Tails Tell Tales: Chapter-Wide Manga Transcriptions with Character Names

Ragav Sachdeva, Gyungin Shin*, and Andrew Zisserman

Visual Geometry Group, Dept. of Engineering Science, University of Oxford
{rs,gyungin,az}@robots.ox.ac.uk

Abstract. Enabling engagement of manga by visually impaired individuals presents a significant challenge due to its inherently visual nature. With the goal of fostering accessibility, this paper aims to generate a dialogue transcript of a complete manga chapter, entirely automatically, with a particular emphasis on ensuring narrative consistency. This entails identifying (i) *what* is being said, i.e., detecting the texts on each page and classifying them into essential vs non-essential, and (ii) *who* is saying it, i.e., attributing each dialogue to its speaker, while ensuring the same characters are named consistently throughout the chapter.

To this end, we introduce: (i) Magiv2, a model that is capable of generating high-quality chapter-wide manga transcripts with named characters and significantly higher precision in speaker diarisation over prior works; (ii) an extension of the PopManga evaluation dataset, which now includes annotations for speech-bubble tail boxes, associations of text to corresponding tails, classifications of text as essential or non-essential, and the identity for each character box; and (iii) a new character bank dataset, which comprises over 11K characters from 76 manga series, featuring 11.5K exemplar character images in total, as well as a list of chapters in which they appear. The code, trained model, and both datasets can be found at: https://github.com/ragavsachdeva/magi

1 Introduction

Manga, a Japanese form of comic art, is celebrated globally for its rich narratives and distinctive graphical style. It engages millions of readers through its compelling visuals and complex character development. However, this visually dependent medium poses significant accessibility challenges for people with visual impairments (PVI). Recent advances in computer vision and machine learning present an opportunity to bridge this gap.

Despite the potential, there has been limited research in the field of improving manga accessibility for visually impaired readers. Our previous work [44] addressed the problem of automatically generating transcriptions for manga images. The contributions include the development of Magi, a model capable of

* Core contribution

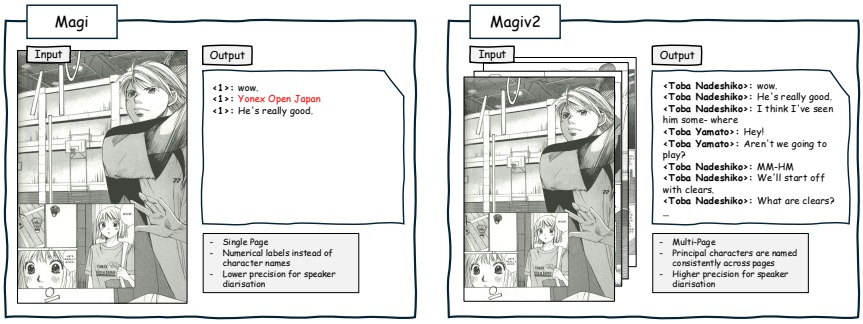

Fig. 1: **(Left)** Magi [44] generates a page-level transcript, with non-essential texts and without character names. **(Right)** Magiv2 (ours) generates chapter-wide transcripts with principal characters consistently named across pages, higher precision for speaker diarisation and only dialogue-essential texts. Manga Pages: ©Saki Kaori

processing a high-resolution manga page to detect characters, texts, and panels, as well as predict character clusters and associate dialogues to their respective speakers. Additionally, it introduced two new datasets, Mangadex-1.5M and PopManga, for training and evaluation purposes.

While the Magi model represents a promising initial step, it remains far from being practically usable due to several critical limitations. First, a major shortcoming is its failure to incorporate character names within the generated transcripts, instead denoting different characters with numerical labels such as 1, 2, 3, etc. As the transcripts are generated on a page-by-page basis, this approach inevitably leads to inconsistent character numbering across different pages. To improve the readability, it is essential to generate chapter-wide transcripts with consistent character names, since numerical labels are non-intuitive and make the transcripts difficult to follow. Second, Magi struggles to reliably associate text with the appropriate speaker, often attributing dialogues to the wrong characters. This misattribution disrupts the flow of conversation, leading to a disjointed and confusing narrative. Improving the association of text with the correct speaker is crucial to maintain the coherence of the dialogue and prevent reader confusion. Third, it is inept at distinguishing between essential and non-essential texts for the dialogue. Non-essential text, such as scene descriptions (e.g., street signs, graffiti, product labels) and sound effects (e.g., "Thud," "Whoosh"), should not be attributed to any character and, if improperly included as dialogues in the transcript, can disrupt the narrative flow.

To address these three limitations, we propose Magiv2, a robust and enhanced model capable of generating chapter-wide manga transcripts with consistent character names. fig. 1 shows the comparison of our model with Magi [44]. Our approach is built upon several key insights. First, recognising the challenges of character naming in manga transcripts—both providing names and ensuring

their consistency—Magiv2 leverages a character bank featuring names and images of principal characters and utilises a training-free, constraint optimisation method to consistently name all characters across the entire chapter, significantly outperforming traditional clustering-based methods. Second, we note that speech-bubble tails are crucial visual cues, intended by manga artists to indicate who is speaking. By making our model tail-aware, we significantly enhance text-to-speaker association performance. Third, we observe that distinguishing between essential and non-essential texts can largely be accomplished visually due to differences in font styles and the placement of texts. We leverage this prior and introduce a lightweight text-classification head on top of our visual backbone that can effectively differentiate dialogues from non-dialogue texts.

In summary, we make the following contributions: (i) We introduce a state-of-the-art model, Magiv2, capable of generating comprehensive manga transcripts across entire chapters, complete with character names and enhanced speaker associations. (ii) We extend the PopManga evaluation dataset by incorporating annotations for character names, speech bubble tails, text-to-tail associations, and text classification. This dataset is used to evaluate the performance of the new capabilities of Magiv2, such as character recognition. (iii) We release a new, meticulously curated character bank dataset for 76 manga series, encompassing more than 11K principal characters, with 11.5K exemplar character images in total. Additionally, this dataset includes detailed metadata such as the series names and specific chapters where each character appears.

