

# Supplementary Materials: Zero-Shot Character Identification and Speaker Prediction in Comics via Iterative Multimodal Fusion

Anonymous Authors

## A LLM PROMPTS

In this section, we show the prompts of large language models (LLMs) for speaker prediction. Through our experiments, we observed that using prompts in Japanese leads to higher prediction accuracy than using prompts in English. This is probably because the input texts are in Japanese; using prompts in the same language could reduce confusion for the LLMs. For the understanding of our prompts, we have translated them into English. The introduced prompts are fed into GPT-4 as system prompts while user prompts only contain the texts in each comic.

### A.1 Character Name Extraction

The prompt we used for extracting character names from dialogues is shown below. Since the names that appear in dialogues might be incomplete, potentially leading to multiple names being extracted for the same character, we instructed the LLMs to use contextual information to infer and output full names wherever possible.

```
1 Given a sequence of manga text in Japanese,
2 identify the names of the characters
3 estimated to appear.
4
5 ### Note
6 - When extracting character names, provide
7 full names if possible, e.g., "Taro
8 Yamada".
9 - If full names are not explicitly
10 mentioned, analyze the context within the
11 text to deduce the full names.
12 - If the name of a character is unknown,
13 describe them by their occupation or their
14 relationship with other characters, e.g.,
15 "the teacher" or "Yamada's mother".
16
17 ### Input/Output format
18 [Input format]
19 Text ID | Text
20
21 [Output format]
22 Character ID | Character Name
```

### A.2 Context Extraction

The prompt we used for extracting a story summary and character profiles is shown below. Here, we take the prompt for LoveHina vol01 as an example. We instructed the LLMs to output the context information in Japanese to make the language consistent as mentioned above.

```
1 Given a sequence of manga text in Japanese
2 and a list of characters who appear in it,
3 produce a story summary and character
4 profiles based on the following steps.
5 Note: The output should be in Japanese.
6
7 1. Summary: summarize the story in the
8 manga.
9
10 2. Characters: For each character listed,
11 provide details about their attributes,
12 including gender, estimated age, role, a
13 brief description, and relationships with
14 other characters.
15
16 ### List of appearing characters
17 Character ID | Character Name
18 A | Keitaro
19 B | Naru
20 ...
21
22 ### Input/Output format
23 [Input format]
24 Text ID | Text
25
26 [Output format]
27 1. Summary:
28 2. Characters:
29 - Keitaro:
30 - Naru:
31 - ...
```

### A.3 Initial Speaker Prediction (w/o candidates)

The prompt of the initial speaker prediction based only on the texts is shown below. We instructed the LLMs to output not only the predicted speaker's name but also a confidence level for that prediction. The data with a low confidence level is not used in the subsequent steps. We defined five confidence levels and provided detailed explanations for the criteria.

```
1 Given a sequence of manga text in Japanese
2 and a list of characters who appear in it,
3 predict the speaker of each text
4 considering the context information.
5 Please also output a confidence level of
6 prediction on a scale of 5.
7
8 ### Note
9 - Not all the given characters might be
10 speaking in the provided text.
```

```

117 6 - The number of output lines should be the
118 ↪ same as the number of input lines.
119 7
120 8 ### List of appearing characters
121 9 Character ID | Character Name
122 10 A | Keitaro
123 11 B | Naru
124 12 ...
125 13
126 14 ### Context information
127 15 1. Summary: Keitaro tries to ...
128 16 2. Characters:
129 17 - Keitaro: Main character who ...
130 18 - Naru: Heroine of the story. ...
131 19 - ...
132 20
133 21 ### Input/Output format
134 22 [Input format]
135 23 Text ID | Text
136 24
137 25 [Output format]
138 26 Text ID | Character Name | Character ID |
139 ↪ Confidence Level
140 27
141 28 ### Confidence level
142 29 Score and criteria:
143 30 1: Completely uncertain, the prediction is
144 ↪ near random.
145 31 2: Low confidence, the probability that the
146 ↪ prediction is correct is less than 50%.
147 32 3: Moderate confidence, the prediction is
148 ↪ likely correct but could be wrong.
149 33 4: High confidence, the prediction is
150 ↪ probably correct, but not 100% certain.
151 34 5: Very high confidence, the prediction is
152 ↪ almost certainly correct.
153 35
154 36 ### Input/Output example
155 37 [Input]
156 38 1 | Hey, Naru.
157 39 2 | What, Keitaro?
158 40
159 41 [Output]
160 42 1 | Keitaro | A | 5
161 43 2 | Naru | B | 4
162

