

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

Yingxuan Li¹, Ryota Hinami², Kiyoharu Aizawa¹, Yusuke Matsui¹

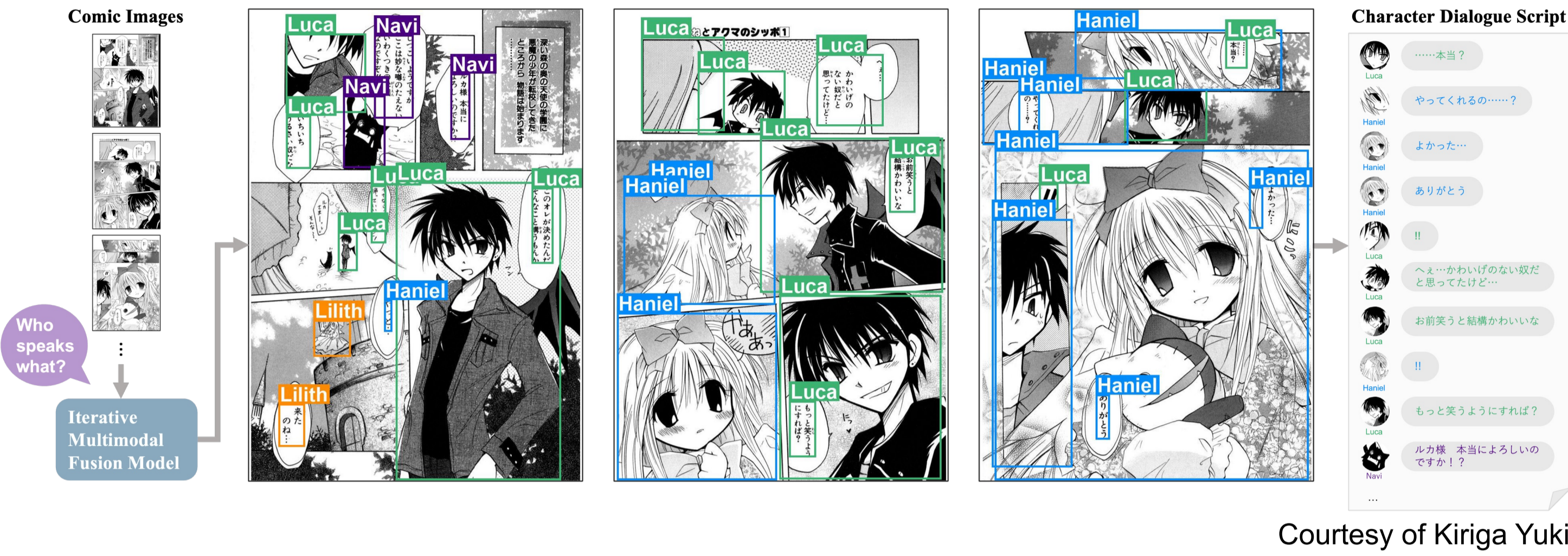
¹The University of Tokyo, ²Mantra Inc.



Introduction

Novel task

- Identify characters and predict speakers of unseen comics **only from images**



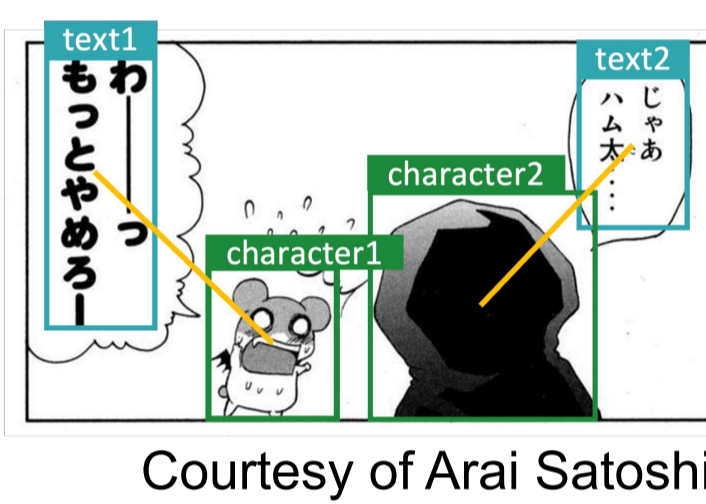
Applications

- Automatic character assignment for audiobooks
- Automatic translation according to characters' personalities
- Inference of character relationships and stories
- ...

