レイアウト構造木を介したマンガページ生成手法：視覚・テキスト特徴の有効性比較

馮思遠 [1]　林克彦 [1]　上垣外英剛 [2]　鷲尾光樹 [3]　平尾努 [4]
[1] 東京大学　[2] 奈良先端科学技術大学院大学　[3] フリーランス　[4] 金沢大学
{9445233883,katsuhiko-hayashi}@g.ecc.u-tokyo.ac.jp　kamigaito.h@is.naist.jp
kkwashio3333@gmail.com　hirao@se.kanazawa-u.ac.jp

概要

マンガページにおけるレイアウトは，読解順序や物語理解に大きな影響を与える重要な構造要素である．しかし，マンガページのレイアウト生成に関する計算的なアプローチは，まだ十分に研究されていない．本稿では，マンガページのレイアウトを木構造として明示的に表現し，各コマの内容に関するテキスト特徴や視覚特徴を使用して，レイアウト構造の生成と具体的な配置座標の生成を行った．実験結果から，ページレイアウトはテキスト内容よりも視覚的構図により強く依存して決定されることが示され，さらに推定されたレイアウト構造が座標生成において非常に有効に機能することが確認された．

1　はじめに

マンガは，テキスト，画像，およびページ内レイアウトが密接に結びついたマルチモーダルな視覚叙事媒体である．とりわけ，1 ページ内に配置された複数のコマの空間構成は，時間的進行や視線誘導，物語的強調に関与する重要な要因であることが，理論研究 [1, 2, 3] および実証研究 [4, 5, 6] の双方から示されてきた．さらに近年では，ページ内レイアウトの特徴のみから作品の識別が可能であることも報告されている [7]．しかし，このようなページレイアウトがどのような情報に基づいて決定されるのかは，計算的観点から十分に明らかにされていない．

近年の計算マンガ研究では，コマ検出 [8, 9]，吹き出し検出 [10, 11]，キャラクター同定 [12, 13]，発話者推定 [14, 15] など，ページ内の局所要素を対象とする解析技術が大きく進展してきた．一方で，ページ全体の分割様式やコマ間の階層的関係としてのレイアウト構造を，明示的な予測対象として扱う研究は依然として限定的である．また，文書や UI 分野では，要素列から配置を生成する手法や系列変換としてレイアウトを扱う枠組みが提案されており [16, 17]，大規模言語モデルを用いた配置推論も報告されている [18, 19, 20]．しかし，これらはいずれも単一画面内の要素配置を主対象としており，複数の視覚単位へとページを分割するマンガレイアウトとは問題設定が異なる．さらに，マンガ生成研究では，テキスト記述からコマ画像を生成する手法が提案されている一方 [21, 22]，ページレベルのコマ配置構造はテンプレートや外部指定に依存する場合が多く，十分に扱われてこなかった．

そこで本稿では，マンガページのレイアウトをページ分割およびコマ間関係から成る中間的な構造表現として明示化し，テキスト特徴および視覚特徴を入力としてこれを推定する手法を提案する．一連の実験を通じて，レイアウト構造推定においては，コマキャプションの語義的内容よりも，コマ画像が有する視覚的構図情報が支配的な役割を果たすことを示す．さらに，推定されたレイアウト構造に基づく座標生成により，ページ配置生成におけるその有効性を検証する．

2　提案手法

本稿では，マンガページのレイアウト生成を，**（1）ページ全体の分割構造の推定**と**（2）推定された構造に基づくコマ配置の決定**から成る二段階の問題として捉える．全体の処理フローを図 1 に示す．

レイアウト構造推定　第 1 段階では，読み順に並べられたコマ列を入力として，ページ全体の空間的分割関係を表す**レイアウト木**を推定する．本稿では，先行研究 [23] に倣い，レイアウト木を縦分割（Row），横分割（Column），およびコマに対応する葉ノード（Panel）から成る再帰的な階層構造として定義する（図 2）．この木構造は，S 式として線形化す

