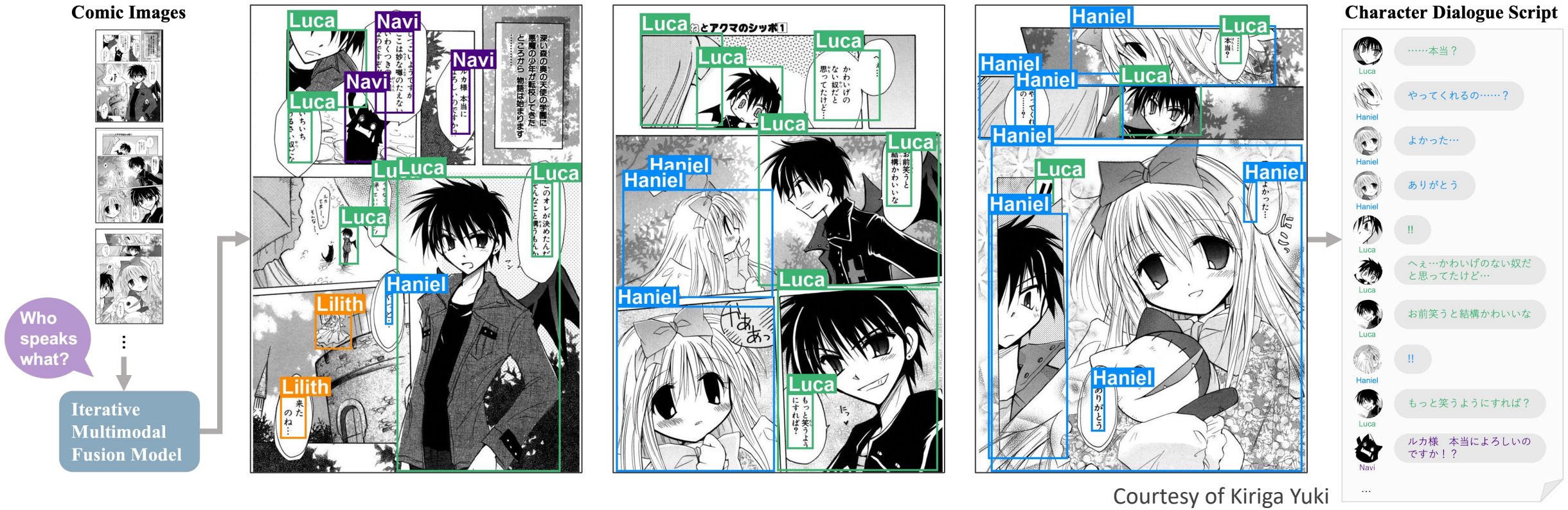
Zero-Shot Character Identification and Speaker Prediction in Comics via Iterative Multimodal Fusion

Yingxuan Li [1], Ryota Hinami [2], Kiyoharu Aizawa [1], Yusuke Matsui [1]

[1] The University of Tokyo, [2] Mantra Inc.

Topic

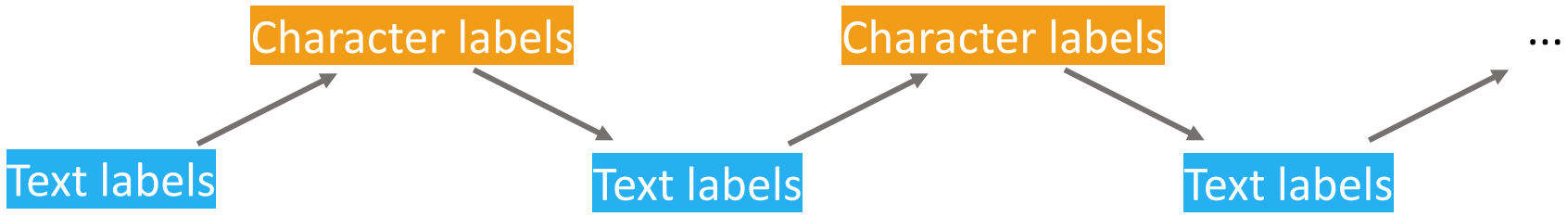
- **Zero-Shot character identification and speaker prediction in comics**
  - Identify characters and predict speakers of unseen comics ***only from images***