```

#### A.4 Iterative Speaker Prediction (w/ candidates)

The prompt of the iterative speaker prediction using the speaker candidates is shown below. A speaker candidate is obtained for each text based on the character identification results. We integrated image-based predictions and textual information by providing LLMs with speaker candidates. We also supplied the prediction probability for each speaker candidate so that LLMs can use the information, which leads to 2.0% improvement in accuracy as shown in Table 3 of the main paper.

```

1 Given a sequence of manga text in Japanese
↪ and a list of characters who appear in it,
↪ predict the speaker of each text
↪ considering the context information.
2 Please also output a confidence level of
↪ prediction on a scale of 5.
3
4 For each text, you will be given a speaker
↪ candidate, which is obtained from the
↪ image-based prediction, along with a
↪ probability for that prediction. Use this
↪ as a reference.
5
6 ### Note
7 - Not all the given characters might be
↪ speaking in the provided text.
8 - The number of output lines should be the
↪ same as the number of input lines.
9 - If no image-based predictions are given,
↪ predict the speaker based on the text and
↪ the context.
10 - The image-based predictions may not
↪ always be correct. Exercise caution,
↪ especially when the prediction probability
↪ is low.
11
12
13 ### List of appearing characters
14 Character ID | Character Name
15 A | Keitaro
16 B | Naru
17 ...
18
19 ### Context information
20 1. Summary: Keitaro tries to ...
21 2. Characters:
22 - Keitaro: Main character who ...
23 - Naru: Heroine of the story. ...
24 - ...
25
26 ### Input/Output format
27 [Input format]
28 Text ID | Text | Speaker Candidate
↪ (Prediction Probability)
29
30 [Output format]
31 Text ID | Character Name | Character ID |
↪ Confidence Level
32
33 ### Confidence level
34 Score and criteria:
35 1: Completely uncertain, the prediction is
↪ near random.
36 2: Low confidence, the probability that the
↪ prediction is correct is less than 50%.
231
232

```

```

37 3: Moderate confidence, the prediction is
38 likely correct but could be wrong.
39 4: High confidence, the prediction is
40 probably correct, but not 100% certain.
41 5: Very high confidence, the prediction is
42 almost certainly correct.
43
44 ### Input/Output example
45 [Input]
46 1 | Hey, Naru. | Keitaro (0.56)
47 2 | What, Keitaro? | Naru (0.8)
48
49 [Output]
50 1 | Keitaro | A | 5
51 2 | Naru | B | 4

```

## B CHARACTER REGION CLASSIFIER

This section describes the details of character region classification. We utilize the ResNet50 model [3], adapted to classify character regions. Our training method consists of a pre-training phase on general comics and a fine-tuning phase for individual unseen comics in the test set. The steps of pre-training, fine-tuning, and testing are detailed below.

**Pre-training.** We pre-train the models for comic character classification using Manga109 dataset [1, 7]. The model are initialized with weights from a model trained on ImageNet [2, 3]. The training set includes 349 characters from the Manga109 training set, which consists of separated titles from the test set. Character regions are cropped and resized to 270×270, followed by data augmentation including random 256×256 crops, horizontal flips, and rotations. Training lasts for 50 epochs using AdamW optimizer [5] with a batch size of 32. The learning rate is set to  $1 \times 10^{-4}$ , and it is reduced by a factor of 0.1 after the 20th and 40th epochs.

**Fine-tuning.** In the character identification step of our framework, the pre-trained model is fine-tuned for each comic in the test set using the pseudo labels generated from speaker prediction results. We split the whole data into train and validation sets, allocating 10% of the data randomly for validation. We run fine-tuning with 10 epochs and select the model with the highest validation accuracy. The learning rate is set to  $1 \times 10^{-4}$ , with other parameters and augmentation strategies being consistent with the pre-training phase.