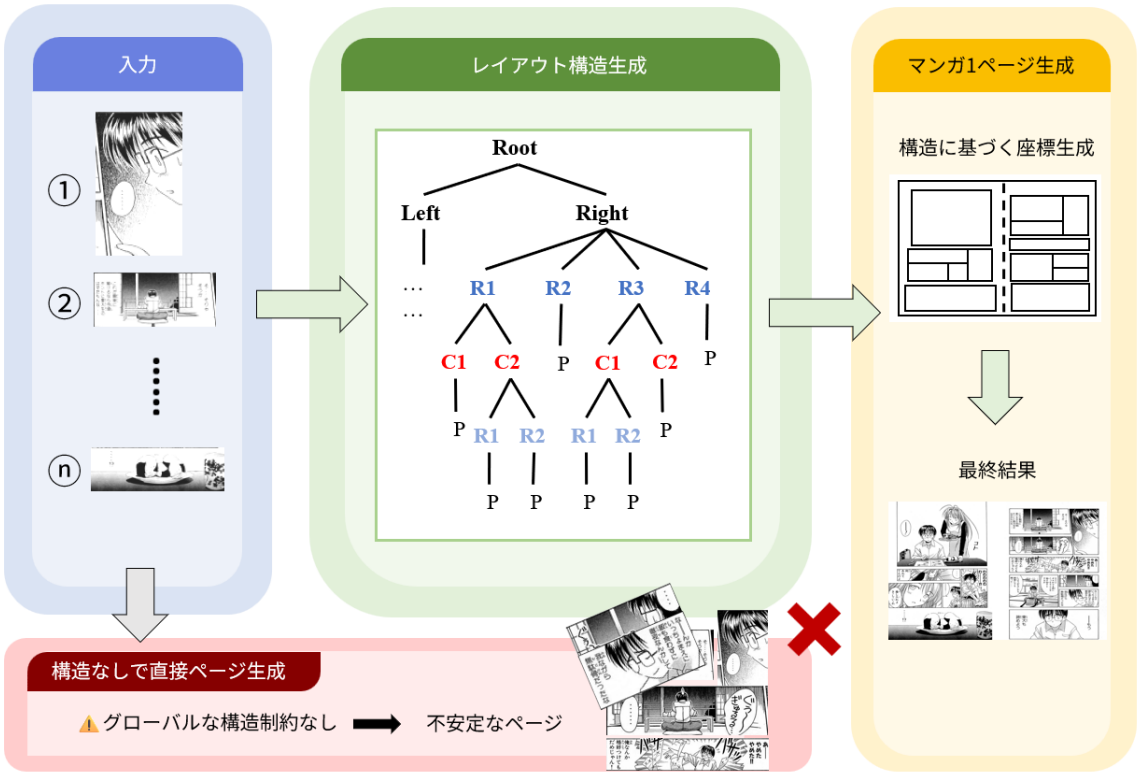

図 1　提案するマンガページ生成フレームワークの全体像.（『ラブひな』© 赤松 健）

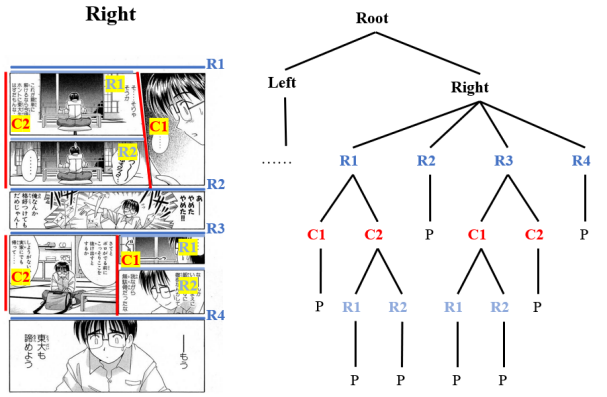

図 2　マンガページに対する再帰的な空間分割（左）と，それに対応するレイアウト木（右）の例．本図では，マンガページのうち右半ページのみを対象として示している．ページは左右領域に分割された後，縦分割（R）および横分割（C）を用いて再帰的に分割され，葉ノード（P）が各コマに対応する．（『ラブひな』© 赤松 健）

ることで，系列モデルによる生成を可能にする．

構造に基づく座標生成　第 2 段階では，第 1 段階で推定されたレイアウト木を中間表現として用い，ページ分割構造に基づいてコマ間の相対関係や占有領域といったページ全体の大域的な配置骨格を定める．その上で，各コマの内容特徴を参照しながら境界位置やサイズを調整し，視覚的バランスと局所的整合性を反映した配置を生成する．

