# The Manga Whisperer: Making Comics Accessible to Everyone

Ragav Sachdeva  Andrew Zisserman

Visual Geometry Group, Dept. of Engineering Science, University of Oxford

## Abstract

*In recent decades, Japanese comics, known as Manga, have gained global popularity, yet their visual nature presents challenges for individuals with visual impairments. In this study, we aim to break down this barrier to make manga accessible to everyone. Our focus is on diarisation i.e. automatically generating a dialogue transcript for manga. For code, datasets, and pre-trained model, visit:* https://github.com/ragavsachdeva/magi.

## 1. Introduction

Automatically understanding and describing manga to visually impaired individuals is highly challenging due to varying character depictions, diverse viewpoints, occlusions from speech balloons, non-human characters, and uncertain text placement. Additional complications arise from artistic layout, visual effects, and resolution. While humans rely on context and deductive reasoning to comprehend such complexity, machines face significant hurdles in this endeavor.

In this work, our objective is *diarisation* – to be able to generate a transcription, page by page, of who said what in the veridical order, to convey the story on that page. This involves solving a series of problems including panel detection, panel ordering, text detection, OCR, character detection, character identification and text-to-speaker association. To this end, we develop a model, *Magi*, that addresses these challenges using a unified architecture. Additionally, we create a challenging evaluation benchmark, called *PopManga*, comprising of manga pages (with annotations for bounding boxes for characters, texts and panels, character clustering labels and text to speaker matches) from 80+ popular manga by various artists known for their complexity and detailed story-telling, and demonstrate superiority of our method over existing solutions.

## 2. Detection and Association

Given a manga page, our goal is to produce a transcript of who said what and when in a fully automatic way. Therefore, as a requirement, the model must be aware of the various components that constitute a manga page and how they are related. We formulate this as a graph generation problem and propose a unified model, *Magi*, that is able to simultaneously detect panels, text blocks and characters (nodes of the graph), and perform character-character matching and text-character matching (edges of the graph). The architecture is illustrated in Figure 1.

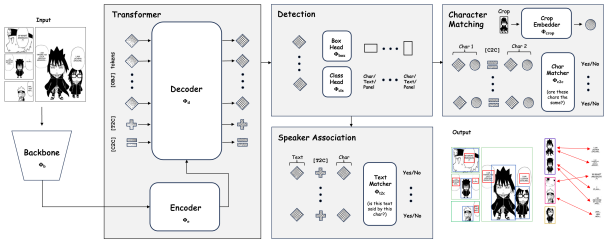


**Figure 1. The *Magi* Architecture.** Given a manga page as input, our model predicts bounding boxes for panels, text blocks and characters, and and associates characters of the same identity, and text to the characters who speak it.

Our model ingests a high resolution manga page as input and (1) predicts bounding boxes for panels, text blocks and characters, and (2) associates the detected (a) character-character pairs, where a positive association between two character boxes implies that they are the same character, and (b) text-character pairs, where a positive association between a text box and character box implies that this text is said by this character. The image is first processed by a CNN backbone, followed by a transformer encoder-decoder resulting in $N \times$ `[OBJ]` + `[C2C]` + `[T2C]` tokens. The `[OBJ]` tokens are processed by the detection heads (box and class) to obtain the bounding boxes and their classifications into one of *character*, *text*, *panel* or *background* class. The `[OBJ]` tokens corresponding to detected objects are then processed in pairs, along with `[C2C]` and `[T2C]`, by a character matching module and a speaker association module respectively resulting in character clusters and speaker information.

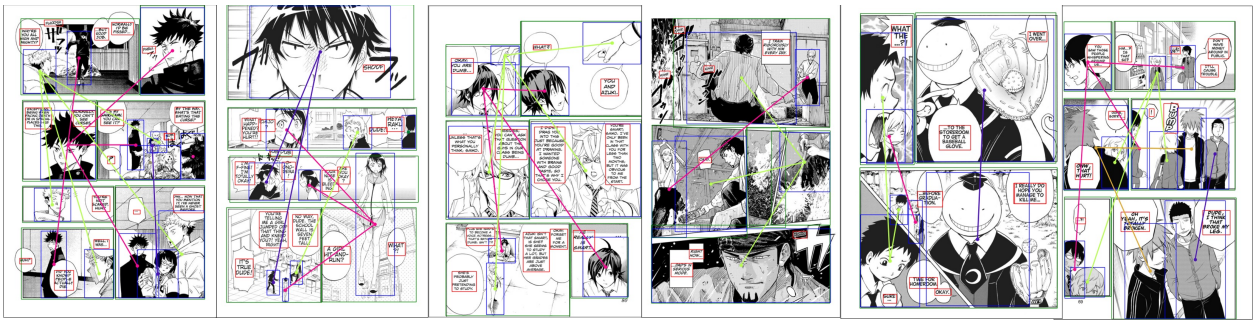

**Figure 2.** Bounding box predictions determined by the *Magi* model for characters, text blocks and panels, as well as clustering predictions (as nodes and edges). Best viewed digitally.