With these contributions, over 10,000 published manga chapters (from series in PopManga) can now be transcribed directly using our model, with the potential for more as the character bank dataset grows.

2 Related Work

Comic understanding. Using computer vision to analyse and understand comic books has been extensively explored. Several datasets have been contributed to facilitate this research including Manga109 [1, 3, 22], DCM [33], eBDtheque [10], PopManga [44] etc. There are several existing works that propose solutions for panel detection [12, 34–36, 41, 44, 50], text/speech balloon detection [3, 34, 35, 37, 44], depth-estimation [5], character detection [16, 17, 35, 44, 48], character re-identification, [38, 44, 46, 49, 52], speaker identification, [41, 44], captioning [39], and transcript generation [44].

Person identification using a character bank. Identifying and naming people in images or videos has been a long studied research problem. Often times this either requires complex reasoning, e.g. inferring the name of a person based on how other people address them, or prior context and memory, e.g. a person may have been introduced previously and this information needs to be remembered. Given the complexity of this task, a common approach is to rely on an external character bank which can be used for matching query character images with a gallery [4, 6, 14, 15, 24, 32, 51].

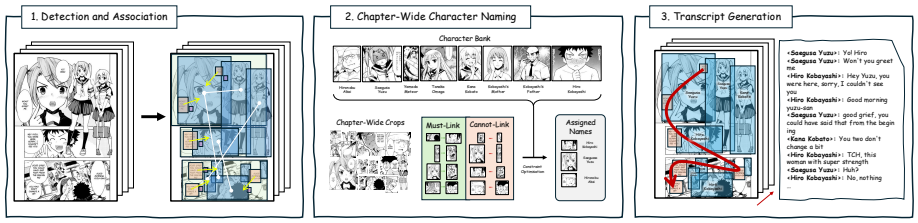

Fig. 2: **Inference pipeline.** Given a manga chapter, along with a character bank: (1) each page is processed independently to detect various elements and their relationships, such as character and text boxes, and their association. Next, (2) using the character bank, names are assigned to detected character crops across all pages using a constraint optimisation approach. Finally, (3) the transcript is generated by performing OCR, ordering all the texts and removing non-essential texts. Manga Pages: ©Takuji.

Comic Accessibility for PVI users. Several efforts have been made to understand the challenges faced by PVI when accessing comics [20, 40, 45] and solutions have been proposed in the form of tactile books [31], textured images [8], audiobooks [47] etc. Recent works have also explored the use of computer vision and machine learning to automatically caption simple comic strips [39] and generate dialogue transcripts of more complex ones [44].
We compare with prior works that are available publicly.

3 Overview of the Chapter-Wide Inference

Given a manga chapter, our goal is to generate a transcript of all pages while ensuring narrative consistency. However, processing all pages simultaneously to generate a dialogue transcript in a single forward pass is computationally prohibitive (a typical manga chapter comprises 15 to 30 pages), necessitating a segmented approach. To mitigate the computational burden and efficiently generate a chapter-wide transcript with accurate speaker attribution, we employ the following three-step process. The complete inference pipeline is shown in fig. 2.

1. Detection and Association. The initial step involves processing each manga page independently, framed as a graph generation problem. This step is similar to [44], but with modifications to incorporate additional elements. In our graph, "nodes" represent bounding boxes of detected characters, texts, panels, and notably, speech bubble tails. The "edges" represent pairwise relationships between character-character, text-character, and text-tail. More details regarding the architecture and training strategy are described in section 4.

2. Chapter-Wide Character Naming. Given the crops of detected characters from all pages of a manga chapter, along with a character bank comprising images and names of principal characters, the goal is to assign each character crop to the

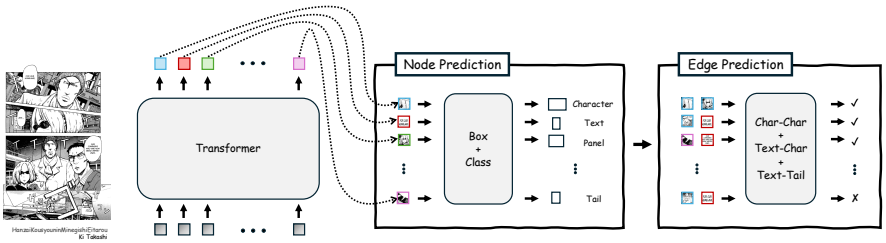

Fig. 3: **Simplified Detection and Association Architecture**. The input to the model is an RGB image of a manga page. The transformer decoder outputs several feature vectors which are used to predict bounding boxes for characters, texts, panels and tails ("nodes"). These features are further processed in pairs to predict character-character, text-character and text-tail associations ("edges").

correct principal character in the character bank, if one exists, otherwise assign it to the "other" class. This step simultaneously allows naming of speakers in the final transcripts, and consistent character identification across pages, if the assignments are correct. We formulate this as a constraint optimisation approach and provide more details in section 5.

3. Transcript Generation. Finally, the gathered information is compiled to generate the chapter-wide transcript. This is a four step process: First, the detected text boxes are filtered such that the texts that are classified as non-essential are removed; Second, the remaining text boxes are organised in their reading order (by first sorting the pages, then sorting panels on each page [44] and finally sorting texts within each panel [13]); Third, Optical Character Recognition (OCR) is used to extract texts from the manga pages; Finally, the transcript is generated by utilising the text-character associations predicted in section 4 and character names predicted in section 5. We provide more details in supp. mat.