**Testing.** In testing, we used 10-crop testing [4] with crops sized at 256×256 from the image resized to 270×270. To handle the instability of training with noisy labels, we used a simple ensemble approach: we trained five models using the same training data, where the output probability of classification is calculated by averaging the outputs from all models. As future work, we might be able to employ specific methods to handle noisy labels to improve the results [6, 8].

## C MAIN RESULTS

In this section, we present more examples of our experimental results, which were obtained using the same experimental settings as

the results of Table 1 in the main paper. As mentioned in Section 4.2, to validate the effectiveness of our proposed iterative multimodal fusion method, we conducted evaluations in two aspects: **Unimodal vs. Multimodal** and **One-Step vs. Iterative**. Additionally, we analyze the limitations of our current method through several failure examples to show the direction of future work.

### C.1 Unimodal vs. Multimodal

To show the effectiveness of using multimodal information, we compare our proposed multimodal method with the methods using only visual information or textual content, as shown in Figure A and Figure B. Color of the bounding box indicates the predicted character label. Red labels indicate failure predictions. The results of *LLM only* and *K-means+SGG* baseline are shown as text-only and image-only results, respectively. As explained in the main paper, *K-means+SGG* used the ground truth to map each cluster into character labels because the image-only approach cannot identify character labels. For the *LLM only* method, we did not perform character identification.

From Figure A, we can observe that LLMs struggle to make accurate predictions in the case that the texts lack distinctive character features. However, when combining these with image-based predictions, visual information provides essential cues about the speakers, enabling the model to make correct predictions. The results in Figure B show that an image-only method failed to predict the case that the speaker is not the character closest to the text. In contrast, the multimodal method can predict correctly by using textual content such as story context. These results show the effectiveness of using multimodal information for these tasks, which accords with how humans do when reading comics, i.e., identifying the speaker of dialogue using both visual and textual information.

### C.2 One-Step vs. Iterative

Experimental results of character identification and speaker prediction under different iteration times are shown in Figure C and Figure D. The accuracy for both character identification and speaker prediction improves as the number of iterations increases. This demonstrates that enhanced results of character identification can positively influence the accuracy of speaker prediction, and vice versa, further validating the effectiveness of our iterative approach.