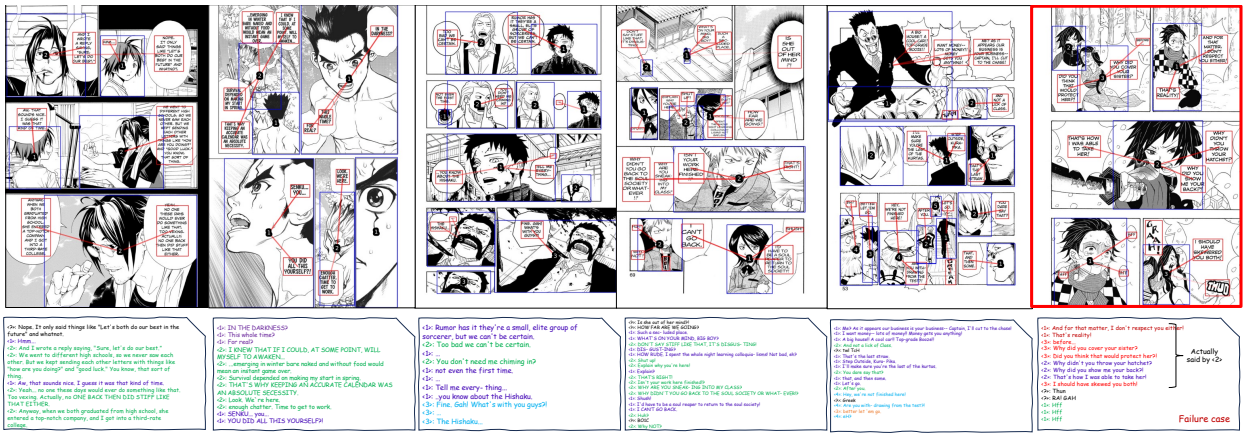

**Figure 3.** Transcripts generated by the *Magi* model. Each predicted text box is associated to a predicted character box using a line. The opacity of the line reflects the confidence of the model (the darker the line, the more confident the model is). Each predicted character box has a number at its centre based on the clustering predictions. Note that all the dialogues are in the correct reading order. For text to speaker predictions that have a low confidence score ($< 0.4$) we replace the predicted speaker with $\langle ? \rangle$ in the generated transcript and let the reader infer it from context. Best viewed digitally.

## 3. Transcript Generation

Once the bounding boxes for panels, characters and text boxes have been extracted, along with the character clusters and speaker associations, generating a transcript from them is really just an OCR + Sorting problem. To sort the text boxes into their reading order, we leverage the fact that manga pages are read from top to bottom, *right to left*. Given this, we order the text boxes in two steps: (a) order the panels to give the relative ordering of text boxes belonging to different panels, (b) order the text boxes within each panel. After ordering the text boxes, we perform OCR to extract the content of the texts and finally generate the transcript using all the computed data. Please refer to the arXiv version of the paper for more details.

## 4. Experiments

For training and evaluation we utilise three datasets: an existing dataset *Manga109*, and two new datasets that we introduce, *PopManga* and *Mangadex-1.5M*.

We show qualitative results in Figures 2 and 3. Quantitatively, (a) for *detection* our method attains an average precision of $\approx 0.85$ for characters, $\approx 0.92$ for texts and $\approx 0.93$ for panels, (b) for *character clustering* our method achieves an AMI of $\approx 0.65$ and MAP@R of $\approx 0.84$, and (c) for *speaker association* our method achieves Recall@#text of $\approx 0.84$. With these results, our method establishes itself as the state of the art. For more details regarding the experiments, results and comparisons with baselines, please refer to the arXiv version of the paper.

## 5. Conclusion

In this study, our primary objective was to improve the accessibility of manga for individuals with visual impairments. Tackling the complex task of diarisation, we have laid the groundwork for a fully automated transcription of manga content, enabling active engagement for everyone, irrespective of their visual abilities. Looking ahead, we anticipate leveraging the language understanding capabilities of Large Language Models (LLMs) to enhance diarisation by incorporating conversation history and context.