4 Detection and Association

Given a manga page, the objective of this section is to detect the various components that constitute a manga page—particularly the panels, characters, text blocks and tails (i.e., localise where they are on the page), and also to associate them: character-character association (i.e., character clustering), text-character association (i.e., speaker diarisation), and text-tail association. We cast this as generating a graph, as described in step 1 of the inference process of section 3. The model architecture for this detection and association is illustrated in fig. 3.

Briefly, the model ingests a high resolution manga page as input, which is first processed by a CNN backbone, followed by a transformer encoder-decoder resulting in N object feature vectors. These feature vectors are processed by the detection head to regress a bounding box and classify it into character, text, panel, tail or background. This completes the *nodes* part of the graph.

To generate text-character *edges*, the features corresponding to detected text boxes and character boxes are processed in pairs by a speaker association head to make a binary prediction (whether the edge exists or not). Similarly, a tail association head processes pairs of text and tail feature vectors resulting in text-tail *edges*. Character-character *edges* are obtained by processing pairs of detected character feature vectors along with their respective crop embedding feature vectors (obtained by a separate crop embedding module). Finally, the feature vectors corresponding to detected texts are processed by a linear layer to classify them into essential vs non-essential. Further details regarding the model architecture and implementation are provided in section 4.2.

4.1 Semi-Supervised Training

A significant challenge in training the graph generation model is the quality and completeness of the training data. We utilise two datasets: (i) Mangadex-1.5M, which is unlabelled, and (ii) PopManga (Dev), which is partially labelled[1].

To address the challenge of partially annotated training data, we approach the training of our graph generation model through semi-supervised learning. We begin by curating a small subset of the PopManga (Dev) set, and endow it with both tail-related annotations and text classification labels. Additionally, we extract *partial* labels for the Mangadex-1.5M dataset using Magi [44], which of course lack tail-related and text classification annotations. Given this combination of data—small subset with complete annotations and larger subsets with partial pseudo-annotations—we adopt the following training strategy.

Initially, we warm up our model by training it on the large-scale partially-labeled pseudo-annotations. We then fine-tune this model on the smaller, fully annotated dataset. Subsequently, this model is used to mine *complete*, but possibly noisy, annotations for the large-scale data. We then re-train our model *from scratch* on this newly annotated large-scale data, which remains pseudo-annotated but now more comprehensively annotated, followed by additional fine-tuning on the small, completely annotated data. This cycle is repeated multiple times to refine our model. Detailed training recipe is provided in the supp. mat.

This training approach ensures robust detection and association capabilities despite the incomplete initial annotations. This style of training paradigm is common in noisy label learning literature, where training the model on noisy but large scale data (in our case pseudo-annotations), followed by fine-tuning on clean data, and then re-mining the pseudo-annotations, results in improved training data, which in turn can be used to train a better model [2, 7, 21, 42, 43]. A crucial aspect of this methodology is the re-training of the model *from scratch* after each round of mining pseudo-annotations, which is essential to avoid confirmation bias and prevent the model from overfitting on its own predictions.

[1] All pages contain character boxes, text boxes, and character clusters, but only a subset of pages include text-to-character associations. Furthermore, none of the pages have annotations for speech bubble tails or labels for text classification.

4.2 Implementation

The graph generation model architecture consists of a ResNet50 [11] backbone, followed by an encoder-decoder transformer with 6 layers each, hidden dimension of 256, 8 attention heads and conditional cross-attention [30]. The crop-embedding module is an encoder-only transformer with 12 layers, hidden dimension of 768, and 12 attention heads. The text-character, text-tail and character-character edge prediction heads are all 3-layered MLPs, and the text-classification head is a simple linear layer. Our training objective for box prediction is the same as in [30]. We further apply Binary Cross Entropy loss to the outputs of our edge-prediction as well as text-classification heads. Additionally, we apply Supervised Contrastive Loss [18] to the per-page embeddings from the crop-embedding module. We trained our mode, on 2×A40 GPUs using AdamW [28] optimiser with both learning rate and weight decay of 0.0001, and batch size of 16.

5 Chapter-Wide Character Naming

The objective in this section to *assign* each detected character crop in the chapter to one of the characters in the character bank (introduced in section 6), unless they are "other". This is the second, chapter-wide, step of the inference process.

The question is, how to optimise this assignment objective? Naively, this can be accomplished greedily by computing the similarity of each crop with each character in the character bank and taking the `argmax`. However, we can do better, by leveraging additional *constraints* (must-link and cannot-link) from per-page associations. Specifically, the graph computed in section 4, provides us with character-character edges which can be transformed into must-link constraints, i.e. these crops are of the same characters and must be assigned the same identity, and cannot-link constraints, i.e. these crops are of different characters and must be assigned to different identities. These per-page must-link and cannot-link constraints provide a stronger signal than simple crop-based similarity as they factor in surrounding visual cues, e.g. two characters in the same panel are likely to be different characters, regardless of the visual similarity. This assignment problem can be formulated as a Mixed Integer Linear Programming problem, for which there are several existing solvers e.g. COIN-OR Branch and Cut Solver (CBC) [9, 29].

Problem Definition. Formally, suppose there are n character crops in a particular chapter and k characters in the character bank. Additionally, suppose that we have a set of must-link constraints M, representing pairs of crops that must be assigned to the same character, and a set of cannot-link constraints C, representing pairs of crops that must not be assigned to the same character. We further define a $(k+1)$th character, which is a dummy character to capture "other" when the crop is of a character that is not in the character bank.

Variable. Let x_{ij} be a binary variable that equals 1 if character crop i is assigned to character j in the character bank, and 0 otherwise.