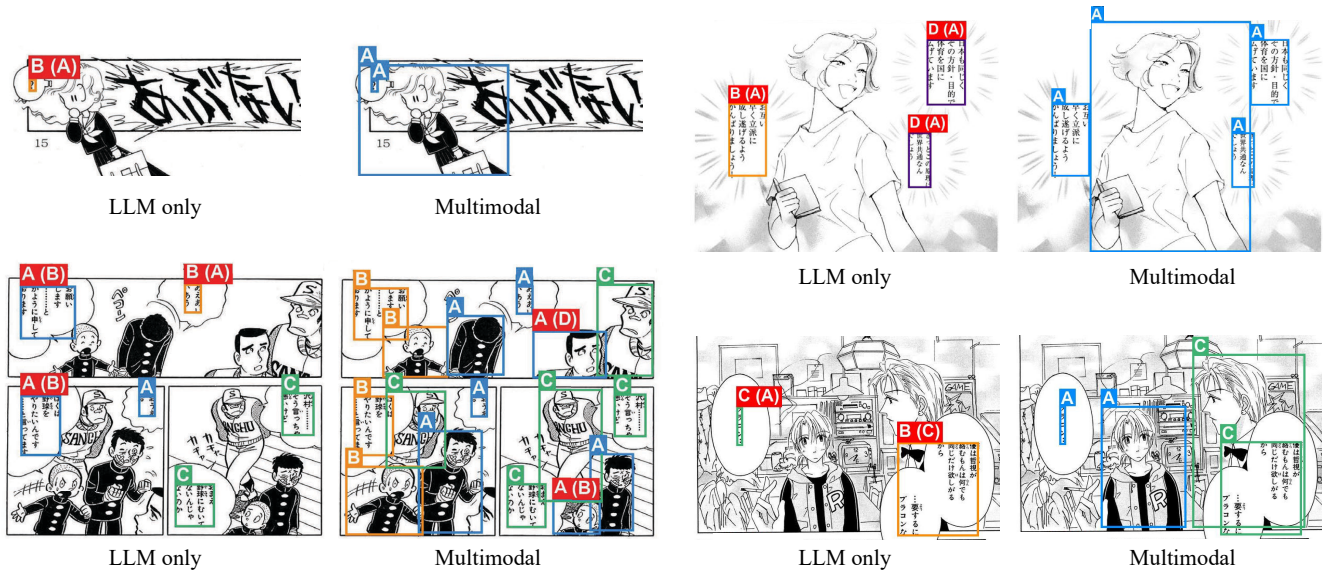


Figure A: Speaker prediction results obtained using a single modality (textual information) and multiple modalities combined. Color of the bounding box indicates the predicted character label. Red labels indicate failure predictions. Courtesy of Tashiro Kimu, Hikochi Sakuya, Yoshimori Mikio, Karikawa Seyu.

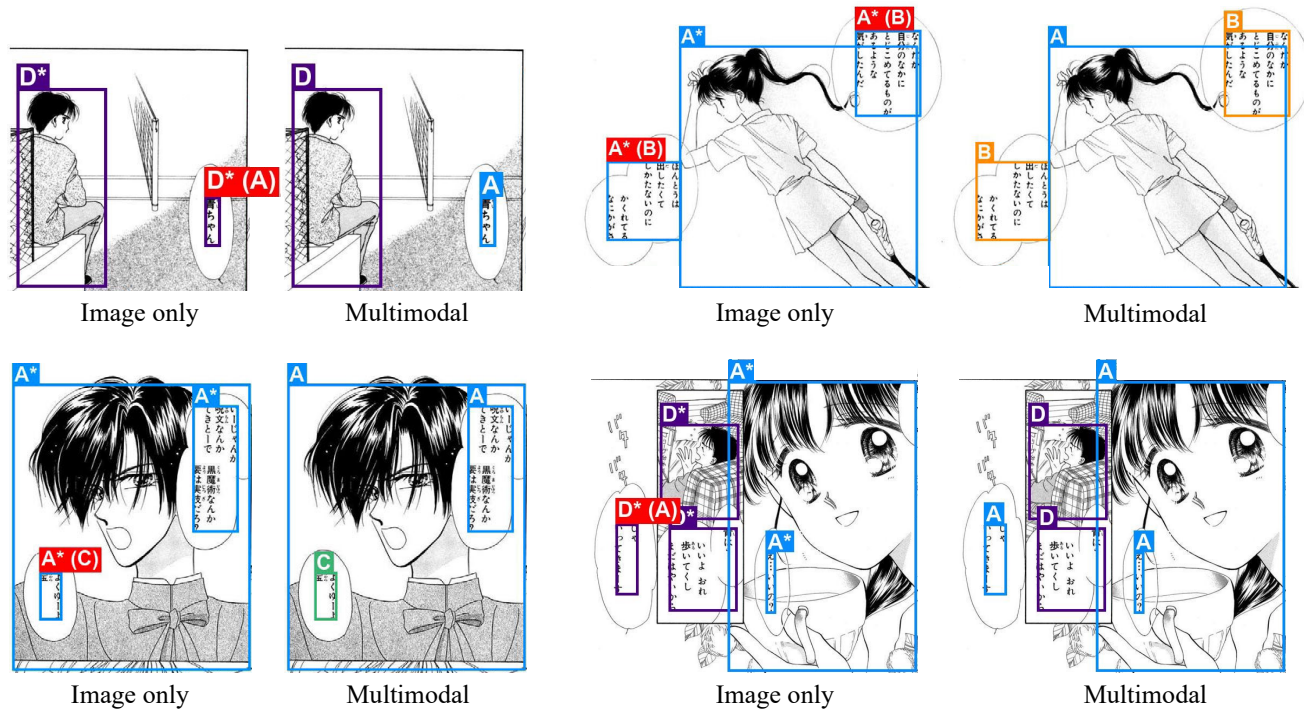


Figure B: Speaker prediction results obtained using a single modality (visual information) and multiple modalities combined. The labels on the boxes (e.g., 'A') are character labels. Labels in brackets are the ground truth. (e.g., 'A (B)' is the case where the ground truth is B but the prediction is A.) \* indicates that the image-based method used the ground truth of character labels. Courtesy of Ayumi Yui, Hanada Sakumi.



