G-13

生成 AI による絵コンテ制作とそれを用いた動画生成支援システムの開発

Development of a Video Generation Support System Using Storyboards Created by Generative AI

石橋 孝太郎 †　　原田 利宣 ‡
Kotaro Ishibashi　Toshinobu Harada

1. はじめに

1.1 背景・目的

高齢者や外国人，言語機能に困難を抱える人々とのコミュニケーションでは，言語に依存しない視覚的な表現が有効な補助手段となり得る．中でも動画は，静止画像では表しにくい動きや感情の変化を伝えられることから，非言語コミュニケーションに適していると考える．ここで非言語コミュニケーションとは，言語や文化に依存せずに意味を伝える手段である．

近年，生成 AI の進展により，自然言語を入力することで高品質な動画を生成する技術が発展している．例えば，Runway[1] はテキストや画像を入力として高品質な動画を生成するサービスである．生成 AI を用いることで，デザイン経験の浅い人，撮影機材や編集ソフトを扱うことができない人でも高品質な動画を生成することが可能となった．しかし，現状の動画生成 AI では，出力される動画をプロンプトや始点画像で制御しており，動画の内容を複数カットを含む細部まで制御することは難しい．

そこで本研究では，動画生成 AI を用いて非言語コミュニケーションのための動画生成を支援するシステムの開発を目的とする．ケーススタディとして小説文を採用する．この理由は，小説文は時系列に沿って展開し，複数の登場人物が登場し行動や感情の変化が描写されるためである．このことから，小説文は非言語コミュニケーションとして内容がどの程度伝わるかを検証する対象に適していると考えた．

また本研究では，動画生成 AI の出力を細部まで制御する手段として絵コンテと映画の文法を用いる．絵コンテとは，映画やテレビドラマなどの映像作品の撮影前に用意されるイラストによる表のことである．絵コンテを用いることで，映像内の構図や場面転換を細かく制御することができると考えた．次に映画の文法とは，映画のシーン内の構図およびシーン間のつなぎ方のことである．本研究ではシーンを単一のカメラ位置から連続して撮影されたショット 1 本分とする．映画の制作では，カメラを回す時間が短いショットをつなげることで一つのシーンを構成する方法も一般的であるが [2]，複雑化を避け，研究対象を明確にするため，カメラを長く回して撮影されたショット 1 本分をシーンとして扱う．映画の文法を用いることで，生成される映像を客観的に制御できると考えた．

本研究で開発するシステムではまず，入力した小説文の隣接する文の類似度を計算し，変化点を検出することで物語をシーンごとに区切る．次に，分割した各シーンを画像生成 AI を用いて絵コンテを生成する．最後に絵コンテと映画の文法を知識ベースとして動画生成 AI へ入力し，動画を生成する．

1.2 研究の位置付け

動画による表現について，大野らの研究ではピクトグラムを対象とし，同じ意味を表す静止画と動画ピクトグラムを見せ，意味を答えてもらうアンケートを実施した．その結果，方向性・目的語を持つ動詞において，動画ピクトグラムの方が正答率が高くなることが分かった [3].

動画生成 AI の細かい制御の難しさについて，動画生成 AI モデルを提供する Runway 社の公式のプロンプトガイド [4] では，複数のシーンや詳細なカメラの演出を一度にプロンプトで指示すると，モデルが多くの異なる要素や矛盾する指示を調整しようとし，意図しない結果が生じる可能性がある，と記載されている．同様の記述は他社の動画生成 AI のプロンプトガイドにもみられ，現在の動画生成 AI の出力を細部まで制御することは難しいと考える．

小説文を入力とした動画生成について，星名らの研究では，入力されたテキストを一文ごとに解析することで情報を導出し，その情報をもとに用意した TVML によるアニメーションを生成するシステムを開発した．その結果，想定通りのアニメーションの生成に成功したが，事前に生成する動画に合わせた 3D モデルを用意しなければならないことが課題として挙げられている [5].