Objective Function. The objective is to compute the optimal assignment of crops to characters in the character bank, i.e. computing x_{ij}, which is achieved by

$$\min_{x} \sum_{i=1}^{n} \sum_{j=1}^{k+1} d_{ij} x_{ij}, \quad \text{where} \quad d_{ij} = \begin{cases} \eta & \text{if } j = k+1, \\ \|e_i - e_j\| & \text{otherwise} \end{cases} \tag{1}$$

and e_i, e_j are embeddings for crop i and character j, respectively, and η is a fixed outlier-threshold hyperparameter (in practice, $\eta = 0.75$).
Constraints. The objective function above is minimised subject to the following constraints:

$$\sum_{j=1}^{k+1} x_{ij} = 1, \quad \forall i \in \{1, \ldots, n\} \tag{2}$$

$$x_{u,j} - x_{v,j} = 0, \quad \forall (u, v) \in M, \quad \forall j \in \{1, \ldots, k+1\} \tag{3}$$

$$x_{u,j} + x_{v,j} \leq 1, \quad \forall (u, v) \in C, \quad \forall j \in \{1, \ldots, k\} \tag{4}$$

where eq. (2) ensures that each crop is assigned to exactly one character, eq. (3) enforces the must-link constraints, and eq. (4) enforces the cannot-link constraints.

Note that in eq. (4), $j \neq k+1$. This is because there may be two different characters (hence *must not be* linked) that are not in the character bank (hence *must be* linked to "other"). In other words, a cannot link constraint between crops u, v is applied such that these crops must not be assigned to the same character, unless they are assigned to "other" i.e. $(k+1)$th character.
In section 7, we compare the proposed constraint optimisation approach with traditional clustering based approaches and demonstrate that our method significantly outperforms the baselines.

6 Datasets: PopCharacters and PopManga-X

The recently introduced PopManga [44] dataset provides annotations for character boxes, text boxes, per-page character clusters and speaker associations. It is divided into three splits: Dev, Test-S and Test-U, with around 2000 images of manga pages in Test, and S & U meaning that other chapters from the series are Seen or Unseen during training.

In this section we detail two data related contributions: (a) We compile a character bank of principal manga characters in PopManga. Please see fig. 4 for some examples and dataset statistics; (b) We extend the annotations of the PopManga test set to facilitate the evaluation of the new Magiv2 model capabilities, such as character labelling. Please see fig. 5 for an overview of the extended test dataset along with statistics on various types of available annotations.

PopCharacters. We introduce a new character bank dataset, called PopCharacters, comprising principal characters[2] in PopManga. For each principal char-

[2] We define principal characters as those who play crucial roles in the main story of a series. Please refer to the supplementary material for more details.

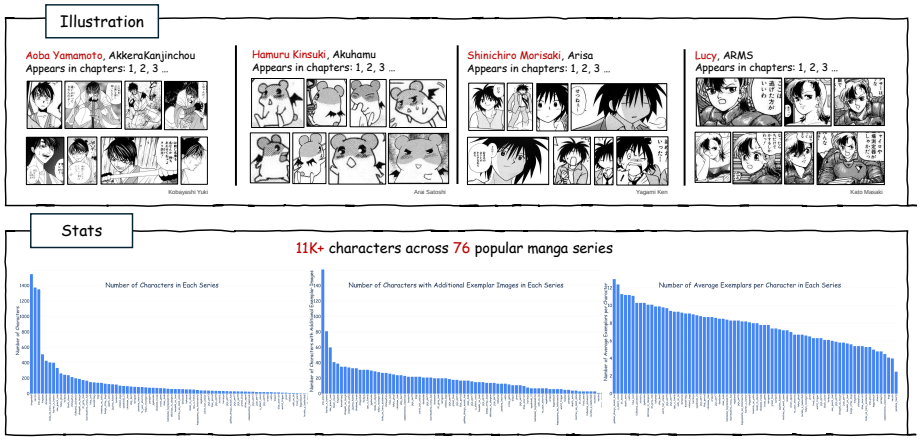

Fig. 4: **PopCharacters—the proposed character bank dataset. (Top)** We illustrate the type of data available in PopCharacters using four different character from Manga109 [1], including their name (in red), followed by the name of the series and the list of chapters they appear in. **(Bottom)** We display histograms showing the number of characters (left), the number of frequently occurring characters with additional exemplar images (middle), and the average number of exemplars per frequently occurring character (right) in each series for PopCharacters.

acter in PopCharacters, we provide (i) the character's name, (ii) a set of web-scraped thumbnail images of the character, and (iii) the series it belongs and a list of manga chapters the character appears in. Additionally, for a subset of the characters in PopCharacters, which appear far more frequently than others, we also provide a set of exemplar images queried from within the manga chapters and verified by human-in-the-loop. The purpose of this dataset is two-fold: (i) it enables Magiv2 to transcribe hundreds of thousands of manga pages (that are commercially available), with names for principal characters; and (ii) it provides a valuable resource for training models on tasks such as character recognition and character clustering. Further details on the dataset curation process is provided in the supp. mat.

PopManga-X. In the PopManga test splits (i.e., Test-S and Test-U), we provide new annotations for speech-bubble tail bounding boxes, associations of text boxes to tail boxes, and text categories (essential vs non-essential), providing the test-bed for tail-related predictions and text classification (see fig. 5). Moreover, we label each character box in the test splits with the name of the character (consistent with PopCharacters), thereby offering global character clusters across the series and permitting the evaluation of chapter-wide character cluster predictions. To differentiate this extended dataset with more types of annotations from the original, we call this PopManga-X (more in supp. mat.).