3　データセットと実験設定

データセット構築　本稿では，Manga109 コーパス [8, 9, 24, 25] に含まれる全 109 作品（10,602 ページ・103,849 コマ）を対象に，LLaVA-v1.6-34B [26] を用いて各コマ画像から英語キャプションを自動生成し，**Manga109Caption データセット**を構築した．キャプション生成に用いたプロンプトの詳細は付録に示す．4 コママンガ作品を除外した上で，残る作品を作品単位で訓練／開発／テスト = 8:1:1 に分割し，それぞれ 83，11，10 作品を含む．

タスク定義　読み順に並べられた N 個のコマ特徴表現の列 $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ を入力とし，ページ全体の分割構造を表すレイアウト木と各コマの配置座標を出力する問題として定式化する．本タスクは，レイアウト構造推定とそれに基づく座標生成の二段階から構成される．

コマ特徴表現（入力）　特に断りのない限り，テキスト特徴には Sentence-Transformers all-MiniLM-L6-v2 [27]，視覚特徴には CLIP ViT-B/32 [28] を用いて，各コマから特徴を抽出する．

レイアウト構造推定モデル　コマ特徴表現の列 $\mathbf{X}$ を入力とし，S 式の構造トークン列 $\mathbf{Y} = (y_1, \ldots, y_T)$

表 1　レイアウト構造推定の性能.

設定	TED ↓
(Base) Random	0.40
(Base) kNN (text)	0.40
(Base) kNN (vis)	0.39
(Main) Text-only	0.33
(Main) Visual-only	0.21
(A) Weak prompt	0.33
(B) CLIP ViT-L/14	0.20
(B) MAE ViT-B/16	0.20
(B) MAE ViT-L/16	0.18
(C) High freq.	0.19
(C) Low freq.	**0.15**

を生成する条件付き系列生成問題として定式化する．エンコーダには BiLSTM を用いてコマ系列を文脈化し，その出力として得られる文脈表現 $\mathbf{H}$ を，Transformer デコーダが参照しながら構造トークンを自動回帰的に生成する．推論時には，括弧構造の整合性と葉ノード数一致を保証する文法制約付きデコーディングを行う．学習には交差エントロピー損失を用い，評価指標として Tree Edit Distance（TED）[29] を採用する．モデルの詳細は付録に示す．

レイアウト座標生成モデル　コマ特徴表現の列 $\mathbf{X}$ と，レイアウト木から得られる構造的位置表現列 $\mathbf{S} = (\mathbf{s}_1, \ldots, \mathbf{s}_N)$ を入力とし，各コマの配置座標 $b_i = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$ を予測する回帰問題として定式化する．本モデルは，構造的位置表現を処理する Transformer ベースの構造エンコーダと，内容特徴に基づいて配置を補正する調整モジュールから構成される．学習には IoU 損失を用い，評価指標として mean Intersection over Union（mIoU）[30] を採用する．モデルの詳細は付録に示す．

4　実験結果と分析

本節では，提案する二段階フレームワークに基づき，(1) レイアウト構造推定と，(2) レイアウト木に基づく座標生成の結果を示す．

4.1　レイアウト構造推定

（Main）主結果　表 1 に，レイアウト構造推定の TED を示す．ここで用いた Random および kNN ベースライン（Base）は，学習による系列生成を行わず，各コマの特徴空間（テキスト特徴または視覚特徴）における近傍探索に基づいて，データセット中の既存レイアウト木を直接参照する手法である．視覚特徴のみを用いたモデル（Visual-only）は，

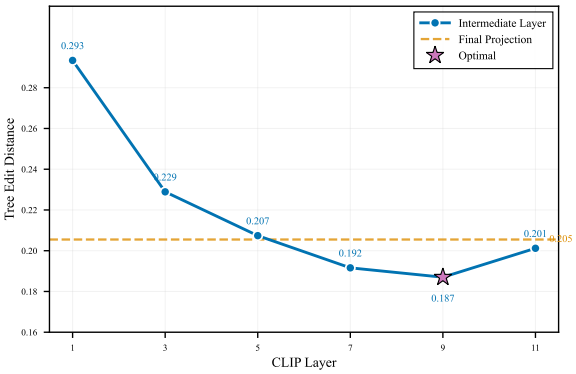

図 3　CLIP ViT の層別特徴を用いた TED の比較.

