大規模言語モデルによるセリフを含む物語の可視化手法

瀧本隼矢　竹内孔一
岡山大学大学院 環境生命自然科学研究科
p00z6xwg@okayama-u.ac.jp

概要

近年，大規模言語モデル（LLM: Large Language Model）を活用した物語の視覚的表現が注目されているが，物語のセリフや感情，キャラクターの行動といった本質的要素を取り入れる試みは十分に行われていない．本研究では，LLM を用いて Chain of Thought（CoT）Prompting により物語を解析・分割し，画像生成プロンプトを構築する新たな手法を提案する．本手法の特徴は，物語中のセリフを含めた可視化に焦点を当てる点にある．CLIP Score による評価実験では，提案手法が物語本文からそのまま画像生成する手法に比べ，セリフの有無に関わらず視覚的情報が画像に正確に反映されることを示す．

1　はじめに

近年，大規模言語モデル（LLM: Large Language Model）を用いた物語を視覚的に表現する手法が議論されている [1] [2][3]．しかし，物語の内容を的確に解析し，視覚的な情報を抽出・再構成する手法には，いまだ多くの課題が残されている．その中でも，キャラクターのセリフや感情，行動といった物語の本質的な要素を取り入れる試みは十分に行われていない．

本研究では，LLM を活用し，物語のセリフを考慮しながら，物語を Chain of Thought（CoT）Prompting により解析・分割し，画像生成プロンプトを作成する新たな手法を提案する．提案手法の最大の特徴は，セリフを含めた物語の可視化にある．従来の CoT Prompting による物語の分割手法は，イベントやハイライトの識別に焦点を当てるものの，セリフを十分に考慮していないため，物語の雰囲気やキャラクターの感情などをうまく表現出来ていなかった [1] [2][3]．そこで提案手法では，セリフを含む形で物語を解析・分割し，各シーンに対して描画情報を整理し，画像生成プロンプトを作成する．

ChatGPT[4] で作成した短編物語から各シーンの画像を生成し，CLIP Score[5] で評価し，提案手法は物語本文を直接分割する方法に比べ，セリフの有無に限らず視覚的情報が画像に反映されることを示す．

2　関連研究

近年，物語を可視化する研究が議論されている．物語をシーンごとに分割し，画像をストーリーに沿って作成する DreamStory[1] や，キャラクターの一貫性に着目した Diffusion ベースのモデルの StoryDiffusion[2]，CoT Prompting を活用し，文章から一つの画像を生成する Narrative-to-Image[3] がある．Narrative-to-Image では与えられたナラティブに沿った画像を，CoT Prompting で生成しており，キャラクターの描画は行わないものの，文章の内容に合った画像生成を実現している．DreamStory や StoryDiffusion は，特に物語を可視化することに注目しており，動画生成も可能になっている．これらの手法では，登場するキャラクターが異なるシーンでも一貫しているかについて，特に議論している．物語を可視化する画像生成手法では，動画生成も同時に可能となる応用が期待できる．提案手法では既存手法のような視覚的な情報のみの可視化に加え，セリフを考慮することで物語の雰囲気やキャラクターの感情を表現した画像生成を行うことができる．

3　提案手法

提案手法ではまず物語を可視化するために，LLM を用いてキャラクターの行動，セリフや物語の話の流れの区切となる部分ごとに物語を分割する．次に，同様に LLM を用いてその分割したシーンから描画情報を生成し，その情報を元に画像生成プロンプトを生成する．最後に，画像生成モデルに画像生成プロンプトを入力することで，そのシーンに合った画像を生成する．図 1 に提案手法のフローを示す．