Motivation

Limitations of previous studies

- Speaker prediction:** Focused only on predicting the correspondence [1]
- Character identification:** Required annotations and specific classifiers for each comic title [2]



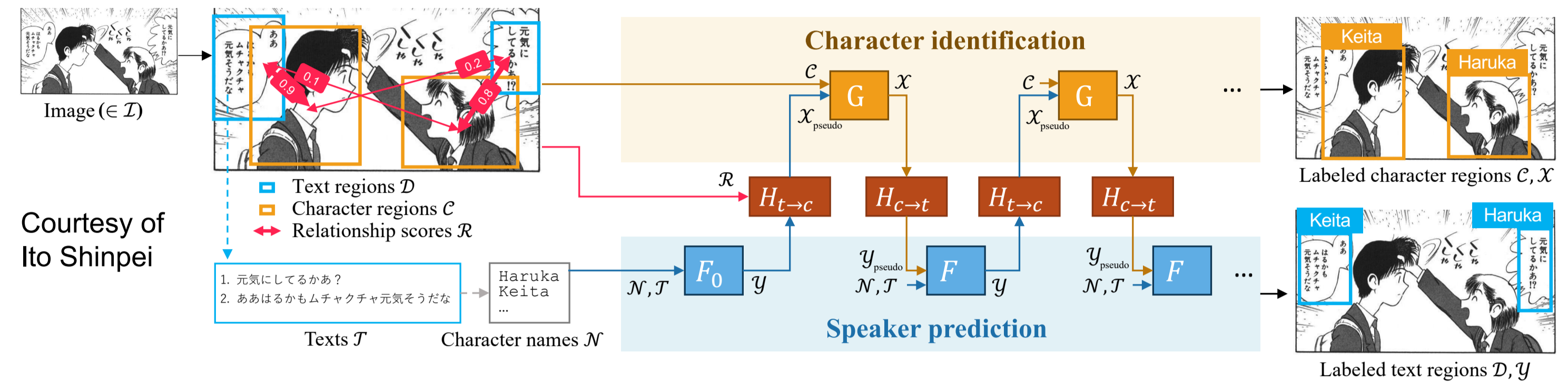
Research focus

- Predict character names for both text and character regions
- Tackle **zero-shot** tasks without requiring any annotations
- Enhance real-world applicability

Approach

Iterative multimodal fusion

- Leverage large language models (LLMs)
- Merge text-based LLM predictions with image-based classifiers
- Alternately refine each module using results from the other



