーション化のタスクは，未知の絵伝に対応することを目指すものではなく，手元の絵伝に特化したアニメーション化モデルを構築することを目指すものである．そのため，アニメーション化したい場面を含む絵伝に特化してチューニングすることが望ましい．

4. 実　　験

実験を通じて，Stable Diffusion およびその拡張機能を組み合わせた提案手法による絵伝アニメーション化の有用性を評価する．以降，実験の条件，方法，結果を述べ，考察する．

4.1　条　　件

本実験では，Stable Diffusion およびその周辺機能を以下のような条件で使用した．なお，乱数のシード値は 0 に固定した．

(注1)：DeepL: https://www.deepl.com/ja/translator/ ［2025/8/3 閲覧］
(注2)：ChatGPT: https://openai.com/ ［2025/8/3 閲覧］

a） Stable Diffusion：

アニメーション化のためのモデルとして，Stability AI により提供されている．Stable Diffusion v1.5(注3)を使用した．

b） 絵伝中の一つの場面に対応する部分画像：

アニメーション化する絵伝中の各場面の範囲は絵解き者の専門家が人手で設定し，512×512 画素に拡縮した．

c） テキストプロンプト：

Stable Diffusion に与えるプロンプトとしては，絵解き者の専門家が実際の絵解きのために作成した日本語の絵解きシナリオテキストを DeepL で英訳したものを使用した．ネガティブプロンプトには，前節で述べたとおり，ChatGPT によって自動生成したものを用いた．

d） LoRA：

LoRA によるスタイル特化モデルの構築にあたり，学習データとして杭全神社蔵「聖徳太子絵伝」の第一幅から第十幅を使用した．各幅の絵伝画像は 9,902×24,002 画素であり，これを縦 12，横 5 の格子状に分割することで 1 幅あたり 60 枚，計 600 枚の部分画像を得た．これらの部分画像を，512×512 画素にそれぞれ拡縮して学習に用いた．LoRA によるファインチューニングでは，学習率は 1.0×10^{-4}，総ステップ数は 10,000 回，繰り返し回数は 20 回，その他のハイパパラメータは拡張機能 LoRA の標準設定に従った．

e） AnimateDiff：

Stable Diffusion の拡張機能 AnimateDiff を使用する際には，アニメーション長を全 16 フレーム，フレームレートを 8 fps に設定し，約 2 秒間の動画像を生成した．拡散過程における Sampling steps は 50 とした．

4.2 方　　法：

杭全神社蔵「聖徳太子絵伝」の第一幅から第五幅を対象に，そこに描かれている場面のうち 30 場面を上述の条件でアニメーション化した．比較評価のために，表 1 に示すように，下記 4 種類の方法で Stable Diffusion を利用してアニメーション化した結果について，解説場面を描写したものとして適切かどうかを選好実験により比較評価した．

- 比較 1：Stable Diffusion にテキストプロンプトのみ入力（SD）
- 比較 2：LoRA でファインチューニングした Stable Diffusion にテキストプロンプトのみを入力（SD ＋ LoRA）
- 比較 3：Stable Diffusion にテキストプロンプトと解説場面の部分画像を入力（SD ＋ img）
- 提案：LoRA でファインチューニングした Stable Diffusion にテキストプロンプトと解説場面の部分画像を入力（SD ＋ img ＋ LoRA)

具体的には，場面毎に，各方法による生成結果を無作為に並べて評価者に提示し，「場面のタイトル，絵解きシナリオテキスト，元画像を参考に，登場人物の見た目や動き等が適切かどうか」という観点で順位付けしてもらった．実験協力者は 20 代男性 5 名であった．この結果に基づいて，手法毎，順位毎に選択された回数を集計した．

表 1　本実験における比較手法の詳細.

手法	Stable Diffusion で利用する拡張機能			
	txt2img	img2img	AnimateDiff	LoRA
比較 1	✓		✓	
比較 2	✓		✓	✓
比較 3	✓	✓	✓	
提案	✓	✓	✓	✓

表 2　各手法における順位別選択回数.