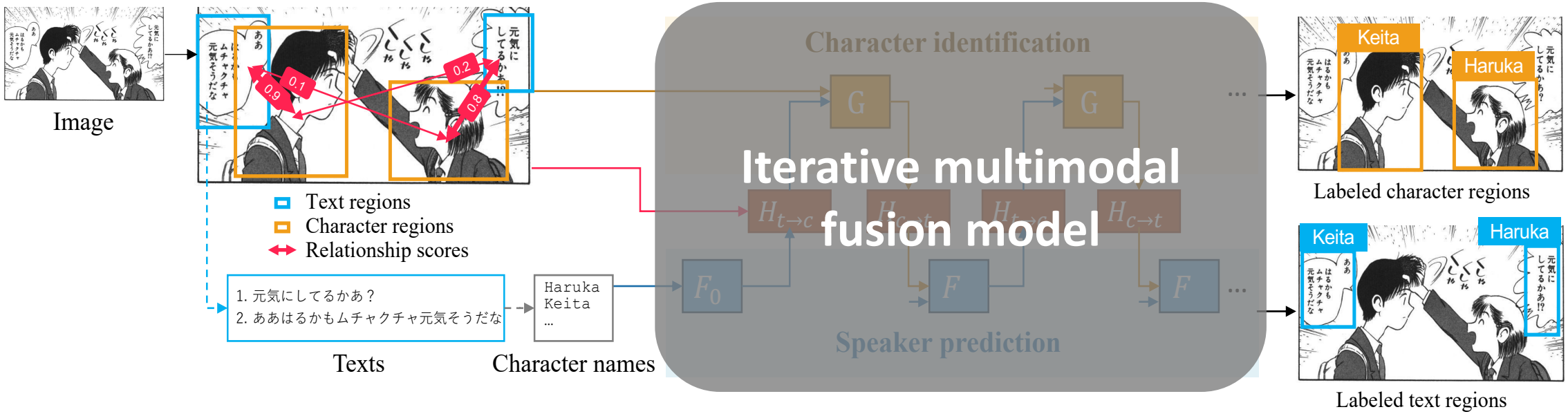
Courtesy of Kiriga Yuki

Proposal

- **Research focus**
  - Predict character names for both text and character regions
  - Tackle ***zero-shot*** tasks without requiring any annotations
  - Enhance real-world applicability
- **Proposed method**
  - **Multimodal fusion approach:** Merge text-based large language model (LLM) predictions with image-based classifiers
  - **Iterative process:** Alternately refine speaker prediction using character labels, and character identification using text labels

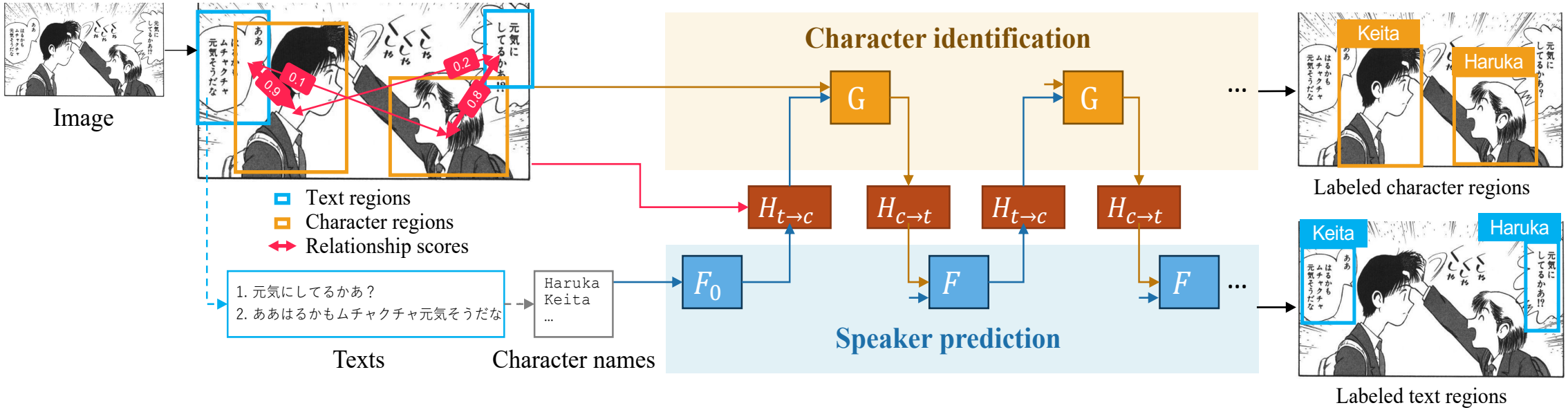


Overall framework



- **Data preprocessing**
  - **Object detection:** Get character regions and text regions
  - **Relationship prediction:** Get initial relationship scores
  - **OCR:** Get texts
  - **Character name extraction:** Get target labels