Data preprocessing

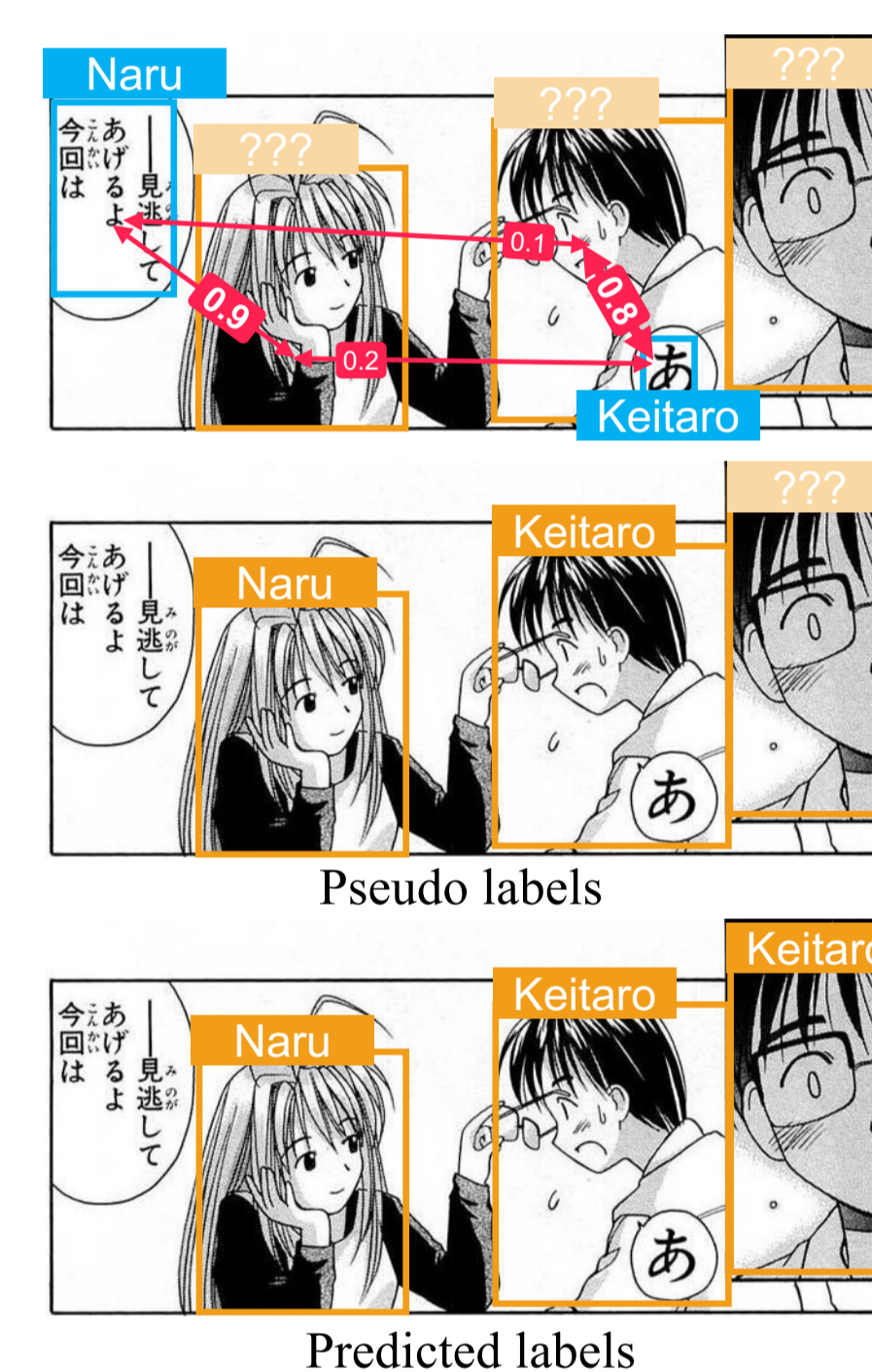
- Object detection: $\mathcal{I} \mapsto \mathcal{C}, \mathcal{D}$
- Relationship prediction: $\mathcal{I}, \mathcal{C}, \mathcal{D} \mapsto \mathcal{R}$
- OCR: $\mathcal{I}, \mathcal{D} \mapsto \mathcal{T}$
- Character name extraction: $\mathcal{T} \mapsto \mathcal{N}$