以上のように，非言語コミュニケーションの手段としては動画は有効であると考えられる．一方で，既存の動画生成 AI の出力を制御することは難しい．また，既存の小説文を入力とした動画生成システムは，3D モデルの事前準備が必要であり，汎用性に課題があることがわかる．本研究では，生成 AI を活用することで動画生成のための事前準備を不要にし，絵コンテと映画の文法を用いることで生成される動

† 和歌山大学大学院
‡ 和歌山大学

画を制御する．

2. 使用技術

本研究では，Google 社が開発した自然言語処理モデルである BERT[6]，同じく Google 社が開発した LLM である Gemini[7]，Runway 社が開発した動画生成 AI である Runway を用いる．本章では，それぞれの技術について説明する．

2.1 BERT

BERT とは，fine-tuning することで様々な自然言語処理タスクに対して高い性能を示した事前学習済みモデルである．文章を単語へ分割し，単語間および文章間の関係性を学習することで，文脈的にあるべき単語を高い精度で予測することができる．BERT は文章の分類や機械翻訳など，幅広い自然言語処理に用いられている．

本研究では，BERT の派生モデルである Sentence-BERT（以下，SBERT）[8] の学習済みモデルを使用する．SBERT とは，意味の近い文章のペアを学習データとし，それらから得られる文章ベクトルが類似するように BERT をファインチューニングする手法である．本研究では，Hugging Face Transformers の日本語に対応した事前学習済み SBERT モデルである intfloat/multilingual-e5-large[9] を用いて入力した小説文を 1 文ずつベクトル化し，隣接する文のコサイン類似度を算出する．その類似度が低い部分をシーンの分割点とする．

2.2 Gemini

Gemini は Google 社が開発したマルチモーダルな生成 AI モデルである．本研究では，開発システムから API を経由して Gemini を使用する．使用用途は自然言語を入力とした文章の解析・関連語の生成とピクトグラムの生成，画像を入力とした絵コンテの清書の 3 つである．使用する学習済みモデルは，文章の解析に gemini-2.0-flash-lite，ピクトグラムと絵コンテの生成には，自然言語と画像の両方の入力が可能なモデルであり，画像生成が可能なモデルである gemini-2.0-flash-exp-image-generation を用いる．

2.3 Runway

Runway は Runway 社の提供する動画生成 AI である．本研究では，入力した 2 つの画像の間を補間し，動画を生成する目的で使用する．使用する学習済みモデルは，視点と終点の画像を指定可能なモデルであり，公式のプロンプトガイドにてカメラのスタイルの制御が可能であると明記されている Gen3 Alpha Turbo を用いる予定である．このモデルが出力できる動画の長さは 5 秒と 10 秒に制限されており，本研究では 5 秒の動画を生成する．