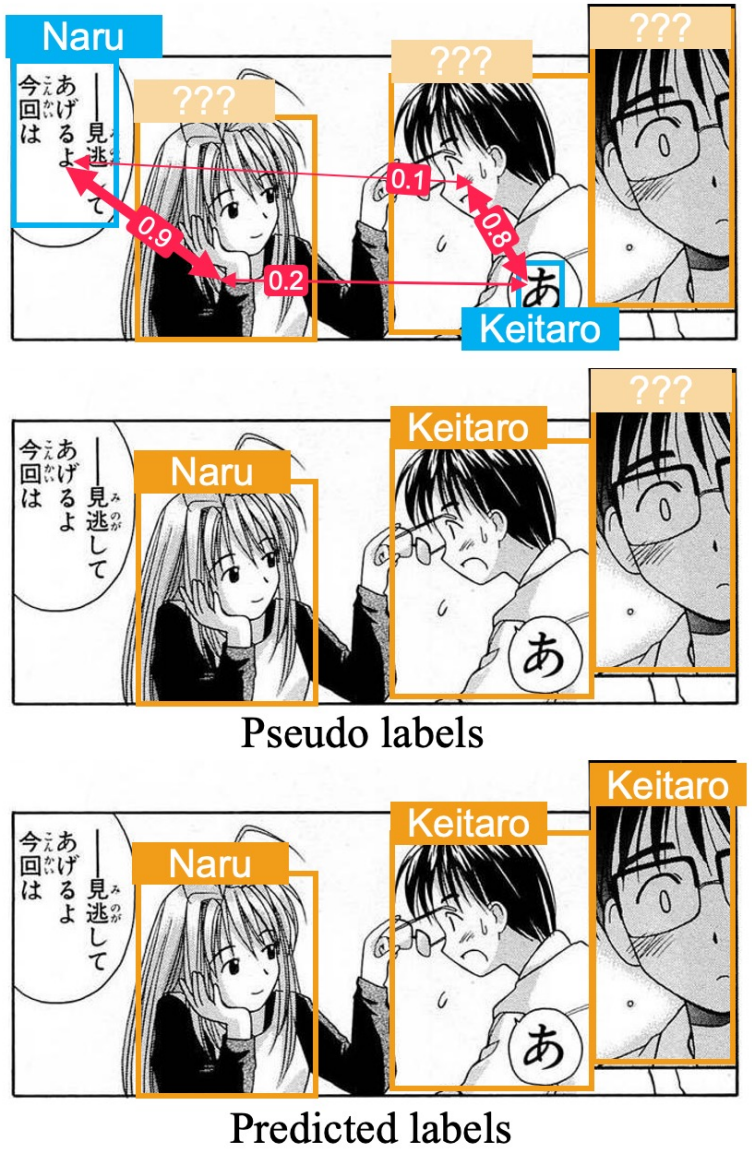
Overall framework



- **Main pipeline:** Three modules
  - **Speaker prediction (F):** Textual info. + Character labels → Text labels
  - **Character identification (G):** Visual info. + Text labels → Character labels
  - **Label propagation ($H_{t\to c}$, $H_{c\to t}$):** Text labels ↔ Character labels