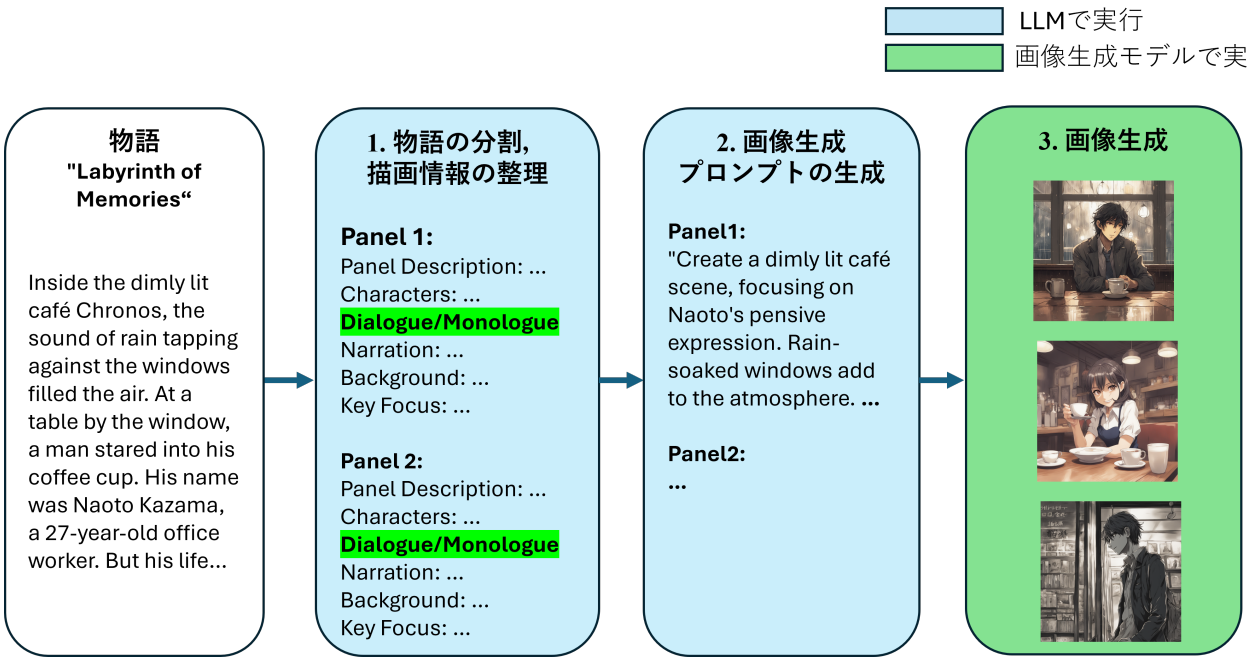

図 1　提案手法のフロー

3.1　LLM による情報生成

LLM でタスクを解く際，CoT Prompting が有効であることが一般的に知られており [6]，特に複雑な推論が必要となる際に，推論のステップを介することでより正確な出力を得ることが可能である．物語を可視化するための画像を生成するために，物語を各シーンに分割し，それぞれのシーンに対応した描画情報，視覚的情報を明確にする必要がある．以下に，CoT Prompting を活用し，それらの情報を LLM から生成する手法を示す．

3.2　描画情報の整理

本節では，物語から画像を生成するために必要な描画情報を LLM を用いて整理する．

入力する物語の文章全体をそのまま 1 枚の画像として描画することは難しい．同様に文章全体を分割し，それに対応した画像を生成するには，その分割の基準を定めなければならない．そのため，それぞれどのようなシーンに対して，どのような基準で描画の情報を分割していくか，明確にする必要がある．そこで，提案手法では以下の手順でセリフを考慮したシーン分割，そのシーンに対応した描画情報の生成を行う．

1. 文章全体を分析
2. イベントやハイライトの識別
3. 描画情報の生成

まず，文章全体を分析するために物語の話全体を LLM に入力する．次に，物語の話の展開，キャラクターの感情や行動，セリフ，設定の変化などを検出し，それらの要素で物語を分割する．最後に，それぞれシーンごとに分割した物語に対して描画情報を割り当てる．LLM がこれらの手順に従うよう，CoT Prompting を活用する．

3.2.1　物語の分割

物語をシーンごとに分割するために，以下の要素を検出する．

- キャラクターの具体的な行動や物語の出来事
- キャラクターの感情や行動
- 物語の進行やキャラクターの心情を表すセリフ
- 時間や場所，雰囲気，登場人物の変化などの環境の変化