手法	選択順位			
	1 位	2 位	3 位	4 位
比較 1（SD）	2	10	23	115
比較 2（SD ＋ LoRA）	19	32	78	21
比較 3（SD ＋ img）	24	84	36	6
提案 （SD ＋ img ＋ LoRA）	105	24	13	8

4.3 結　　果

実験結果を表 2 に示す．提案手法では 1 位の選択率が全体の 7 割にあたる 105 件であり，絵伝内の各場面のアニメーションとして他の手法よりも適切なものを生成できたと考えられる．一方，最も低評価（4 位）の割合が高かったのは Stable Diffusion に拡張機能を追加しない手法 1 であった．これらの結果から，img2img や LoRA を導入する提案手法の総合的な有用性を確認した．

また，各手法によってアニメーション化の例を図 4 および図 5 にそれぞれ示す．図 4 は「黒馬に乗って富士山頂を飛ぶ」場面に対する生成結果，図 5 は「経論を読む太子と臣下二人」場面に対するアニメーション化結果である．以降，それぞれの結果に注目して，提案手法の有用性を考察する．

a）「黒馬に乗って富士山頂を飛ぶ」場面：

黒馬に乗って富士山頂を飛ぶのシーンのアニメーション化結果について，使用したプロンプトは以下のとおりである：

- 絵解きシナリオテキスト原文：「黒駒に乗って東国にあそび富士山頂を飛ぶ。富士山、馬に乗った太子、従者」
- 英訳：“Fly to the east on a black horse and fly over the summit of Mt, Fuji, a prince on horseback, and a retinue”
- ネガティブプロンプト：“modern clothing, anime style, sci-fi, cartoonish, surrealism, horror, low resolution, modern buildings, distorted anatomy, text, watermark, futuristic elements”

図 4 におけるアニメーション化結果を見ると，全手法で馬に乗る男性が描写されていることを確認できる．比較 3（SD ＋ img）および提案手法（SD ＋ img ＋ LoRA）においては，元画像と類似した構図が保持されており，絵解きシナリオテキストに沿った自然な動作がみられる．特に，比較 3 と比較して提案手法では，筆致や画風といった絵伝固有のスタイルがより忠実に再現されており，視覚的にも一貫性があるアニメーション化に成功している．

（注3）：stable-diffusion-v1-5: https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5［2025/8/3 閲覧］

b)「経論を読む太子と臣下二人」場面：

経論を読む太子と臣下二人のシーンのアニメーション化結果について，使用したプロンプトは以下のとおりである：

- 絵解きシナリオテキスト原文：「経論披見。経論を読む太子、臣下二人」
- 英訳：“Reading the sutras,The Crown Prince and two of his subjects reading the sutras”
- ネガティブプロンプト：“modern clothes, futuristic elements, anime style, distorted faces, extra limbs, blurry, low resolution, cartoonish, modern furniture, text, watermark, surreal, horror”

図 5 におけるアニメーション化結果を見ると，各比較手法と比べて，提案手法は構図の保持，スタイルの再現性，動きの自然さのいずれにおいても優れていることを確認できる．特に，絵伝画像の特徴的な構図や筆致がより良く保持されている．

4.4 考　　察

実験結果を総括するとともに，提案するアニメーション化手法の改善点について考察する．まず，選好実験の結果から，提案手法（SD ＋ img ＋ LoRA）は各拡張機能を除いた比較手法よりも登場人物の外見や動作，構図の再現性，全体の一貫性において最も高く評価された．よって img2img および LoRA（および AnimateDiff）の導入は，絵解きのための絵伝アニメーション化というタスクにおいて有用であったといえる．

一方，本実験ではアニメーションの長さを約 2 秒としたが，生成結果の中には，動作の大きさや変化に限界があり，人物の表情変化や複雑な動作を十分に描写できない場面も見られた．また，プロンプトの記述内容によって動作の明瞭さや自然さに差が生じることもあった．これに関して，絵解きシナリオテキストを英訳してプロンプトを作成した場合と，ChatGPT によりプロンプトを作成した場合の結果の例を図 6 に示す．これを見ると，絵解きシナリオテキストに基づく生成結果では不自然な雑音が生じているのに対して，自動生成されたプロンプトに基づく結果では比較的滑らかで一貫性があるアニメーションになっていることがわかる．よって，今後は，絵解きシナリオテキストからの最適なプロンプト生成方法を検討する必要がある．