- Introduction
- **Approach**
- Experiments
- Discussion and Conclusion

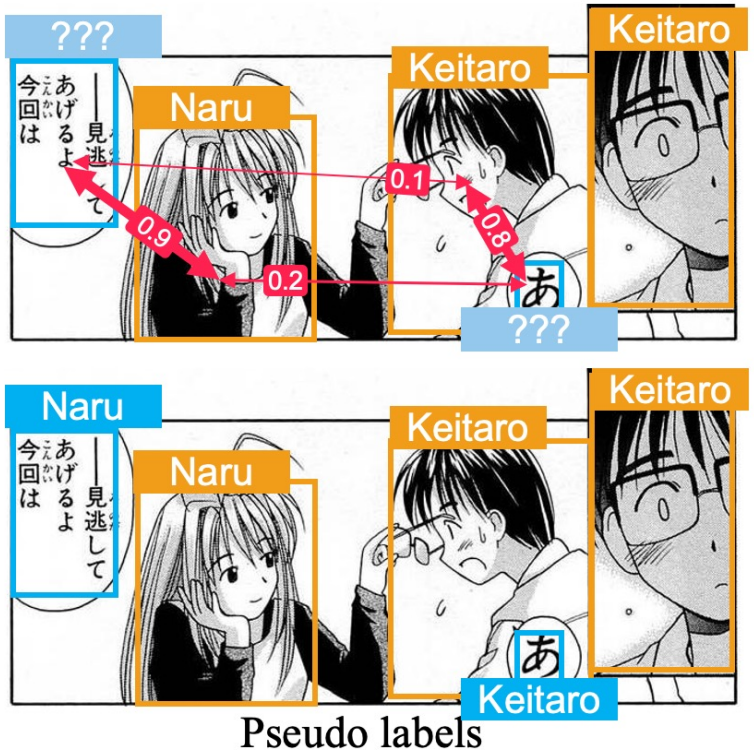
Iterative character identification



- **Step 1: Pseudo label generation**
  - Select the character–text pair with the highest relationship score
  - Take the text region's label as the pseudo label
- **Step 2: Character identification**
  - Fine-tune a classifier with pseudo labels
  - Apply the classifier to character regions

Courtesy of Akamatsu Ken

Iterative speaker prediction



Pseudo labels

Input

```
Given dialogs and character names in comics,
annotate the speaker of each dialog.

Characters:
A | Keitaro
B | Naru

Context:
Story: Keitaro tries to ...
Characters:
- Keitaro: Main character who ...
- Naru: Heroine of the story. ...

Dialog:
1 | Hey, Naru.     | Keitaro (0.56)
2 | What, Keitaro? | Naru(0.8)
```

Character names

Context information (obtained beforehand by LLMs)

Texts

Speaker candidates

Prompt → LLM

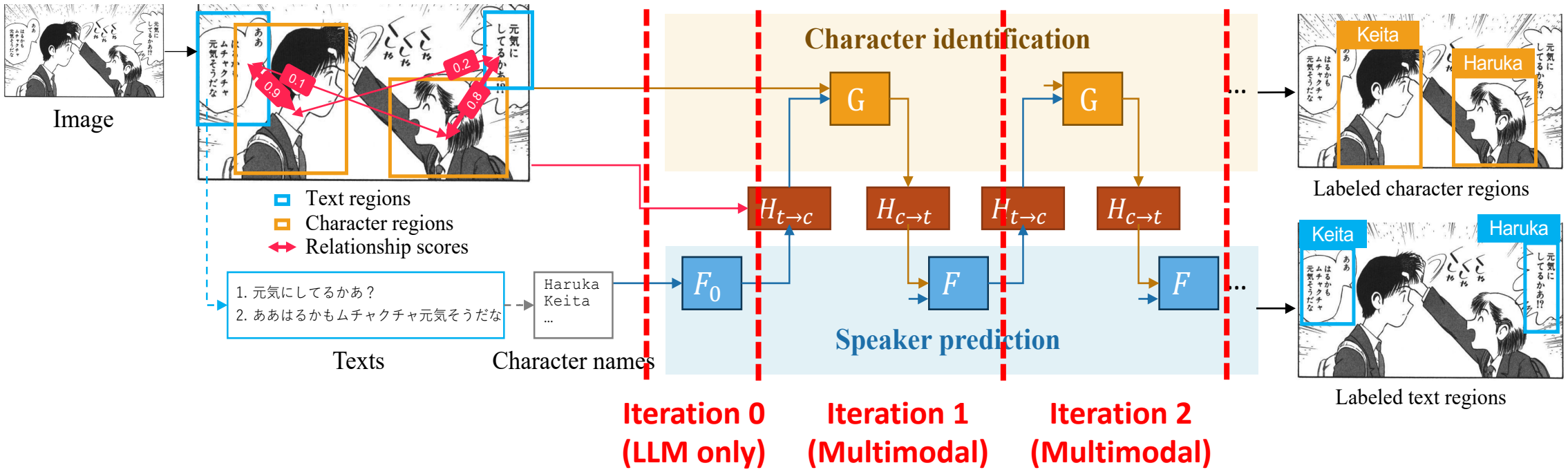
Output

```
L | Character | Id | Confidence
1 | Keitaro   | A  |     5
2 | Naru      | B  |     4
```

Predicted labels

- **Step 1: Pseudo label generation**
- **Step 2: Speaker prediction**
  - Treat pseudo labels as candidate speakers

Iterative process



- **Iterative performance enhancement**
  - Predictions from the previous step → Close to true labels
  - Generated pseudo labels → Reliable training data
  - Updated prediction results → High accuracy