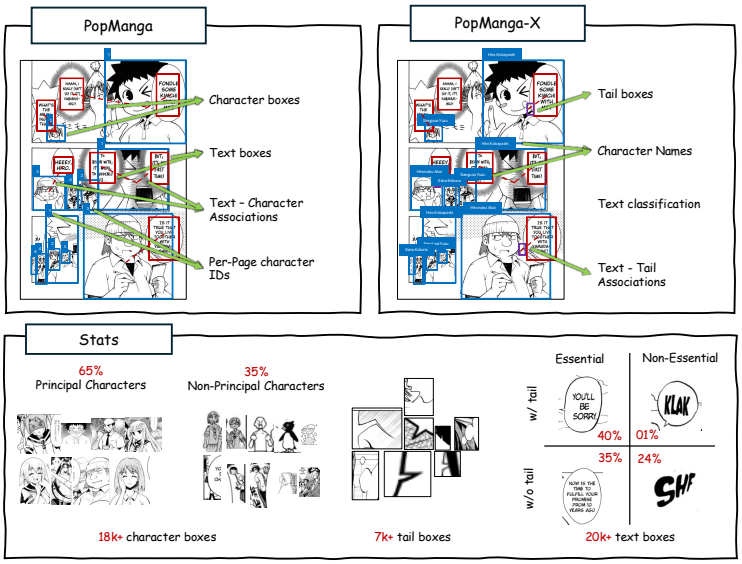

Fig. 5: **PopManga-X**. **(Top-Left)** Ground-truth annotations in PopManga. **(Top-Right)** Ground-truth annotations *added* to test splits of PopManga, now referred to as PopManga-X. **(Bottom)** Statistics on various elements of PopManga-X test splits. Manga images are from MeteoSan ©Takuji and only for illustration purposes.

7 Results

To achieve high-quality chapter-wide manga transcriptions, we identify three core tasks in section 3: (i) per-page detection and association, (ii) chapter-wide character naming, and (iii) transcript generation. In the following, we report our model's performance on the first two tasks and note that their quality directly determines the quality of the third. In other words, to evaluate the quality of the generated transcript, it is enough to evaluate the first two tasks only, as they reflect the correctness of the transcript. Qualitative results are shown in fig. 6.

7.1 Detection and Association/Graph Generation

Given a single manga page as input, here we evaluate the performance of the graph generation model in terms of (i) predicting panels, texts, characters, and tails, (ii) predicting text-character edges, text-tail edges, character-character edges, and (iii) classifying predicted texts into essential vs non-essential

Nodes: To measure the quality of predicted "nodes" i.e. predicted bounding boxes, we use the standard object detection evaluation measures. In table 1, we report the results for the average precision metric [23] for detecting character boxes, text boxes, tail boxes and panel boxes. We compare our results

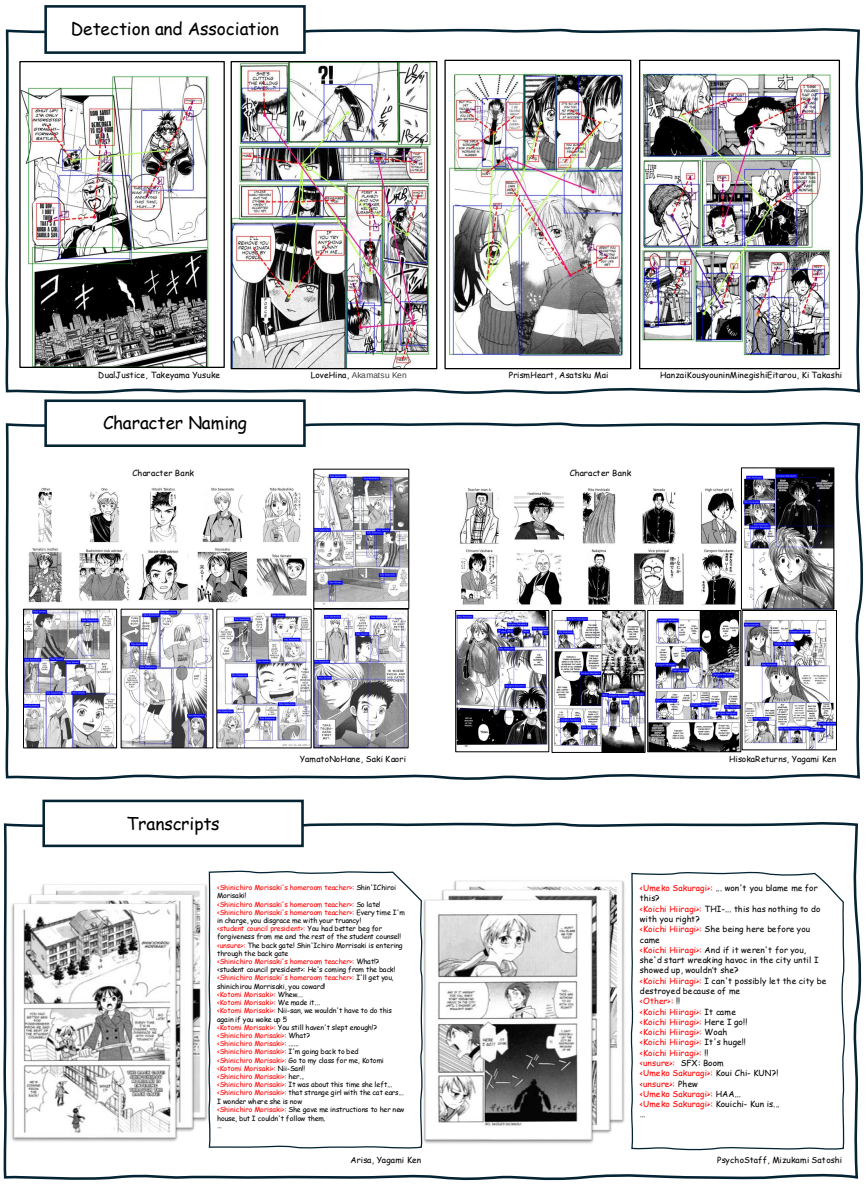

Fig. 6: **Qualitative Predictions. (Top)** We show the predictions from our graph generation model—panels (in green), character (in blue), texts classified as essential (in red), and speech bubble tails (in purple). The text-character edges (dashed red lines), text-tail edges (dashed purple lines) and character-character edges (unique colour for each connected component) are also shown. **(Middle)** We show the prediction for character names across multiple pages of two different manga series, demonstrating character naming consistency. **(Bottom)** We show the final, generated, multi-page transcripts using our method.

against, Magi [44], DASS [48] and zero-shot results from GroundingDino [26], on PopManga-X and Manga109 [1].