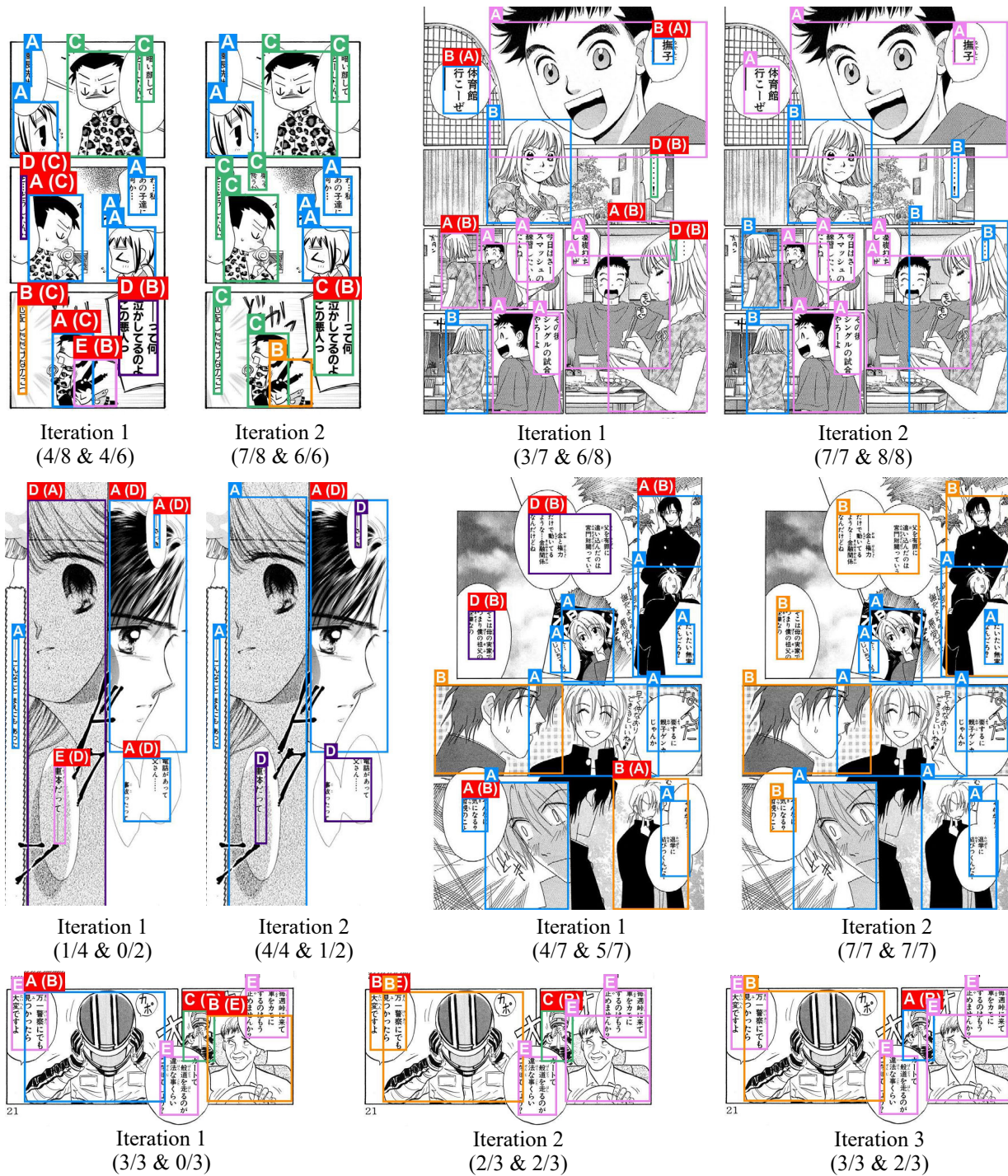
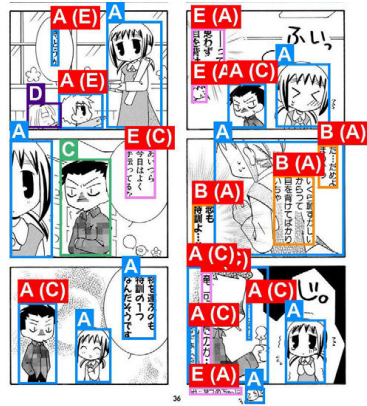
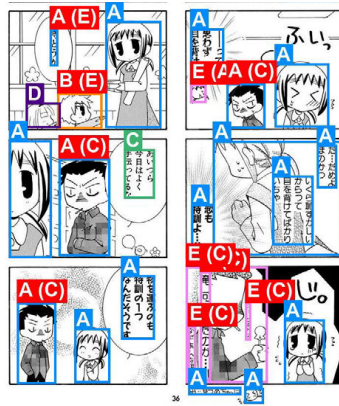