これらの要素を検出した後，それぞれのイベントを整理し，自然な流れでシーンとして構築することで，セリフを考慮した物語の描画情報の生成が可能になる．

3.2.2　描画情報の生成

キャラクターの感情，行動，セリフ等に着目した後，そのシーンに対する描画情報を生成する．提案手法では，各描画シーンに対する描画情報として以

表 1　物語本文と分割後から生成した描画情報の例

物語本文	物語分割後から生成した描画情報
The night sky was always her ally. For 17-year-old Kotono, gazing at the stars was her only solace. Her life in a small town was bound by the constraints of academics and her family's expectations, and under the starlit sky was the only place where her mind could roam freely.	**Description:** Kotono stands on her balcony, gazing at the starry night sky. **Characters:** Kotono. **Dialogue/Monologue:** None. **Narration:** "For Kotono, the night sky was her sanctuary, a place where her mind could wander freely." **Background:** A view of the starry sky with a silhouette of Kotono. **Key Focus:** Kotono's serene expression and the beautiful night sky.
But tonight was different. The sky, usually adorned with countless stars, was eerily pitch-black. Not a single star was visible. Overcome by a mix of shock and unease, Kotono dashed out of her house and hurried to the nearby hill. "Why...?"	**Description:** Kotono is shocked as she notices the absence of stars in the sky. **Characters:** Kotono. **Dialogue/Monologue:** Kotono (thought bubble): "Why...?" **Narration:** "But tonight, something was amiss. The sky, usually filled with stars, was an eerie darkness." **Background:** A dark, starless sky with Kotono's shocked expression. **Key Focus:** Kotono's expression of disbelief and the empty sky.

下の形式をとる．

- Description（シーンに対する説明）
- Charactors（シーンに登場するキャラクター）
- Dialogue,Monologue（キャラクターのセリフ）
- Narration（シーンのナレーション）
- Background（シーンの背景）
- Key Focus（重要視する描画情報）

これらの形式を採用することにより，キャラクターのセリフ，物語の話の流れに沿った描画情報を生成する．物語から描画情報を生成するこれらの手順に沿って得られる描画情報例を表 1 に示す．

3.3　画像生成プロンプトの生成

3.2 節で生成した描画情報を元に，画像生成プロンプトを生成する．各シーンに対してキャラクターの見た目，感情，行動，背景の情報として環境，照明，トーンなどを分析する．そして，視覚的情報を特定し，画像生成に使用する感情，テーマを絞り込む．最終的に，これらを元に画像生成プロンプトを生成する．

4　評価実験

4.1　評価方法

ChatGPT で作成した 4 章構成の英語の短編物語 4 本に対し，物語本文からそのまま画像生成する方法と，提案手法を比較した．前者では，物語本文を適切と考えられる部分で分割し，そのまま画像生成プロンプトとして扱った．また，画像生成の際は出力される画像の画風を統一するために，プロンプトの先頭に「anime-style illustration」を追加した．各手法について，LLM は Command R[7] の「`c4ai-command-r-08-2024-Q4_K_M.gguf`」モデル，画像生成に使用するモデルは Stable Diffusion XL（SDXL）[8] を使用した．

表 2　ChatGPT で生成した物語を直接画像生成する手法と提案手法の CLIP スコア比較

物語名	SDXL	LLM + SDXL (ours)
The Weaver of Stars	0.3231	**0.3474**
The Secret of the Clock Tower	0.3218	**0.3517**
Labyrinth of Memories	0.3223	**0.3499**
The Crystal of Light and the Hero of Darkness	0.3228	**0.3625**
Overall Average	0.3225	**0.3527**

評価指標として，テキストと画像との特徴の類似度を表す CLIP Score を用いる．物語本文からそのまま画像生成する手法では，画像生成に使用するために分割した物語本文をテキストとし，それによって生成される画像との CLIP Score を計算する．提案手法では，物語本文から生成された画像生成プロンプトをテキストとし，それによって生成される画像との CLIP Score を求める．

4.2　実験結果

表 2 に，各短編小説に対して各手法の CLIP Score を示す．各短編物語ごとの CLIP Score，全体の CLIP Score を比較しても，提案手法が上回った．

表 3 に，セリフを含むシーンについて各手法の CLIP Score を示す．表 2 と同様に，提案手法の CLIP Score が上回った．物語本文からそのまま画像生成する手法では，表 2 の CLIP Score から表 3 の CLIP Score は低下したが，提案手法では表 2 と表 3 の CLIP Score はほぼ変動しない結果となった．