Edges: Our model is trained to predict three different kinds of edges: (i) character-character, (ii) text-character and, (iii) text-tail. To evaluate the character-character edges, we treat it as a per-page clustering problem and report the same metrics as [44], namely AMI, NMI, P@1, R-P, MRR and MAP@R, on PopManga-X and Manga109 [1], in table 3. For text-character and text-tail edges, we treat it as a binary classification problem (whether the predicted edge is correct or not) and report the average precision metric on PopManga-X in table 2. We compare our results against [44].

Table 1: **Detection Results.** We report the average precision results, which have an upper bound of 1.0.

	PopManga-X (Test-S)			PopManga-X (Test-U)			Manga109	
method	Char	Text	Tail	Char	Text	Tail	Body	Panel
DASS [48]	0.8410	-	-	0.8580	-	-	**0.9251**	-
Grounding-DINO [26]	0.7250	0.7922	-	0.7420	0.8301	-	0.7985	0.5131
Magi [44]	0.8485	0.9227	-	0.8615	0.9208	-	0.9015	0.9357
Magiv2 (Ours)	**0.8544**	**0.9372**	**0.8766**	**0.8720**	**0.9353**	**0.8737**	0.9046	**0.9405**

Text Classification: Finally, we evaluate the performance of our model on categorising the detected texts into essential vs non-essential. In this work, we classify a text as essential, if it is a spoken dialogue, interjection or an internal thought by a character, or context added by the narrator. Everything else is non-essential e.g. sound-effects, editorial footnotes, scene-texts etc. We report the average precision results on PopManga-X in table 2. We use the confidence scores from Magi [44] as a baseline.

Discussion. When compared with prior works, our model achieves (i) better per-page character clustering results, particularly in the crop-only setting, which we attribute to the semi-supervised learning training scheme, where mining better pseudo labels in turn improves the model's performance; (ii) significant improvement in speaker diarisation (i.e. text-character matching) results, which is largely attributed to the introduction of speech-bubble tails; (iii) comparable bounding box detection results. Furthermore, our work unlocks new functionality, not supported by prior works, including (i) detecting tail boxes, text-tail associations, and (ii) text classification into essential vs non-essential, which can be used to improve the quality of the generated transcripts.

Table 2: **Text-Related Results.** We report the average precision results, which have an upper bound of 1.0.

	Text - Character Association		Text - Tail Association		Text Classification	
method	PopManga-X (Test-S)	PopManga-X (Test-U)	PopManga-X (Test-S)	PopManga-X (Test-U)	PopManga-X (Test-S)	PopManga-X (Test-U)
Magi [44]	0.5248	0.5632	-	-	0.9617	0.9692
Magiv2 (Ours)	**0.7499**	**0.7512**	**0.9838**	**0.9830**	**0.9897**	**0.9914**

Table 3: **Per-Page Character Clustering Results.** We report results using several metrics. They all have an upper bound of 1.0.

method	AMI	NMI	MRR	MAP@R	P@1	R-P
PopManga-X (Test-S)						
Magi (crop only) [44]	0.4892	0.7178	0.9008	0.7840	0.8423	0.8008
Magiv2 (crop only) (Ours)	**0.5826**	**0.8120**	**0.9275**	**0.8401**	**0.8831**	**0.8526**
Magi [44]	0.6574	0.8501	0.9312	0.8439	0.8884	0.8555
Magiv2 (Ours)	**0.6745**	**0.8610**	**0.9431**	**0.8669**	**0.9066**	**0.8770**
PopManga-X (Test-U)						
Magi (crop only) [44]	0.4862	0.7326	0.9061	0.7926	0.8477	0.8076
Magiv2 (crop only) (Ours)	**0.5711**	**0.8108**	**0.9321**	**0.8491**	**0.8898**	**0.8598**
Magi [44]	0.6527	0.8503	0.9347	0.8557	0.8936	0.8656
Magiv2 (Ours)	**0.6650**	**0.8579**	**0.9508**	**0.8818**	**0.9202**	**0.8898**
Manga109 (Body)						
Magi (crop only) [44]	0.5690	0.7694	0.9237	0.8259	0.8721	0.8389
Magiv2 (crop only) (Ours)	**0.6204**	**0.8152**	**0.9400**	**0.8646**	**0.9002**	**0.8737**
Magi [44]	0.6345	0.8202	0.9383	0.8567	0.8966	0.8667
Magiv2 (Ours)	**0.6456**	**0.8336**	**0.9514**	**0.8812**	**0.9179**	**0.8895**

7.2 Chapter-Wide Character Naming/Character Identification

Here we evaluate the efficacy of our method in forming chapter-wide character clusters and evaluate whether the same characters across pages are assigned the same name. For evaluation, we utilise test splits of PopManga-X where the input this time is an entire manga chapter. There are 50 chapters in the two test sets in total. For each chapter, we curate a chapter-specific character bank from the PopCharacters dataset, comprising principal characters that appear in this chapter. This chapter-specific character bank consists of names of principal characters along with exactly 1 exemplar image per character. Furthermore, all non-principal characters in the chapter, i.e. characters for which we do not have a name and exemplar image, are grouped into a single "other" category. Given a manga chapter and chapter-specific character bank, we report the accuracy of character naming, in table 4.

Naive Baseline. A straightforward solution to the chapter-wide character naming problem is to formulate it as a clustering problem. Assuming that the number of characters in the character bank, k, is equal to the number of ground truth clusters in the chapter (which may not be true in practice), a simple approach is to compute embeddings for each character crop in the chapter and then cluster them into k clusters.