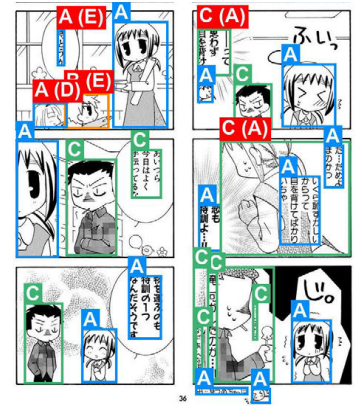
Figure C: Results of character identification and speaker prediction across different iterations. (Accuracy of speaker pred. & character id.). Courtesy of Tenya, Saki Kaori, Ayumi Yui, Karikawa Seyu, Matsuda Naomasa.



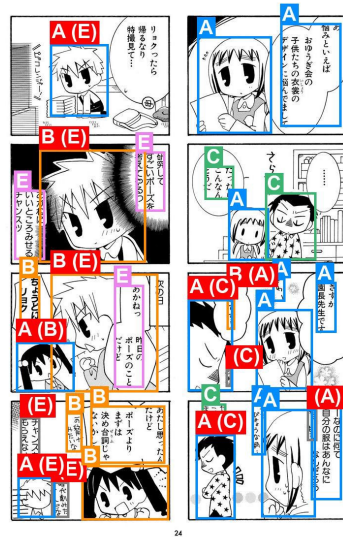
Iteration 1 (1/11 &amp; 8/14)



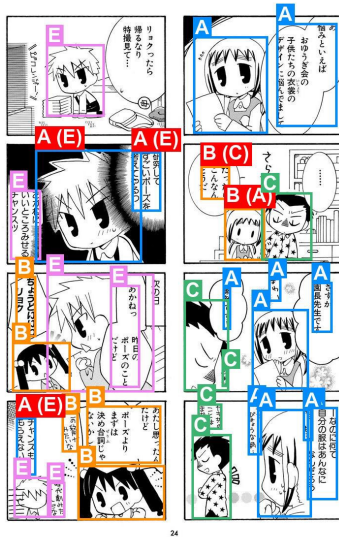
Iteration 2 (4/11 &amp; 7/14)



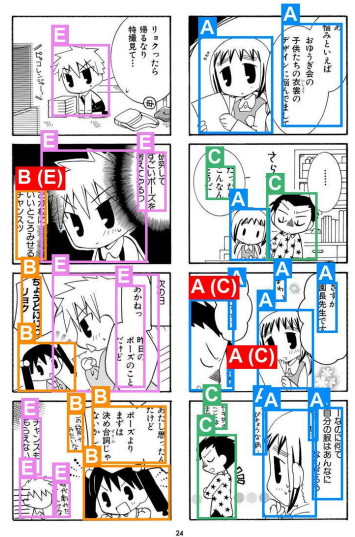
Iteration 3 (9/11 &amp; 11/14)