5. ま と め

本報告では，動的で魅力的な絵解きを支援するために，絵伝に描かれた各場面をアニメーション化する手法を検討した．提案手法では，絵伝中の場面に対応する部分画像と絵解きシナリオテキストから Stable Diffusion に基づく画像生成モデルによりアニメーション化する．絵伝の画風を維持した自然で滑らかなアニメーション化のために，text-to-image を基本とする Stable Diffusion を LoRA によりファインチューニングし，image-to-image タスクや時系列画像の出力に対応させる各種拡張機能を組み合わせた．評価実験では，提案手法および各拡張機能を除いた 3 種類の比較手法と比較し，提案手法によって，各場面をよく描写したアニメーション化ができることを確認した．

今後の課題としては，アニメーションの長さの指定やテキストプロンプト生成に関する最適化が挙げられる．また，実際の絵解きにおける生成結果の有用性についても調査していく．

謝　　辞

本研究を実施するにあたり，「聖徳太子絵伝」のスキャンデータおよび絵解き解説シナリオテキストをご提供くださった龍谷大学の阿部泰郎教授および岐阜大学の末松美咲准教授に感謝する．

文　　献

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.10674–10685, June 2022.

[2] D.P. Kingma and M. Welling, “Auto-encoding variational Bayes,” Proceedings of the 2nd International Conference on Learning Representations, pp.1–14, April 2014.

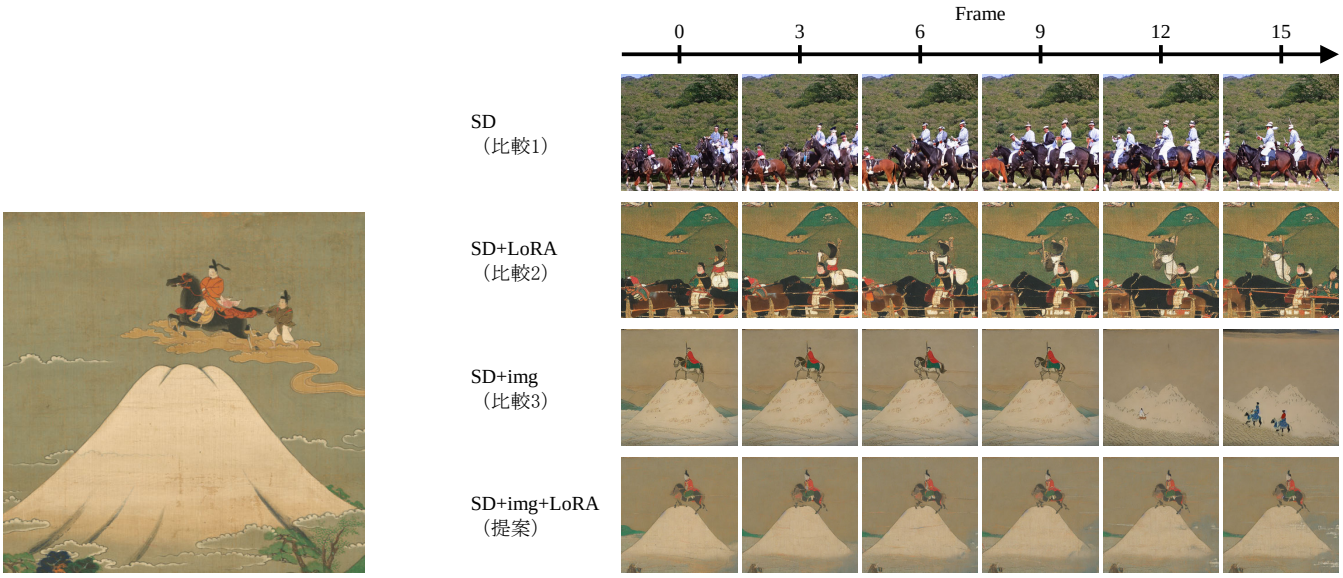
[3] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” Proceedings of the 38th International Conference on Machine Learning, pp.8748–8763, July 2021.

[4] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” Advances in Neural Information Processing Systems, vol.33, pp.6840–6851, Dec. 2020.

[5] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, vol.3, pp.234–241, Nov. 2015.