We take two approaches to implement clustering based baselines: (i) simple K-means clustering [27], with $k + 1$ clusters (an extra cluster for "other" characters, not in the character bank); and (ii) first filter out all "other" characters using outlier/anomaly detection [25] and then perform simple K-means clustering with k clusters. Once the clusters have been computed they are assigned to character names by using Hungarian matching [19] between the embeddings for cluster centres and exemplar images in the character bank.

Discussion. Compared to traditional clustering-based methods, our approach performs significantly better. Clustering methods have two key shortcomings: (i) visually similar but distinct characters are often grouped into the same cluster, affecting other cluster assignments since the number of clusters is fixed; (ii) these methods fail to use spatial cues, such as different characters in the same panel. Our proposed method is more robust to such shortcomings as evident by the superior performance. We also observe that using ground truth must-link and cannot-link constraints for each page, significantly improves the quality of the results. This finding has a very ***significant implication***—in the future, it is sufficient to improve the *per-page* model in order to improve the *chapter-wide* results. This is of great value because training a per-page model is much more tractable. A key ***limitation*** of this approach, however, is that it groups all non-principal characters into a single "other" category. It is not designed to disambiguate 'unnamed person 1' from 'unnamed person 2'. We leave this as future work.

Table 4: **Character Naming Results.** We report the accuracy results, which have an upper bound of 1.0.

embedding model	method	notes	PopManga-X (Test-S)	PopManga-X (Test-U)
Magi	K-means [27]	nclusters= $k+1$	0.3351	0.3820
Magiv2	K-means [27]	nclusters= $k+1$	0.3801	0.4223
Magi	iForest [25] + K-means [27]	nclusters= k	0.4549	0.4646
Magiv2	iForest [25] + K-means [27]	nclusters= k	0.5101	0.4942
Magi	Constraint Optimisation (Ours)	Predicted per-page constraints	0.6637	0.7058
Magiv2	Constraint Optimisation (Ours)	Predicted per-page constraints	**0.7273**	**0.7530**
Magi	Constraint Optimisation (Ours)	GT per-page constraints	0.7445	0.7975
Magiv2	Constraint Optimisation (Ours)	GT per-page constraints	0.7987	0.8526

8 Conclusion

We present a solution for generating chapter-wide manga transcriptions with consistent character names and clearer narrative, and contribute a new SOTA model, a training-free constraint optimisation approach to chapter-wide character naming, and new datasets to facilitate further research and comparisons. With these contributions it is now possible to transcribe over 10,000 manga chapters that are currently available commercially, complete with character names.

Acknowledgments. This research is supported by EPSRC Programme Grant VisualAI EP/T028572/1 and a Royal Society Research Professorship RP/R1/191132. This work was partially supported using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service, provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/T022159/1), and DiRAC funding from the Science and Technology Facilities Council. Gyungin Shin would like to thank Zheng Fang for the enormous support.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Aizawa, K., Fujimoto, A., Otsubo, A., Ogawa, T., Matsui, Y., Tsubota, K., Ikuta, H.: Building a manga dataset "manga109" with annotations for multimedia applications. IEEE MultiMedia **27**(2), 8–18 (2020). https://doi.org/10.1109/mmul.2020.2987895
2. Arazo, E., Ortego, D., Albert, P., O'Connor, N., McGuinness, K.: Unsupervised label noise modeling and loss correction. In: International conference on machine learning. pp. 312–321. PMLR (2019)
3. Baek, J., Matsui, Y., Aizawa, K.: Coo: Comic onomatopoeia dataset for recognizing arbitrary or truncated texts. In: European Conference on Computer Vision. pp. 267–283. Springer (2022)
4. Bain, M., Nagrani, A., Brown, A., Zisserman, A.: Condensed movies: Story based retrieval with contextual embeddings. In: Proceedings of the Asian Conference on Computer Vision (2020)
5. Bhattacharjee, D., Everaert, M., Salzmann, M., Süsstrunk, S.: Estimating image depth in the comics domain. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2070–2079 (2022)
6. Chen, Z., Feng, B., Ngo, C.W., Jia, C., Huang, X.: Improving automatic name-face association using celebrity images on the web. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. pp. 623–626 (2015)
7. Cordeiro, F.R., Sachdeva, R., Belagiannis, V., Reid, I., Carneiro, G.: Longremix: Robust learning with high confidence samples in a noisy label environment. Pattern recognition **133**, 109013 (2023)
8. Des livres à voir et à toucher. https://www.lavillebraille.fr/des-livres-a-voir-et-a-toucher/
9. Forrest, J., Lougee-Heimer, R.: Cbc (coin-or branch and cut). https://github.com/coin-or/Cbc, computational Infrastructure for Operations Research (COIN-OR)
10. Guérin, C., Rigaud, C., Mercier, A., Ammar-Boudjelal, F., Bertet, K., Bouju, A., Burie, J.C., Louis, G., Ogier, J.M., Revel, A.: ebdtheque: a representative database of comics. In: 2013 12th International Conference on Document Analysis and Recognition. pp. 1145–1149. IEEE (2013)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. He, Z., Zhou, Y., Wang, Y., Wang, S., Lu, X., Tang, Z., Cai, L.: An end-to-end quadrilateral regression network for comic panel extraction. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 887–895 (2018)
13. Hinami, R., Ishiwatari, S., Yasuda, K., Matsui, Y.: Towards fully automated manga translation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 12998–13008 (2021)
14. Hong, X., Sayeed, A., Mehra, K., Demberg, V., Schiele, B.: Visual writing prompts: Character-grounded story generation with curated image sequences. Transactions of the Association for Computational Linguistics **11**, 565–581 (2023)
15. Huang, Q., Xiong, Y., Rao, A., Wang, J., Lin, D.: Movienet: A holistic dataset for movie understanding. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. pp. 709–727. Springer (2020)