- Introduction
- Approach
- **Experiments**
- Discussion and Conclusion

Quantitative results

- **Main results:** Simplified the tasks to classifying the character labels for the character and text regions
- **Baselines:** Clustering + mapping clusters to ground truth*
- **Data division:** Divided the test set by the difficulty of relationship prediction

	iter	text	img	Speaker pred. *Easy*	Speaker pred. *Hard*	Speaker pred. *Total*	Character id. *Easy*	Character id. *Hard*	Character id. *Total*
Baseline									
K-means+Distance	-		✓	34.5*	31.8*	33.1*	37.0*	36.7*	36.8*
K-means+SGG	-		✓	36.7*	34.8*	35.7*	37.0*	36.7*	36.8*
Proposed									
LLM only	0	✓		41.8	45.1	43.6	-	-	-
Multimodal	1	✓	✓	51.0	51.2	51.1	45.8	39.6	42.4
	2	✓	✓	52.4	**51.3**	**51.8**	48.5	**40.3**	**44.0**
	3	✓	✓	**53.5**	49.8	51.6	**48.9**	37.7	42.8

Our multimodal method shows a significant improvement over unimodal methods

Quantitative results

- **Main results:** Simplified the tasks to classifying the character labels for the character and text regions
- **Baselines:** Clustering + mapping clusters to ground truth*
- **Data division:** Divided the test set by the difficulty of relationship prediction

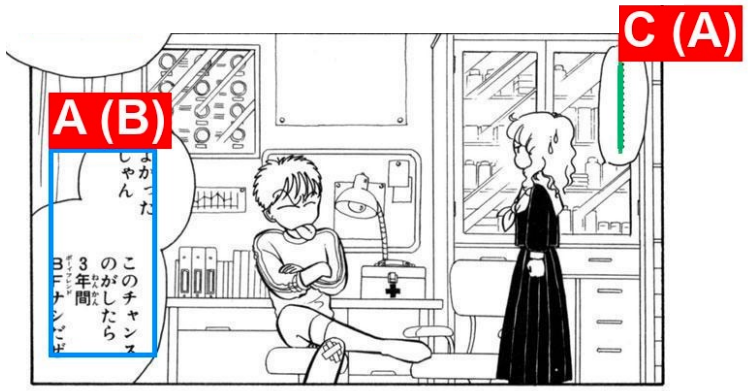
	iter	text	img	Speaker pred. *Easy*	Speaker pred. *Hard*	Speaker pred. *Total*	Character id. *Easy*	Character id. *Hard*	Character id. *Total*
Baseline									
K-means+Distance	-		✓	34.5*	31.8*	33.1*	37.0*	36.7*	36.8*
K-means+SGG	-		✓	36.7*	34.8*	35.7*	37.0*	36.7*	36.8*
Proposed									
LLM only	0	✓		41.8	45.1	43.6	-	-	-
Multimodal	1	✓	✓	51.0	51.2	51.1	45.8	39.6	42.4
	2	✓	✓	52.4	**51.3**	**51.8**	48.5	**40.3**	**44.0**
	3	✓	✓	**53.5**	49.8	51.6	**48.9**	37.7	42.8

Our iterative process is effective on both subsets, particularly on *Easy*

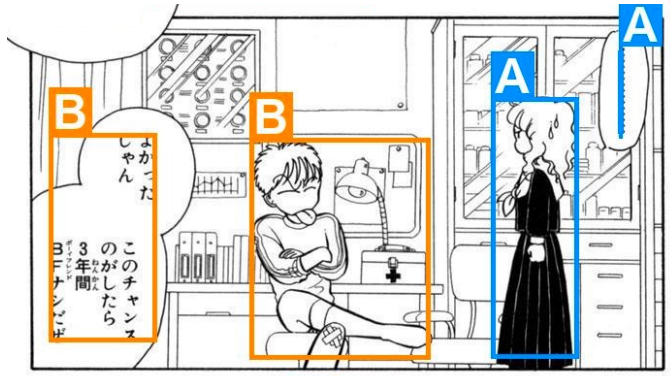
Qualitative results

- **Unimodal vs. Multimodal**

■ Mislabeled predictions (ground truth)