Main pipeline: Three modules

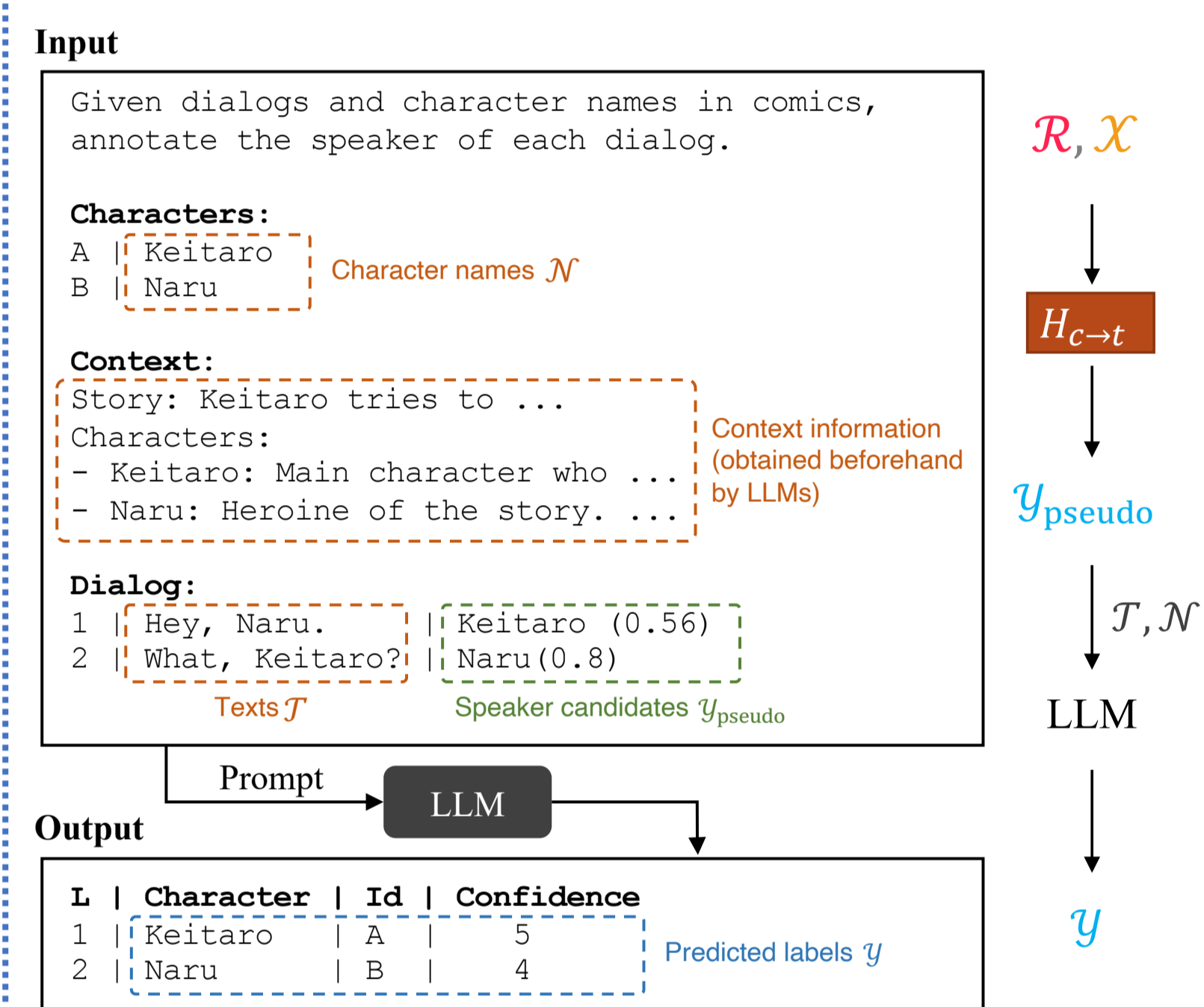
- Speaker prediction module (F)
- Character identification module (G)
- Label propagation module ($H_{t \rightarrow c}, H_{c \rightarrow t}$)

Initial speaker prediction: $\mathcal{T}, \mathcal{N} \xrightarrow{F_0} \mathcal{Y}$

Iterative character identification



Iterative speaker prediction



Experiments

Main results

- Dataset**
 - Annotations:** Manga109 [3] + Manga109Dialog [1]
 - Test set:** 23 volumes that were unseen in the training set
- Task settings**
 - Set object regions (\mathcal{C}, \mathcal{D}), texts (\mathcal{T}), and the name list (\mathcal{N}) to known
- Baselines**
 - Character identification:** Clustering + mapping clusters to ground truth*
 - Speaker prediction:** Previous approaches + character identification results
- Data division**
 - Divided the test set into *Easy* and *Hard* by the difficulty of relationship prediction
 - Easy:** 11 volumes with an accuracy of relationship prediction over 75%