Random や kNN ベースラインを大きく上回る性能を示した一方，テキスト特徴のみ（Text-only）では改善が限定的であった．この結果は，本設定において，ページレベルの分割構造が，コマキャプションに含まれる語義的内容よりも，コマ画像の視覚的構図に強く依存して決定されている可能性を示唆する．

上記結果の解釈を検証するため，レイアウト構造推定が語義情報に依存しているか否かを切り分ける 3 種類の追加分析を行った（表 1）．

(A) キャプション劣化　キャプション生成時の指示を簡略化し，要約に関する制約を設けない劣化版キャプション (weak prompt) を用いた場合でも，TED は元の設定とほぼ同一であった．これは，情報量の違いが性能に影響しないことを示しており，キャプションがレイアウト構造推定において決定的な手掛かりとして機能していない可能性を示唆する．簡略したプロンプトの詳細は付録に示す．

(B) 視覚特徴の語義バイアス　主結果における視覚特徴の優位性が，語義情報の保持能力に起因するものか，あるいはレイアウト構造自体が語義情報に依存しないためかを検証するため，CLIP ViT-B/32 の層別特徴に加え，CLIP ViT-L/14，MAE ViT-B/16，MAE ViT-L/16 を用いた比較分析を行った．その結果，CLIP ViT-B/32 では中間層が最も低い TED を達成し，語義情報が強く統合される最終層では性能が低下した（図 3）．また，画像再構成を目的とする自己教師あり学習に基づく MAE 特徴 [31] においても，CLIP と同等，あるいはそれ以上の性能が得られた．これらの結果は，レイアウト構造推定が「何が描かれているか」という意味内容よりも，視覚的構図や空間的配置といった構造的情報に強く依存して決定されることを示唆している．

(C) 周波数分解　視覚特徴の周波数特性を検証するため，各コマ画像に 2 次元フーリエ変換（2D

図 4　元画像.

図 5　低周波.

図 6　高周波.

FFT）を適用し，周波数領域で成分を分離した．低周波成分（大域的構図，**Low freq.**）では振幅スペクトルの下位 10% を保持することでページ全体の構図や領域配置を残し，高周波成分（局所的形状・エッジ，**High freq.**）では上位 10% のみを保持することで輪郭や細部の形状を強調した（図 4–6）．その結果，いずれの周波数成分を用いた場合も元の視覚特徴より高い精度が得られたが，特に低周波成分のみを用いた条件で最も高い性能が確認された．この結果は，レイアウト木の推定において，局所的な形状やエッジよりも，コマ間の大域的な構図関係がより重要な役割を果たしていることを示唆している．

小括　以上より，レイアウト構造推定においては，テキスト内容よりもコマの視覚的構図に由来する情報が支配的な役割を果たす可能性が示された．

4.2　レイアウト座標生成

図 7 に，レイアウト木を用いた座標生成の結果を示す．ここでは，テキスト特徴のみ (Text-only)，視覚特徴のみ (Visual-only) を用いた条件に加え，レイアウト木を中間表現として導入した場合の効果を比較する．なお，具体的な座標生成結果の可視化例および失敗例に関する分析は，付録に示す．

レイアウト木の効果

正解レイアウト木（GT Tree）を用いた場合，Text-only / Visual-only 条件と比べて mIoU が大きく向上した．この結果は，レイアウト木がページ全体の大域的な配置関係を規定する中核的な構造表現として機能していることを示している．GT Tree + Visual が最良であり，視覚特徴が木構造では表現しきれない局所的な境界位置やサイズ調整を補完している．

予測レイアウト木の使用

推論時に予測されたレイアウト木（Pred. Tree）を用いた場合でも，Visual-only を上回る性能が得られた．これは，構造推定に誤差が含まれる場合でも，レイアウト木がページ構成の骨格として有効に機能しうることを示している．さらに視覚特徴を併用することで性能が回復しており，視覚情報が構造誤差の影響を部分的に緩和できることが分かる．