LLM only



Multimodal

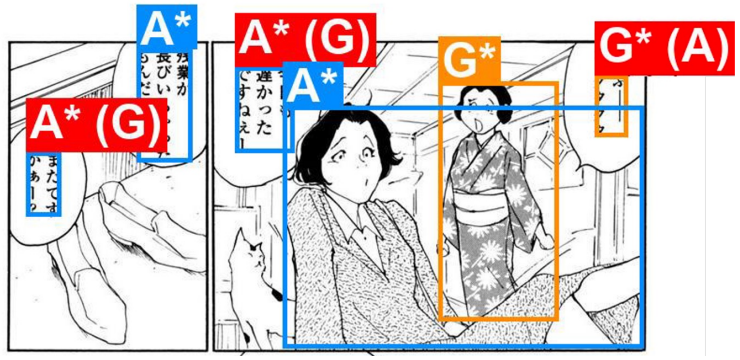
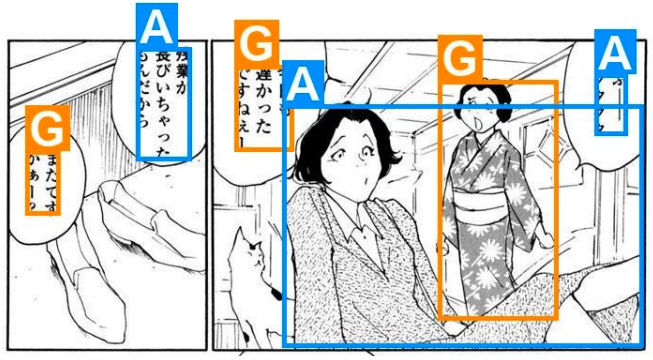


Image only (using GT*)

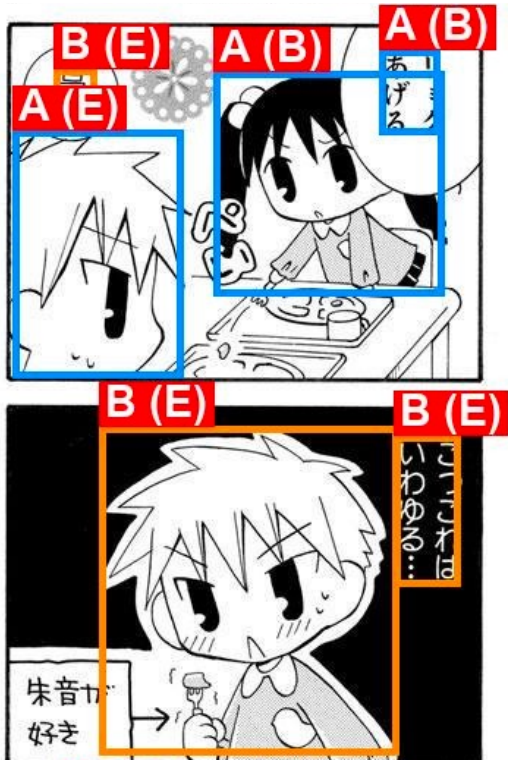


Multimodal

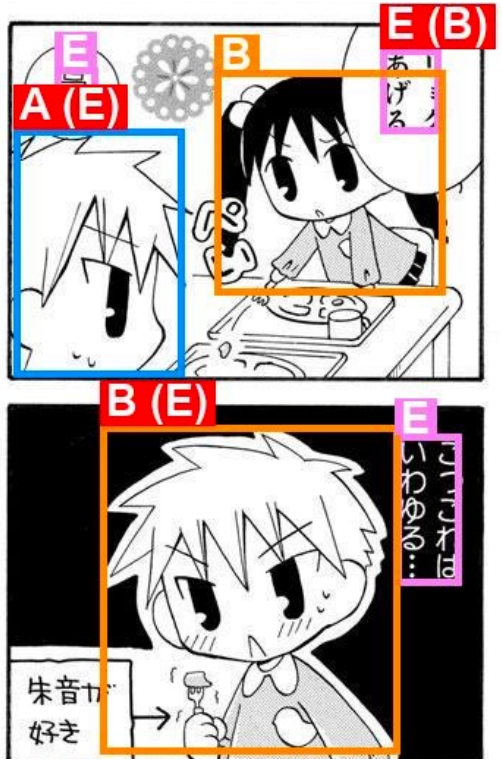
Courtesy of Tashiro Kimu, Hikochi Sakuya

Qualitative results

- **One-step vs. Iterative** (Accuracy of speaker pred. & character id.)