16. Inoue, N., Furuta, R., Yamasaki, T., Aizawa, K.: Cross-domain weakly-supervised object detection through progressive domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5001–5009 (2018)
17. Jiang, J., Chen, B., Wang, J., Long, M.: Decoupled adaptation for cross-domain object detection. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=VNqaB1g9393
18. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)
19. Kuhn, H.W.: The Hungarian Method for the Assignment Problem. Naval Research Logistics Quarterly (1955)
20. Lee, Y., Joh, H., Yoo, S., Oh, U.: Accesscomics: an accessible digital comic book reader for people with visual impairments. In: Proceedings of the 18th International Web for All Conference. pp. 1–11 (2021)
21. Li, J., Socher, R., Hoi, S.C.: Dividemix: Learning with noisy labels as semi-supervised learning. arXiv preprint arXiv:2002.07394 (2020)
22. Li, Y., Aizawa, K., Matsui, Y.: Manga109dialog a large-scale dialogue dataset for comics speaker detection. arXiv preprint arXiv:2306.17469 (2023)
23. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
24. Liu, D., Keller, F.: Detecting and grounding important characters in visual stories. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 13210–13218 (2023)
25. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data (TKDD) **6**(1), 1–39 (2012)
26. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
27. Lloyd, S.P.: Least squares quantization in pcm. IEEE Transactions on Information Theory **28**(2), 129–137 (1982)
28. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
29. Lougee-Heimer, R.: The common optimization interface for operations research: Promoting open-source software in the operations research community. IBM Journal of Research and Development **47**(1), 57–66 (2003). https://doi.org/10.1147/rd.471.0057
30. Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., Sun, L., Wang, J.: Conditional detr for fast training convergence. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3651–3660 (2021)
31. Meyer, P.: Life - a tactical comic for the blind people (2013)
32. Nagrani, A., Zisserman, A.: From benedict cumberbatch to sherlock holmes: Character identification in tv series without a script (2017)
33. Nguyen, N.V., Rigaud, C., Burie, J.C.: Digital comics image indexing based on deep learning. Journal of Imaging **4**(7), 89 (2018)
34. Nguyen, N.V., Rigaud, C., Burie, J.C.: Comic mtl: optimized multi-task learning for comic book image analysis. International Journal on Document Analysis and Recognition (IJDAR) **22**, 265–284 (2019)

35. Ogawa, T., Otsubo, A., Narita, R., Matsui, Y., Yamasaki, T., Aizawa, K.: Object detection for comics using manga109 annotations. arXiv preprint arXiv:1803.08670 (2018)
36. Pang, X., Cao, Y., Lau, R.W., Chan, A.B.: A robust panel extraction method for manga. In: Proceedings of the 22nd ACM international conference on Multimedia. pp. 1125–1128 (2014)
37. Piriyothinkul, B., Pasupa, K., Sugimoto, M.: Detecting text in manga using stroke width transform. In: 2019 11th International Conference on Knowledge and Smart Technology (KST). pp. 142–147. IEEE (2019)
38. Qin, X., Zhou, Y., Li, Y., Wang, S., Wang, Y., Tang, Z.: Progressive deep feature learning for manga character recognition via unlabeled training data. In: Proceedings of the ACM Turing Celebration Conference-China. pp. 1–6 (2019)
39. Ramaprasad, R.: Comics for everyone: Generating accessible text descriptions for comic strips. arXiv preprint arXiv:2310.00698 (2023)
40. Rayar, F.: Accessible comics for visually impaired people: Challenges and opportunities. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 3, pp. 9–14. IEEE (2017)
41. Rigaud, C., Le Thanh, N., Burie, J.C., Ogier, J.M., Iwata, M., Imazu, E., Kise, K.: Speech balloon and speaker association for comics and manga understanding. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 351–355. IEEE (2015)
42. Sachdeva, R., Cordeiro, F.R., Belagiannis, V., Reid, I., Carneiro, G.: Evidentialmix: Learning with combined open-set and closed-set noisy labels. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 3607–3615 (2021)
43. Sachdeva, R., Cordeiro, F.R., Belagiannis, V., Reid, I., Carneiro, G.: Scanmix: learning from severe label noise via semantic clustering and semi-supervised learning. Pattern recognition **134**, 109121 (2023)
44. Sachdeva, R., Zisserman, A.: The manga whisperer: Automatically generating transcriptions for comics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12967–12976 (2024)
45. Samarawickrama, C., Lenadora, D., Ranathunge, R., De Silva, Y., Perera, I., Welivita, K.: Comic based learning for students with visual impairments. International Journal of Disability, Development and Education **70**(5), 769–787 (2023)
46. Soykan, G., Yuret, D., Sezgin, T.M.: Identity-aware semi-supervised learning for comic character re-identification. arXiv preprint arXiv:2308.09096 (2023)
47. Star wars audio comics. https://www.youtube.com/@StarWarsAudioComics/
48. Topal, B.B., Yuret, D., Sezgin, T.M.: Domain-adaptive self-supervised pre-training for face & body detection in drawings. arXiv preprint arXiv:2211.10641 (2022)
49. Tsubota, K., Ogawa, T., Yamasaki, T., Aizawa, K.: Adaptation of manga face representation for accurate clustering. In: SIGGRAPH Asia 2018 Posters. pp. 1–2 (2018)
50. Wang, Y., Zhou, Y., Tang, Z.: Comic frame extraction via line segments combination. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 856–860. IEEE (2015)
51. Xu, M., Yuan, X., Shen, J., Yan, S.: Cast2face: Character identification in movie with actor-character correspondence. In: Proceedings of the 18th ACM international conference on Multimedia. pp. 831–834 (2010)
52. Zhang, Z., Wang, Z., Hu, W.: Unsupervised manga character re-identification via face-body and spatial-temporal associated clustering. arXiv preprint arXiv:2204.04621 (2022)