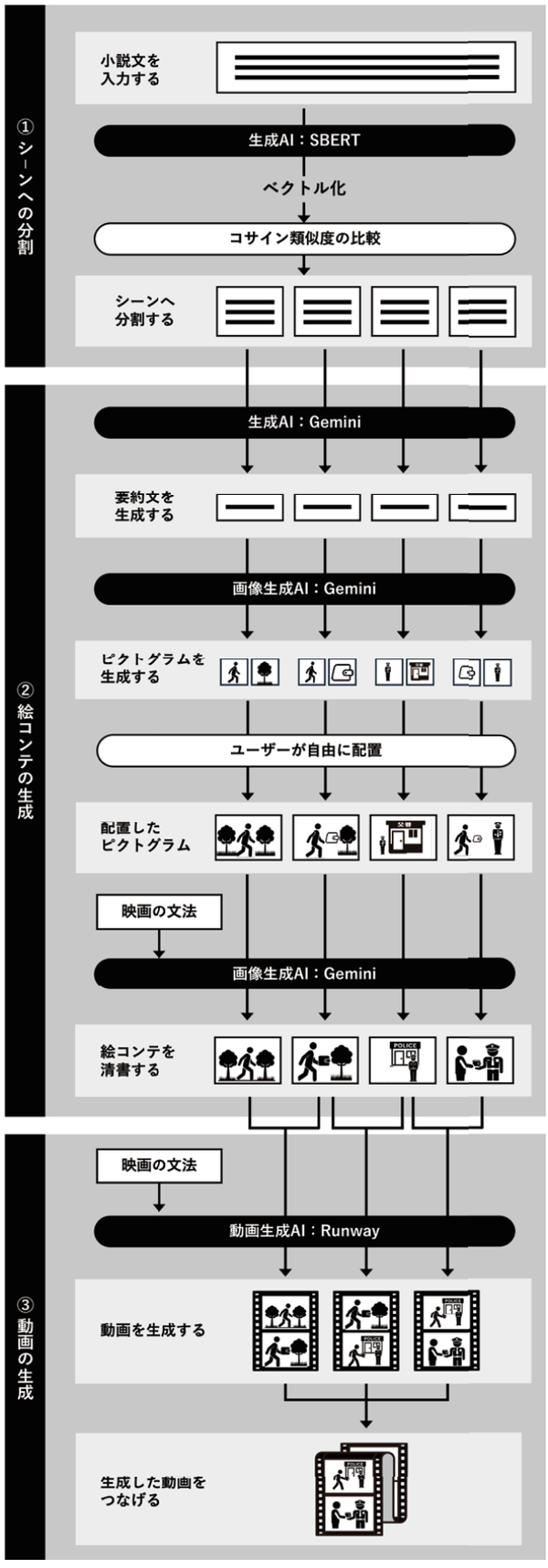


図 1　システムの流れ

3. 開発システム

本章では，開発するシステムの処理の流れについて説明する．システムは①シーンへの分割，②絵コンテの生成，③動画の生成 の 3 つの段階で進める．システムの流れを図 1 に示す．

① シーンへの分割

まず，小説文をシーンへ分割する．具体的には，システムへ入力した小説文を整形し，句点ごとに文単位へ分割する．次に，各文を SBERT を用いてベクトルに変換し，隣接する文同士のコサイン類似度を算出する．この時，類似度が低くなる位置をシーンの境界とみなし，物語を自然な単位にシーンへ分割する．

② 絵コンテの生成

分割した各シーンに対し，Gemini を用いた絵コンテの生成を行う．絵コンテはシステムの利用者がピクトグラムを自由に配置することでラフな画像を作成し，それを Gemini へ入力することで生成する．ピクトグラムを用いる理由は，ピクトグラムがその形状を使って意味や概念を理解させるものであり [10]，言語や非言語コミュニケーションの手段として適していると考えたからである．

具体的には，シーンごとに要約文を作成し，その要約文から主要な言葉（名詞・動詞・形容詞）を抽出する．抽出した言葉ごとに Gemini を用いて関連語を生成する．次に抽出した言葉と生成した関連語を画像生成 AI へ入力して言葉ごとのピクトグラムを生成する．関連語を出力することで，文脈に応じた多様で具体的なピクトグラムを生成することができ，最終的な動画の精度を向上させることができると考えた．生成したピクトグラムを 1 つのキャンバス上に自由に配置し，1 枚の絵を作成する．この絵を画像生成 AI を用いて清書することで絵コンテを作成する．②の作業は全てのシーンで実施する．

③ 動画の生成

②で作成した絵コンテを Runway へ入力することで，動画を生成する．具体的には作成した絵コンテ 2 枚を始点・終点として Runway に入力し，絵コンテ間のフレームを補間する．これを全ての絵コンテで行い，生成された動画を

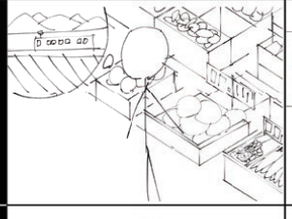
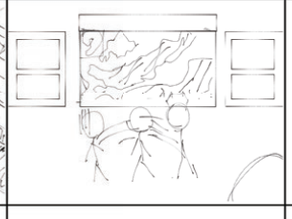
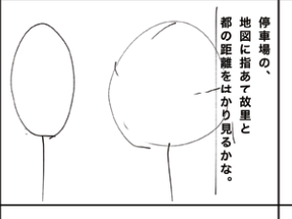
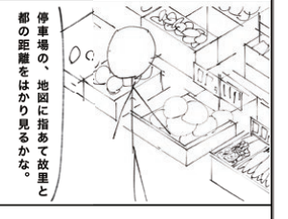
	1枚目	2枚目	3枚目	4枚目
絵コンテ				
リスト番号	0	6	13	40
シーンの開始文	私はよく、ホームシック…	停車場の待合室にはどこの…	ある夜のことであった…	今では、もう停車場…

図 2　実験協力者によって作成された絵コンテ

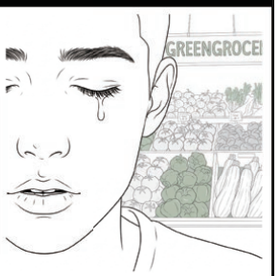
	1枚目	2枚目	3枚目	4枚目
絵コンテ				
リスト番号	0	7	15	34
シーンの開始文	私はよく、ホームシック…	その地図の下に立って…	彼は非常に沈んで…	その帰り道、私は…