Iteration 1 (0/3 & 0/3)



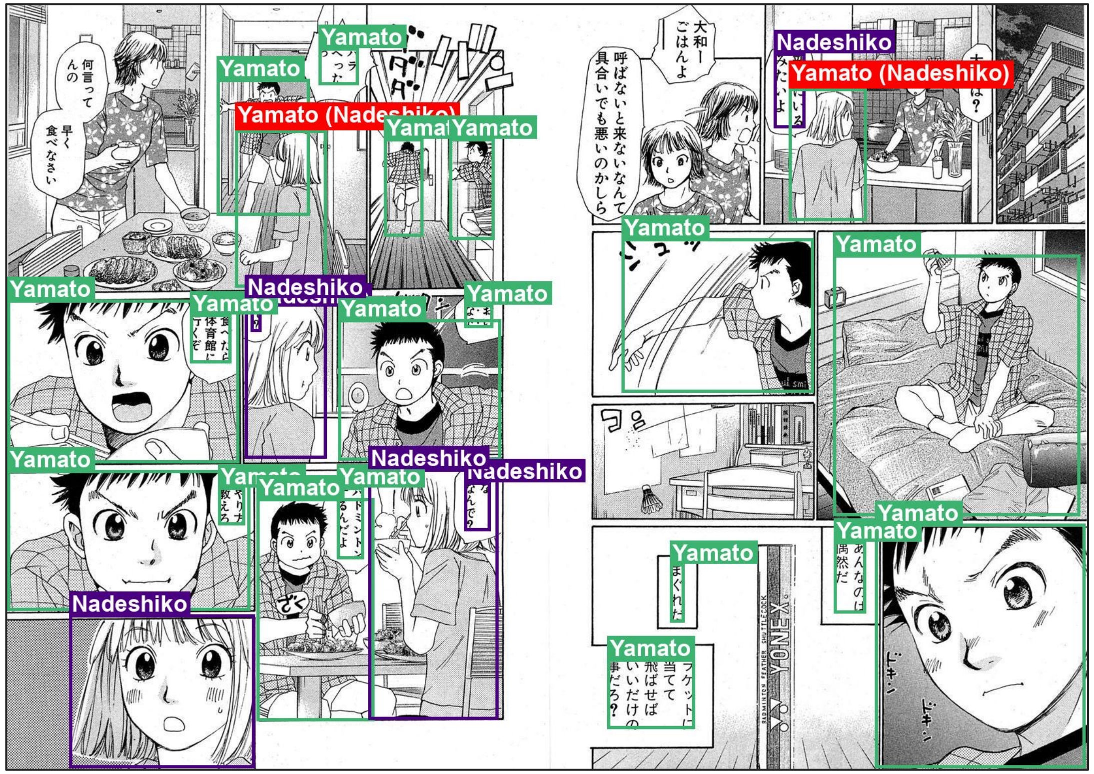
Iteration 2 (1/3 & 1/3)

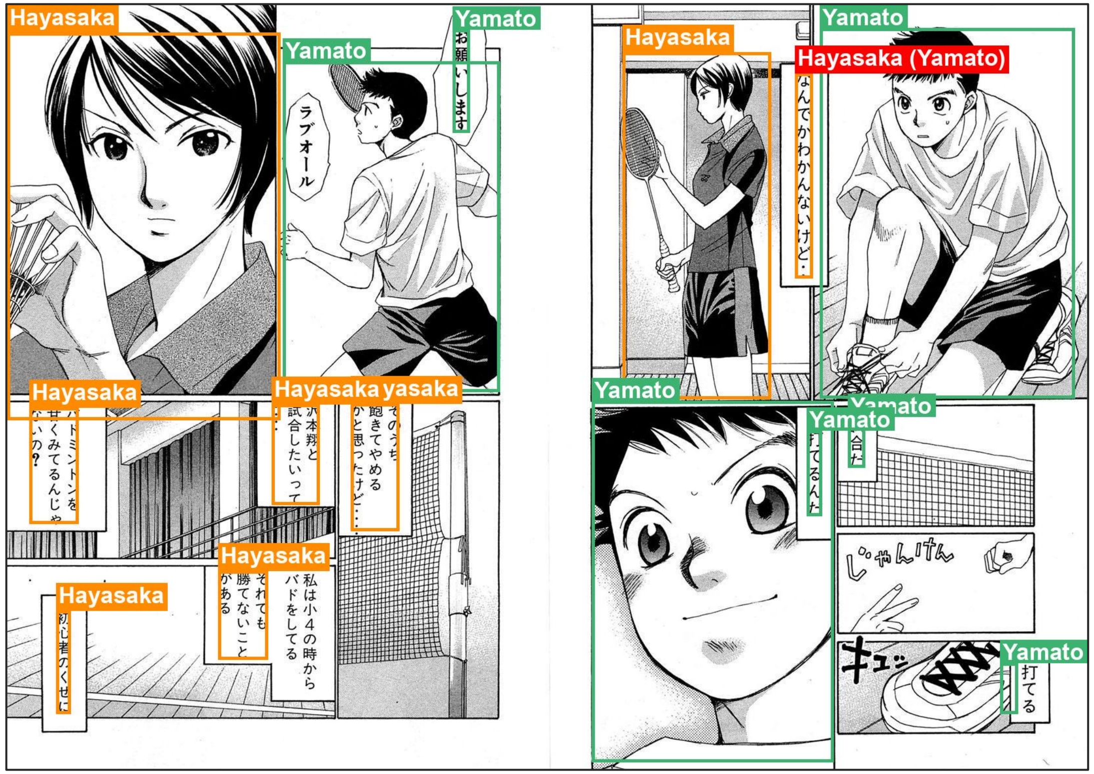
Iteration 3 (3/3 & 2/3)