表 3 セリフが含まれる物語シーンにおける画像生成手法の CLIP スコア比較

物語名	SDXL	LLM + SDXL (ours)
The Weaver of Stars	0.3183	**0.3447**
The Secret of the Clock Tower	0.3163	**0.3508**
Labyrinth of Memories	0.3089	**0.3502**
The Crystal of Light and the Hero of Darkness	0.3123	**0.3631**
Overall Average	0.3138	**0.3534**

4.3 考察

物語本文をそのまま画像生成する手法では，画像生成プロンプトに適していないためか，CLIP Score が低い結果となった．一方，提案手法では画像生成プロンプトへ，より視覚的な情報を組み込むようになり，CLIP Score は上昇したが，画像として表現が難しい，「The woman, with a serious tone」といったものが含まれていた．これはセリフを喋るトーン，声に焦点を当てているため，このような情報が画像生成プロンプトに含まれると CLIP Score が低くなる可能性がある．提案手法における画像生成プロンプトで，より視覚的な表現へ改善することができれば，より CLIP Score も上昇すると考えられる．

提案手法では，他の研究にはない「セリフを含めた物語の可視化」が実現されている．物語中のセリフは，各シーンにつき必ずしも 1 つに限定されるわけではなく，セリフのないシーンや，キャラクターが複数いる場合にセリフが複数存在するシーンなど，さまざまなパターンが含まれる．本手法は，これらの多様なセリフの数にも柔軟に対応している．さらに，セリフを含めた物語の可視化においても，表 3 の結果から定量的に優れた性能を示している．具体的には，物語本文をシーンごとに分割する明確な基準が確立されていない現状において，本提案手法は物語分割の新たなアプローチとして有効であると考えられる．

5 おわりに

本稿では，LLM を用いたセリフを含めた物語の可視化の手法を検討した．物語本文から，LLM を用いて物語を分析し，イベントやハイライトの識別を行い，セリフやキャラクターの感情や行動を考慮しつつ物語を分割した．その後，分割した物語それぞれの描画情報を整理し，フォーマットに沿って画像生成プロンプトに使用するための描画情報を生成する手法を提案した．そして，その描画情報から画像生成プロンプトを生成する手法を検討した．CLIP Score を用いた評価実験では，提案手法を用いることでセリフの有無において物語がより画像に反映されることが分かった．また，他の研究には無いセリフを含めた物語の可視化として，物語本文から画像生成するよりも定量的に良い手法であることを示した．

今後は，物語の描画情報から画像生成プロンプトを作成する際に，登場するキャラクターの表情，感情など，表現が難しい表現を除いていき，より視覚的な情報を生成できるよう検討していきたい．

参考文献

[1] Huiguo He, Huan Yang, Zixi Tuo, Yuan Zhou, Qiuyue Wang, Yuhang Zhang, Zeyu Liu, Wenhao Huang, Hongyang Chao, and Jian Yin. DreamStory: Open-Domain Story Visualization by LLM-Guided Multi-Subject Consistent Diffusion, 2024.

[2] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. StoryDiffusion: Consistent Self-Attention for Long-Range Image and Video Generation. **NeurIPS 2024**, 2024.

[3] 岩崎翔, 粟井修司, 青木俊彦, 石塚昌平, 紺野剛史. Narrative-to-image: ナラティブに合致した画像の自動生成. 人工知能学会全国大会論文集, Vol. JSAI2024, pp. 4Xin220–4Xin220, 2024.

[4] OpenAI. ChatGPT, Accessed:2024-12-20. https://openai.com/chatgpt.

[5] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning, 2022.

[6] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, 2023.

[7] Cohere. Command R, Accessed:2024-12-20. https://docs.cohere.com/docs/command-r.

[8] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis, 2023.

A　付録

本節では，提案手法による画像生成の結果を示す．図 2 に，ChatGPT で生成した物語「Labyrinth of Memories」の提案手法によって生成した画像と，それに対応したセリフを示す．

図 2　提案手法で生成した画像とセリフの例