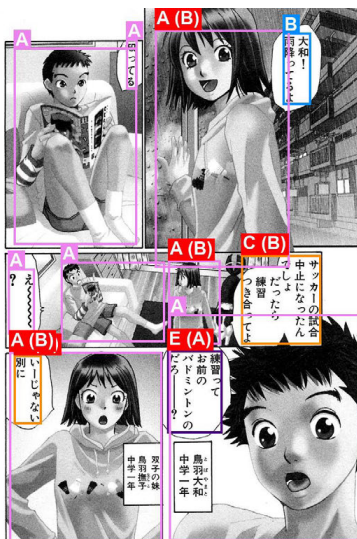
Iteration 1 (12/18 &amp; 6/13)



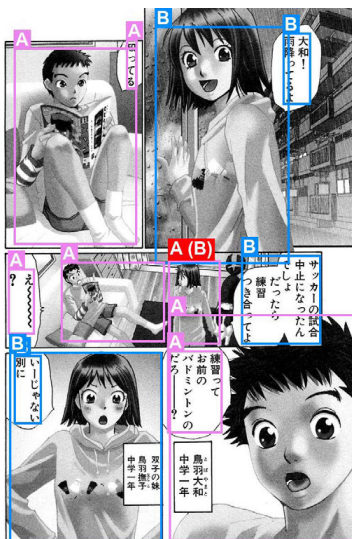
Iteration 2 (15/18 &amp; 11/13)



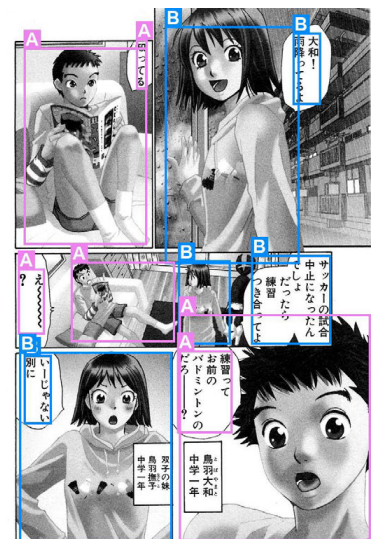
Iteration 3 (16/18 &amp; 12/13)



Iteration 1 (3/6 &amp; 3/6)



Iteration 2 (6/6 &amp; 5/6)



Iteration 3 (6/6 &amp; 6/6)

Figure D: Results of character identification and speaker prediction across different iterations. (Accuracy of speaker pred. & character id.). Courtesy of Tenya, Saki Kaori.



### C.3 Failure Examples

Figure E shows the cases where increasing the number of iterations leads to poorer prediction results. In the first example, for the texts positioned on the left (purple box with label D), the LLMs initially made correct speaker prediction in iteration 1. However, since character B is closer to these texts, its label is propagated to these texts, leading to incorrect predictions in iteration 2. In the second example, the character classifier originally made the correct identification. However, labels of the character C are changed to incorrect label B. This is because the labels to dialogue are propagated to characters even if the character does not speak in this figure.

These examples show that improving the accuracy of either character identification or speaker prediction does not necessarily enhance the performance of the other in the case that the correspondence between text and character regions is not clear. It suggests future directions such as a method to generate reliable pseudo-labels or a robust training method to handle noisy labels.

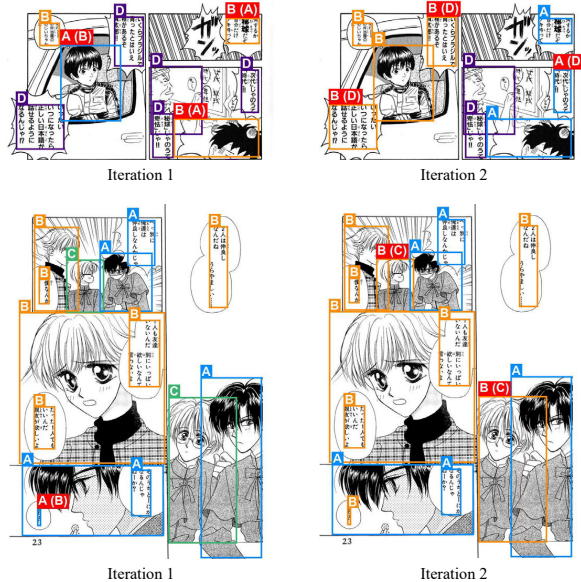


Figure E: Failure examples where prediction results deteriorate with increasing iterations. Courtesy of Matsuda Naomasa, Hanada Sakumi.