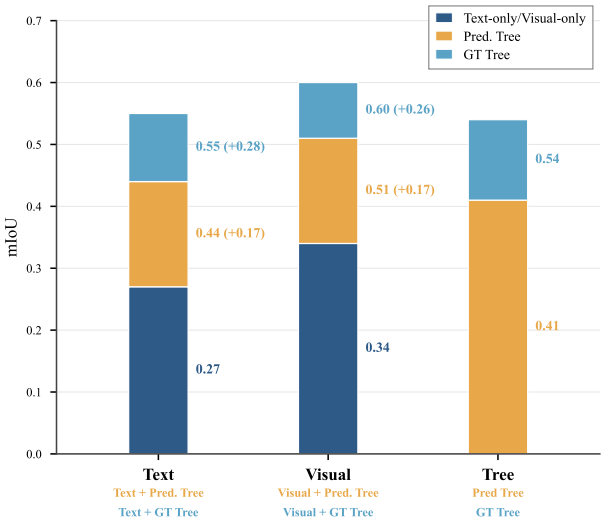

図 7　レイアウト木あり/なしの座標生成結果.

4.3　ページ生成への適用

推定されたレイアウト木に基づいてコマ配置を行うことで，ページ全体のバランスを保ったマンガページが生成された．詳細な生成例は付録に示す．

5　おわりに

本研究では，マンガページのレイアウト構造を明示的に推定し，それに基づいて各コマの配置を決定する二段階アプローチを提案した．特に，レイアウト構造推定と座標生成を分けることで，ページ全体の構造とコマの配置に対する柔軟な調整を可能にした．実験の結果，レイアウト構造はキャプションの語義的内容よりも，視覚的構図に強く依存することが確認された．レイアウト木を中間表現として使用することで，ページ構成の大域的な骨格を安定的に生成できることが分かった．また，予測されたレイアウト木に基づく座標生成が，視覚特徴との組み合わせにより改善されることも確認された．今後の課題としては，より多様なマンガページに対する適用や，予測精度の向上，画像生成の質を高めるための方法の検討が挙げられる．このアプローチは，マンガ自動生成や編集支援システムなど，今後の応用に向けて大きな可能性を秘めている．

謝辞

本研究は JSPS 科研費 JP24K02993 助成を受けたものです.

参考文献

[1] Scott McCloud and Mark Martin. **Understanding comics: The invisible art**, volume 106. Kitchen sink press Northampton, MA, 1993.

[2] Neil Cohn. **The Visual Language of Comics: Introduction to the Structure and Cognition of Sequential Images**. Bloomsbury, 2013.

[3] Neil Cohn. Navigating comics: An empirical and theoretical approach to strategies of reading comic page layouts. **Frontiers in psychology**, 4:186, 2013.

[4] Gunther Kress and Theo Van Leeuwen. **Reading images: The grammar of visual design**. Routledge, 2020.

[5] Clare Kirtley, Christopher Murray, Phillip B Vaughan, and Benjamin W Tatler. Navigating the narrative: An eye-tracking study of readers' strategies when reading comic page layouts. **Applied Cognitive Psychology**, 37(1):52–70, 2023.

[6] Kai Mikkonen and Olli Philippe Lautenbacher. Global attention in reading comics: Eye movement indications of interplay between narrative content and layout. **ImageTexT**, 10(2), 2019.

[7] Siyuan Feng. How panel layouts define manga: Insights from visual ablation experiments. In **Proceedings of the 47th Annual Meeting of the Cognitive Science Society (CogSci 2025)**, Rotterdam, Netherlands, 2025. Cognitive Science Society.

[8] Azuma Fujimoto, Toru Ogawa, Kazuyoshi Yamamoto, Yusuke Matsui, Toshihiko Yamasaki, and Kiyoharu Aizawa. Manga109 dataset and creation of metadata. In **Proceedings of the 1st international workshop on comics analysis, processing and understanding**, pages 1–5, 2016.

[9] Kiyoharu Aizawa, Azuma Fujimoto, Atsushi Otsubo, Toru Ogawa, Yusuke Matsui, Koki Tsubota, and Hikaru Ikuta. Building a manga dataset "manga109" with annotations for multimedia applications. **IEEE MultiMedia**, 27(2):8–18, 2020.

[10] David Dubray and Jochen Laubrock. Deep cnn-based speech balloon detection and segmentation for comic books. **CoRR**, abs/1902.08137, 2019.