図 3　Gemini によって生成された絵コンテ

つなげることで動画を作成する．この時，テキストでは曖昧になりがちなカメラの視点や構図，ショット間のつなぎ方などの映画の文法をプロンプトで制御する．

4. シーンの分割に関する予備実験

本章では，シーン分割と絵コンテの生成について実施した予備実験の概要と結果について記述する．実験は SBERT によるシーンの分割が人間の判断とどの程度一致するかを検証し，手法の妥当性と課題を明らかにすることを目的とした．

4.1 対象作品の選定

実験に用いる作品は，青空文庫 [11] で入手できる小説の中から本文が約 1500 字と短いこと，新字新仮名で表記されていること，複数の登場人物が登場して物語内に明確な場面転換がみられる構成であること，さらに閲覧数が比較的少なく実験協力者に先入観を与えにくいことを条件に選定した．その結果，佐佐木俊朗の「郷愁」を選定した．

4.2 実験協力者によるシーン分割と絵コンテの作成

実験協力者には和歌山大学在学中の 3 年生 1 名を選定した．日常的にイラスト制作を行なっており，絵コンテについて一定の知識があると判断した．

実験ではまず,実験協力者に「郷愁」の全文を読んでもらった後，内容を絵コンテ作成のために複数のシーンへ分割し，シーンが切り替わった直後の文章を答えてもらった．次に，分割した各シーンを表す絵コンテを作成してもらった．この時，シーンは内容を伝えるために必要最低限の個数と指示した．また，絵コンテは登場人物の配置とカメラアングルが把握できる程度の簡単なものと指示し，制限時間は設けなかった．

実験の結果，実験協力者は小説文を 4 つのシーンへ分割し，4 枚の絵コンテを提出した．実験協力者が作成した絵コンテとシーンの開始文のリスト番号，シーンの開始文を図 2 に示す．

4.3 AI を用いたシーン分割と絵コンテの生成

本節では，SBERT を用いたシーン分割と，Gemini を使用した絵コンテの生成について説明する．

まず，シーン分割には 3 章で述べた通り学習済みモデルである intfloat/multilingual-e5-large を用いた．小説文を文単位にリスト化し，学習済みモデルを用いて各文を 1,024 次元のベクトルへ変換した．隣接分同士のコサイン類似度を算出し，類似度の高い箇所と低い箇所の差が 0.02 以上となる位置をシーンの境界として抽出した．この結果，小説文は 4 つのシーンへ分割された．

絵コンテの作成には Gemini を使用した．本実験では，API を経由するのではなく，Google が提供する Web インターフェース上で作業を行なった．本実験で用いたモデルは Gemini 2.5 Pro である．まず，物語の全文のテキストデータを与え，各シーンの文章を個別に入力することで絵コンテの生成を指示した．このとき，画風を「モノクロかつシンプルな」と指示した．これは，実験協力者が作成した絵コンテと画風を合わせるためである．

Gemini によって生成した絵コンテとシーンの開始文のリスト番号，シーンの開始文を図 3 に示す．

4.4 シーン分割の結果の比較

本節では，実験協力者によるシーンの分割と SBERT を用いた分割の一致度を算出し，結果について考察する．

シーン分割の一致度の算出には，WindowDiff[12] を用いた．WindowDiff とは，文章を長さ k の「窓」に区切り，その窓を一文ずつ前へずらしながら，その窓の中に何本の分割線が引かれているかを 2 つの分割結果で比べる手法である．各窓で両者の分割線の本数が異なる場合，誤差を 1 と数え，全ての窓の誤差平均を一致度として算出する．したがって，値が 0 に近いほど 2 種類の分割は一致しており，1 に近いほど一致していないと解釈できる．本実験において窓の長さ k は 10 とし，実験協力者による分割結果のリスト番号と SBERT による分割結果のリスト番号の 2 つを入力して一致度を算出した．

WindowDiff によって算出した分割の一致度は 0.3784 であり，4 割弱の窓で分割の本数が食い違ったことが分かる．つまり，大まかなシーンの切り替わりは一致しているが，細部ではズレが残っていると考えられる．この結果になった理由として，実験協力者と SBERT でシーンの境界の判定規準に違いがあると考えられる．実験協力者は，描写されている場所や時間が切り替わった文でシーンを分割している．反対に SBERT を用いた分割では，シーンの境界の直後の文に「その」「彼」などの指示語があり，登場人物の感情を示す動作が見られる文が検出された．このことから，SBERT を用いた分割では指示語による文同士のつながりは重要視されておらず，登場人物の感情の変化をシーンの境界として分割を行ったと考えられる．このようなシーンの境界に対する判定規準の違いが，WindowDiff で算出した誤差につながったと考えられる．より誤差を減らす方法として，SBERT やコサイン類似度による分割点の検出におけるパラメータの調整や，WindowDiff の窓の長さの調整が考えられる．