### D ZERO-SHOT RESULTS

Figure F shows the results under entirely zero-shot settings where only images are provided as inputs. Experimental settings are described in Section 4.4 of the main paper.

### REFERENCES

- [1] Kiyoharu Aizawa, Azuma Fujimoto, Atsushi Otsu, Toru Ogawa, Yusuke Matsui, Koki Tsubota, and Hikaru Ikuta. 2020. Building a manga dataset “manga109” with annotations for multimedia applications. *IEEE MultiMedia* 27, 2 (2020), 8–18.
- [2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 248–255.

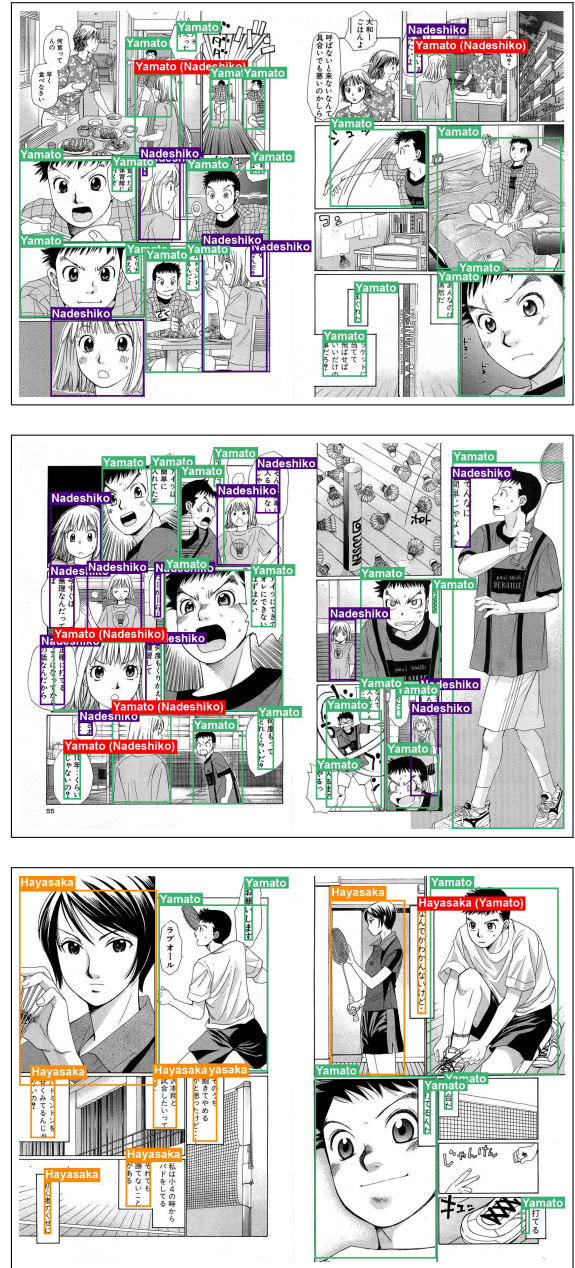


Figure F: Results under an entirely zero-shot setting. Courtesy of Saki Kaori.

- [3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, Vol. 25.
- [5] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. arXiv:1711.05101
- [6] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. 2013. Learning with noisy labels. In *Advances in neural information processing systems*, Vol. 26.

- [7] Toru Ogawa, Atsushi Otsubo, Rei Narita, Yusuke Matsui, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2018. Object detection for comics using manga109 annotations. arXiv:1803.08670
- [8] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2022. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems* (2022).

871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917  
918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928