Courtesy of Ito Tenya

Zero-shot results

- **Results under entirely zero-shot settings**



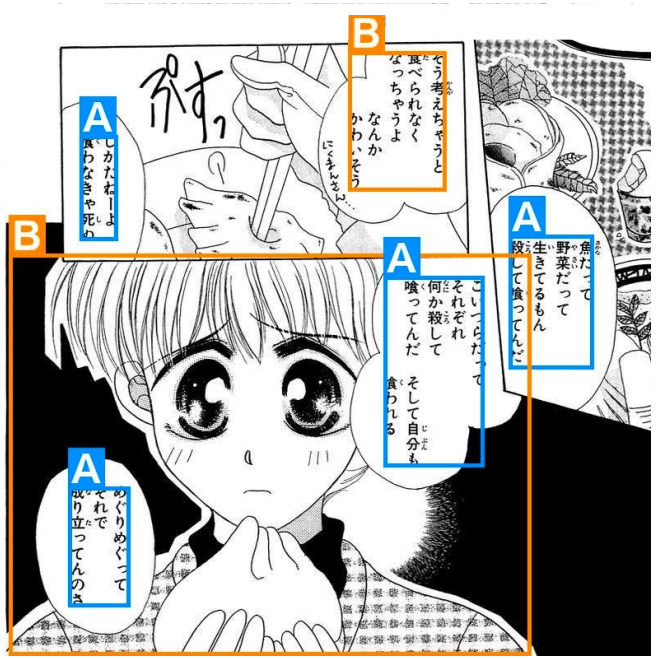


Courtesy of Saki Kaori

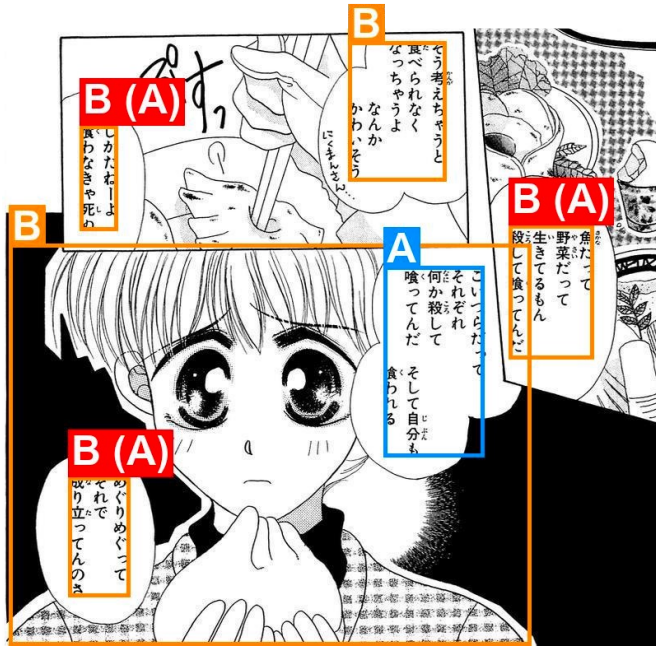
- Introduction
- Approach
- Experiments
- **Discussion and Conclusion**

Discussion

- **Limitations of proposed method**
  - The overall accuracy is limited
  - More iterations do not necessarily lead to higher accuracy



Iteration 1



Iteration 2

Courtesy of Hanada Sakumi

Conclusion

- **New tasks**
  - First to integrate the tasks of character identification and speaker prediction in comics
  - First to tackle zero-shot tasks with direct applications in real-world scenarios
- **Iterative multimodal fusion**
  - Revealing the significant potential of LLMs for comics analysis
  - First approach to use multimodal information for character identification and speaker prediction
- **Future work**
  - Refine the model and enhance accuracy

Explore more

- **Visit our poster presentation!**
  - **Poster session:** Poster Session 3
  - **Posterboard:** P175
- **Explore our project page for more details**

Thank you!

UTokyo MANTRA ACM Multimedia 2024 Melbourne, Australia