4.5 絵コンテの比較

次に絵コンテを比較する．Gemini を用いて生成した絵コンテの特徴として，絵コンテ間の画風が統一されておらず，登場人物に一貫性がないことが挙げられる．そのため，画

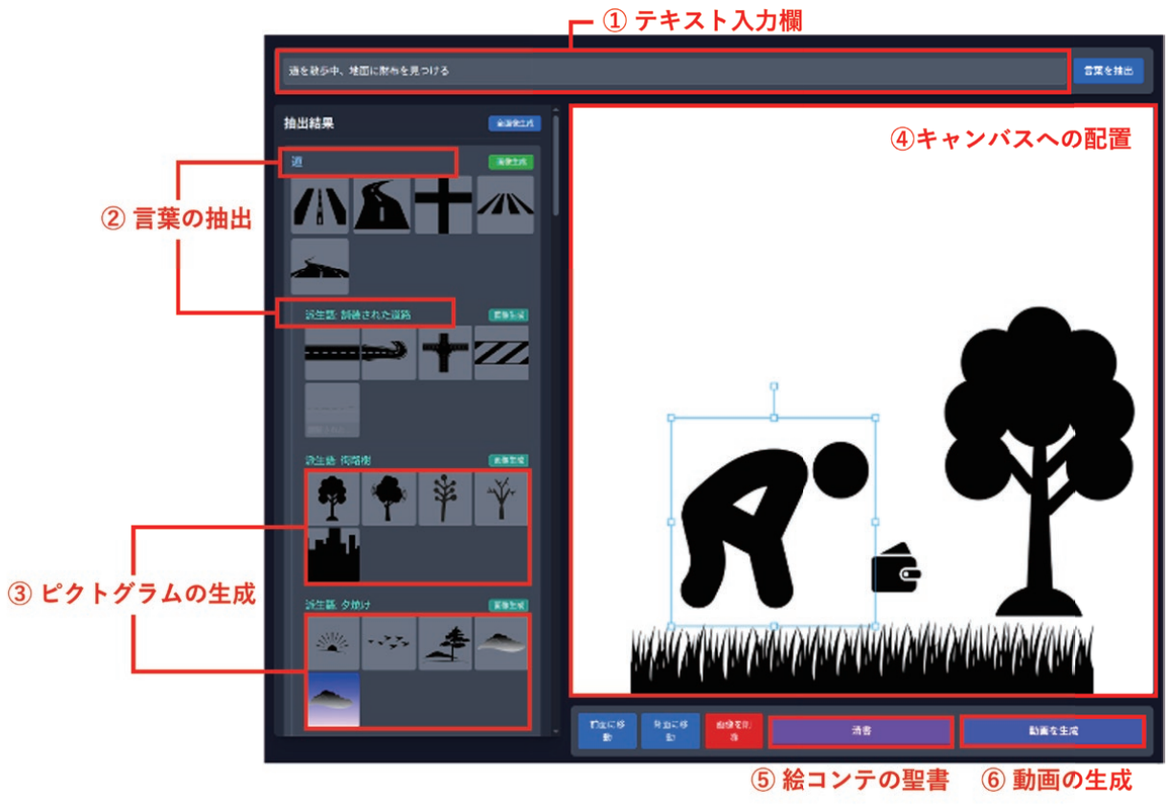


図 4　開発したシステムのインターフェース

風については比較せず，構図やアングル，描写している内容について比較する．

実験協力者が作成した絵コンテと Gemini が生成した絵コンテを比較したところ，被写体の配置やカメラのアングルには違いが見られるものの，主要な登場人物や背景を描写している点は概ね共通していた．

実験協力者は，場所や時間が切り替わる箇所を境界としてシーン分割を行っていたことから，状況を説明するような絵コンテが作成されたと考えられる．SBERT を用いたシーン分割では，感情が変化する箇所を境界として分割していたが，Gemini によって感情が明確に描かれた絵コンテは 1 枚のみであった．結果として，絵コンテはどちらも状況を説明するような描写がなされた．この理由として，シーンを分割する位置が異なっていても，両者とも各シーンにおける「絵コンテに直結する重要な描写」が共通していたからであると考えられる．