[11] Christophe Rigaud, Nhu-Van Nguyen, and Jean-Christophe Burie. Text block segmentation in comic speech bubbles. In **Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part VI**, pages 250–261. Springer-Verlag, 2021.

[12] Ragav Sachdeva and Andrew Zisserman. The manga whisperer: Automatically generating transcriptions for comics. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pages 12967–12976, 2024.

[13] Yonggang Li, Yafeng Zhou, Yongtao Wang, Xiaoran Qin, and Zhi Tang. Dual loss for manga character recognition with imbalanced training data. In **2020 25th International Conference on Pattern Recognition (ICPR)**, pages 2166–2171. IEEE, 2021.

[14] Yingxuan Li, Ryota Hinami, Kiyoharu Aizawa, and Yusuke Matsui. Zero-shot character identification and speaker prediction in comics via iterative multimodal fusion. In **Proceedings of the 32nd ACM International Conference on Multimedia**, pages 7366–7374, 2024.

[15] Ragav Sachdeva and Andrew Zisserman. From panels to prose: Generating literary narratives from comics. **arXiv preprint arXiv:2503.23344**, 2025.

[16] Kamal Gupta, Justin Lazarow, Alessandro Achille, Larry S Davis, Vijay Mahadevan, and Abhinav Shrivastava. Layouttransformer: Layout generation and completion with self-attention. In **Proceedings of the IEEE/CVF International Conference on Computer Vision**, pages 1004–1014, 2021.

[17] Haruka Takahashi and Shigeru Kuriyama. Text2layout: Layout generation from text representation using transformer. **IEEE Access**, 2024.

[18] Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. **Advances in Neural Information Processing Systems**, 36:18225–18250, 2023.

[19] Tao Yang, Yingmin Luo, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. Posterllava: Constructing a unified multi-modal layout generator with llm. **arXiv preprint arXiv:2406.02884**, 2024.

[20] HsiaoYuan Hsu and Yuxin Peng. Postero: Structuring layout trees to enable language models in generalized content-aware layout generation. In **Proceedings of the Computer Vision and Pattern Recognition Conference**, pages 8117–8127, 2025.

[21] Siyu Chen, Dengjie Li, Zenghao Bao, Yao Zhou, Lingfeng Tan, Yujie Zhong, and Zheng Zhao. Manga generation via layout-controllable diffusion. **arXiv preprint arXiv:2412.19303**, 2024.

[22] Jianzong Wu, Chao Tang, Jingbo Wang, Yanhong Zeng, Xiangtai Li, and Yunhai Tong. Diffsensei: Bridging multi-modal llms and diffusion models for customized manga generation. In **Proceedings of the Computer Vision and Pattern Recognition Conference**, pages 28684–28693, 2025.

[23] Ying Cao, Antoni B Chan, and Rynson WH Lau. Automatic stylistic manga layout. **ACM Transactions on Graphics (TOG)**, 31(6):1–10, 2012.

[24] Kiyoharu Aizawa, Azuma Fujimoto, Atsushi Otsubo, Toru Ogawa, Yusuke Matsui, Koki Tsubota, and Hikaru Ikuta. Building a manga dataset "manga109" with annotations for multimedia applications. **IEEE MultiMedia**, 27(2):8–18, 2020.

[25] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. **Multimedia Tools and Applications**, 76(20):21811–21838, 2017.

[26] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**, pages 26296–26306, 2024.

[27] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. **Advances in neural information processing systems**, 33:5776–5788, 2020.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In **International conference on machine learning**, pages 8748–8763. PmLR, 2021.

[29] Kaizhong Zhang and Dennis Shasha. Simple fast algorithms for the editing distance between trees and related problems. **SIAM journal on computing**, 18(6):1245–1262, 1989.

[30] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. **International journal of computer vision**, 88(2):303–338, 2010.

[31] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**, pages 16000–16009, 2022.

[32] 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. **arXiv preprint arXiv:2307.01952**, 2023.