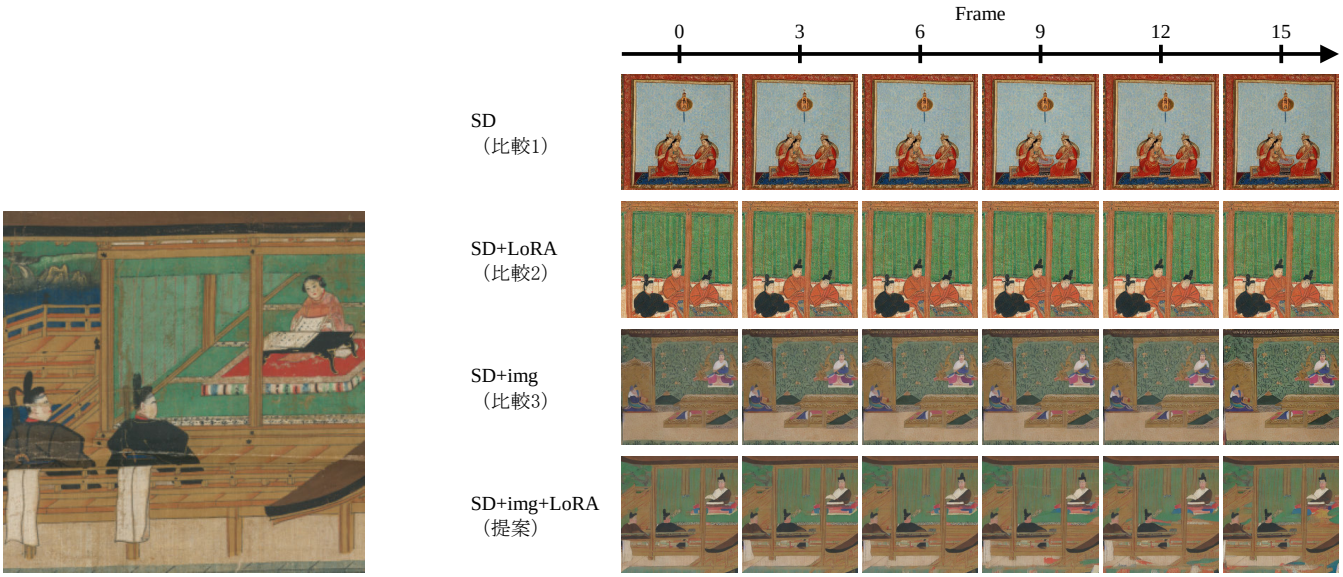
[6] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai, “AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning,” Proceedings of the 12th International Conference on Learning Representations, pp.1–19, April 2024.

[7] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-Rank Adaptation of large language models,” Proceedings of the 10th International Conference on Learning Representations, pp.1–13, April 2022.



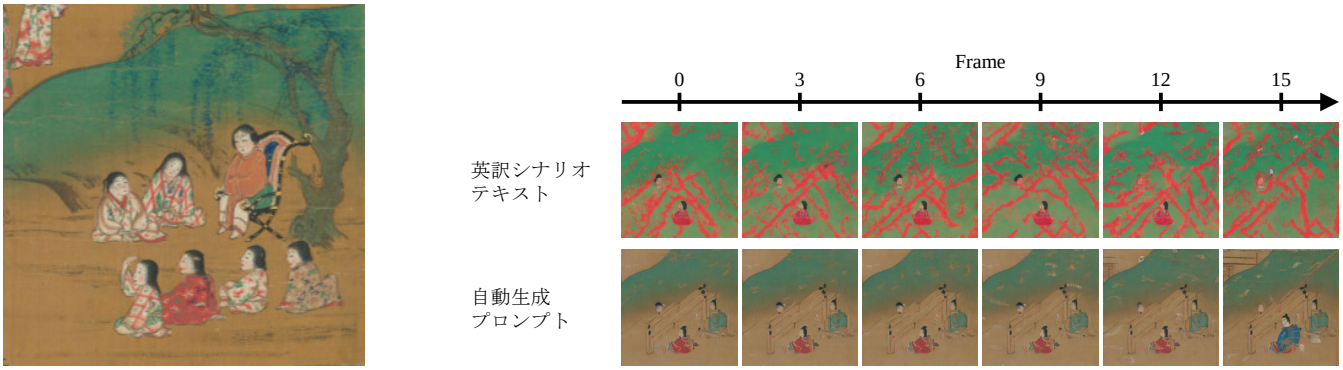
(a) 元画像　　(b) アニメーション化結果

図 4　実験結果の例：「黒馬に乗って富士山頂を飛ぶ」場面のアニメーション化結果.



(a) 元画像　　(b) アニメーション化結果

図 5　実験結果の例：「経論を読む太子と臣下二人」場面のアニメーション化結果.



(a) 元画像　　(b) 生成されたアニメーション

図 6　実験結果の例：プロンプトによる生成結果の違い