以上より，SBERT を用いたシーン分割と Gemini による絵コンテ生成を組み合わせた本手法は，人手による結果と近しいものであり，一定の妥当性を有すると考える．しかし，本実験では実験協力者が 1 名であったため，個人差の影響が大きい可能性がある．今後は協力者の数を増やし，WindowDiff のばらつきや絵コンテの多様性を把握することで，より詳細な検証を行う必要があると考える．

5. システムの開発

本研究では，小説文を入力として動画の生成を行うシステムを開発することを目的としている．本章では Vue と FastAPI を利用して開発したシステムについて説明する．

本システムは，入力された文章から関連語を生成し，各言葉に対応するピクトグラムを一覧で表示する．その後，ユーザーが選択したピクトグラムをキャンバス上に自由に配置することができる．このキャンバスを入力として Gemini を用いて清書を行い，絵コンテの画像を生成する．さらに，この絵コンテの画像を入力として Runway を用いて動画の生成を行う．文章や語句の解析およびピクトグラムの生成には Gemini の API を使用し，画像を入力とした動画の生成には Runway の API を使用している．図 4 にシステムのインターフェースを示す．

6. まとめと今後の課題

本研究では，小説文を入力としてシーン分割・絵コンテ生成・動画生成までを行う支援システムの開発を目的とし，SBERT を用いたシーン分割，Gemini による絵コンテ生成，Runway による動画生成を一連の流れで試作・検証した．

シーンの分割に関する予備実験の結果，SBERT を用いた自動シーン分割と人手による分割は，大きな場面転換は一

致しているが，細部ではズレが残る結果であった．また，絵コンテにおいては，シーンの分割位置が一致していなかったにも関わらず，出力された内容には共通の点が多く見られた．

また，Vue と FastAPI を用いて開発したシステムでは，入力された一文から関連語を抽出し，ピクトグラムを配置・編集する機能や，清書・動画生成までの一連の流れを実装した．ただし現段階では，一つの文に対する処理に限定されており，複数の文を含む物語全体の入力や，複数の動画の出力を行う機能は未実装である．

今後の課題として，シーンの分割精度の向上や実験協力者の数を増やした実験による精度の再検証が必要である．また，現在のシステムにシーンの分割や複数枚の絵コンテを生成する機能を実装することで，より実用的な動画生成支援システムの構築を目指す．

参 考 文 献

［1］ Runway.“Runway| Tools for human imagination.”. Runway. 2025. https://runwayml.com/,（最終閲覧日 2025-07-24）.

［2］ D.Arijon, 岩本憲児 , 出口丈人（訳）. 映画の文法 . 紀伊國屋書店 . 1980

［3］ 宗森純 , 大野純佳 , 吉野孝 . 絵文字チャットによるコミュニケーションの提案と評価 , 情報処理学会論文誌 , 47, 7, 2071－2080, 2006

［4］ Runway.“Creating with Video to Video on Gen-3 Alpha and Turbo”. Runway. 2024. https://help.runwayml.com/hc/en-us/articles/33350169138323-Creating-with-Video-to-Video-on-Gen-3-Alpha-and-Turbo,（最終閲覧日 2025-07-24）.

［5］ 星名研吾 , 野口武紘 , 杉本徹 , 榎津 秀次 . 物語理解シミュレーションの試み：物語テキストからアニメーション自動生成を通して , 日本認知科学界第 29 会大会 , pp1-7, 2012

［6］ Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.“BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171-4186, Minneapolis, 2019. Association for Computational Linguistics.

［7］ Google.“Gemini – A Family of Multimodal Models”. Google DeepMind. 2025. https://deepmind.google/models/gemini/,（最終閲覧日 2025-07-24）.

［8］ Reimers, N., and Gurevych, I.“Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3982–3992, Hong Kong, 2019. Association for Computational Linguistics.

［9］ intfloat.“multilingual-e5-large”. Hugging Face. 2024. https://huggingface.co/intfloat/multilingual-e5-large,（最終閲覧日 2025-07-24）.

［10］ 太田幸夫：ピクトグラムのおはなし , 日本規格協会 , pp.13, 1995.

［11］ 青空文庫 . https://www.aozora.gr.jp/（最終閲覧日 2025-07-24）.

［12］ Lev Pevzner, Marti A. Hearst. A Critique and Improvement of an Evaluation Metric for Text Segmentation. Computational Linguistics, 2002.