A　付録

レイアウト構造推定モデルの実装詳細　各コマ特徴は線形射影により 256 次元へ写像し，2 層の BiLSTM（単方向隠れ状態 256 次元）に入力する．デコーダには 4 層・8 ヘッドの Transformer Decoder を用いた．推論時には greedy decoding を採用し，有限状態機械に基づく hard mask により括弧整合性および葉ノード数一致を強制した．最適化には AdamW を用い，開発集合における正規化 Tree Edit Distance を基準として早期終了を行った．

レイアウト座標生成モデルの実装詳細　構造エンコーダでは，レイアウト木から得られる分割パス系列を 2 層 LSTM により埋め込み，構造特徴と統合した後，4 層 Transformer Encoder（$d_{\text{model}} = 256$, $n_{\text{head}} = 8$）に入力した．調整モジュールでは，2 層 Transformer Encoder を用いて補正量を推定し，tanh により調整幅を 0.3 に制限した．学習には AdamW を用い，IoU 損失を主目的関数とし，Smooth L1 損失および Residual 損失を加えた次の目的関数を最適化した：

$$\mathcal{L} = \lambda_{\text{iou}}\mathcal{L}_{\text{IoU}} + \lambda_{\text{l1}}\mathcal{L}_{\text{SmoothL1}} + \lambda_{\text{res}}\mathcal{L}_{\text{residual}},$$

ここで $\lambda_{\text{iou}} = 1.0$，$\lambda_{\text{l1}} = 0.5$，$\lambda_{\text{res}} = 0.5$ とした．

キャプション生成および劣化設定　各コマ画像に対して LLaVA-v1.6-34B を用い，以下の 2 種類のプロンプトにより英語キャプションを生成した．

通常プロンプト:

“Describe the scene in one concise English sentence. Focus on the actions of the characters, their surroundings, and the overall atmosphere, without unnecessary details.”

劣化プロンプト:

“Describe this panel from a Japanese manga.”

レイアウト木に基づく座標生成の挙動分析　図 8 は，レイアウト木を中間表現として用いた場合，予測された座標が，木構造に基づく分割関係に従って配置されることを示している．レイアウト木を用いない条件ではコマ間の配置関係が崩れるのに対し，レイアウト木を用いた条件では，ページ全体の構造と整合した配置が得られている．一方で図 9 に示すように，レイアウト構造推定に誤差が生じた場合には，座標生成モデルは予測された木構造に忠実に従うため，正解配置とのずれが生じることがある．

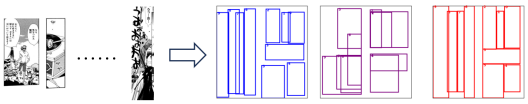

図 8　正解，レイアウト木を用いない条件とレイアウト木を中間表現として用いた条件の予測座標例．（『学園ノイズ』© 猪原大介）

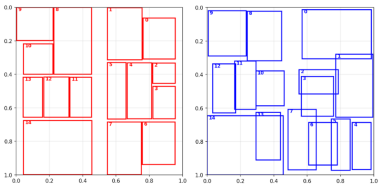

図 9　真のレイアウト木を用いた条件と予測レイアウト木を用いた条件の比較．

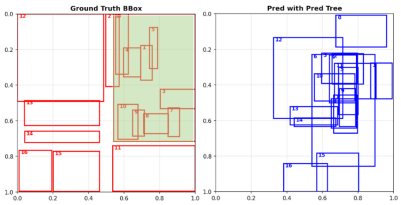

図 10　失敗例．左：正解座標，右：予測座標．緑色で示した領域は，1 つの大きなコマの内部に複数の小コマが配置される **inset panel** 構造に対応する．

図 11　画像生成モデルにより生成された複数のコマ画像を入力とし，提案手法によって単一ページとして再構成した例．

失敗例の分析　1 つの大きなコマの内部に，複数の小コマが配置される構造は **inset panel** 構造と呼ばれる．図 10 には，この **inset panel** 構造において配置の破綻が生じている．現行のレイアウト木表現では，同一の親ノード下に属するコマが同一のレイアウトパスを持つ．そのため，大コマと小コマの包含関係や相対的なサイズ差を表現できず，座標生成段階において構図上の整合性が失われている．

画像生成モデルとの統合によるページ生成　提案手法の最終的な適用例として，テキストから生成されたコマ画像を入力とし，マンガページを再構成するデモを実施した．物語を ChatGPT により場面分割し，各場面の要約文を，Stable Diffusion XL [32] を基盤とするマンガ風画像生成モデル BluePencilXL 3.10 に入力することで，各コマ画像を独立に生成した．得られたコマ画像列を提案したモデルに入力し，ページとして再構成した結果を図 11 に示す．