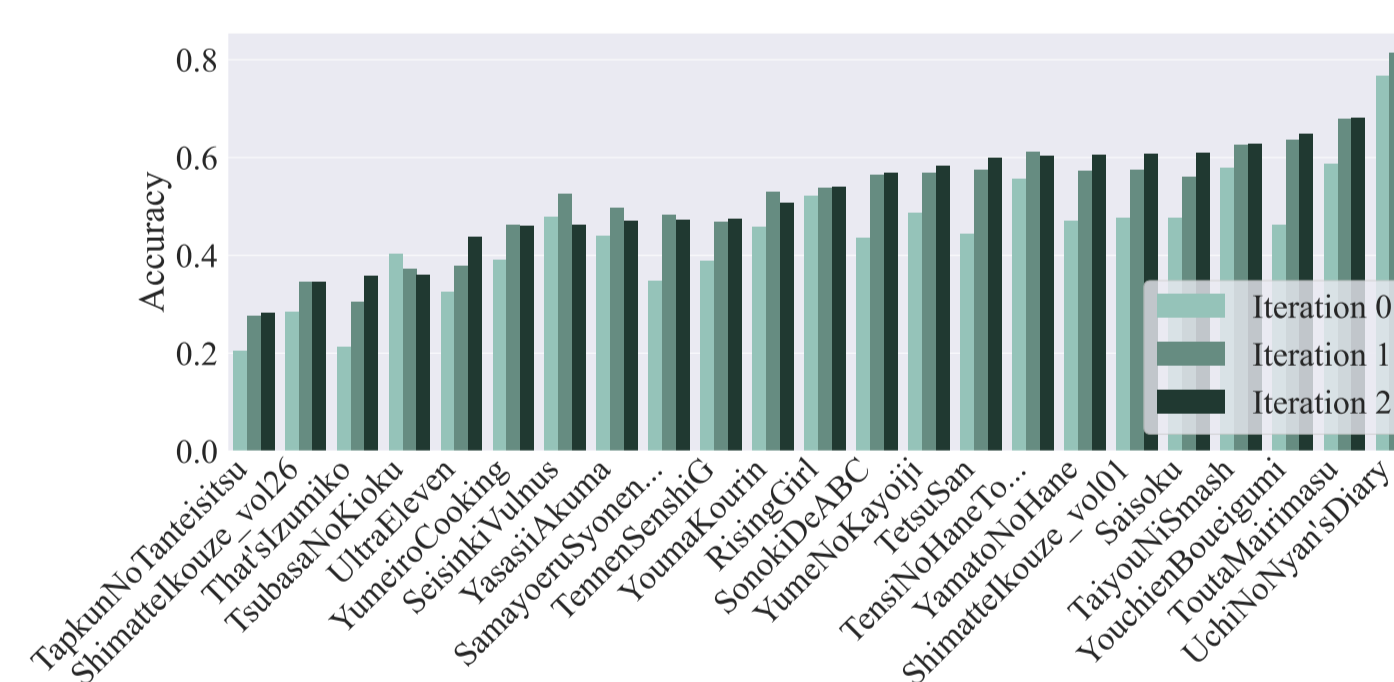
	Speaker pred.			Character id.		
	iter	text	img	Easy	Hard	Total
Baseline						
K-means+Distance	-	✓		34.5*	31.8*	33.1*
K-means+SGG	-	✓		36.7*	34.8*	35.7*
Proposed						
LLM only	0	✓		41.8	45.1	43.6
Multimodal	1	✓	✓	51.0	51.2	51.1
	2	✓	✓	52.4	51.3	51.8
	3	✓	✓	53.5	49.8	51.6
				48.9	37.7	42.8

(a) Results on different test sets. * indicates that the baseline method used the ground truth to map clusters into labels, as explained in the experimental setup.

	iter	Speaker pred.		Character id.	
		Speaker pred.	Character id.		
Baseline					
K-means+GT	-	42.0*	36.8*		
Proposed					
LLM only	0	43.6	-		
Multimodal	1	60.2	53.9		
	2	63.4	55.5		
	3	63.8	56.6		

(b) Results using the ground truth relationships.

Speaker prediction accuracy of each comic title



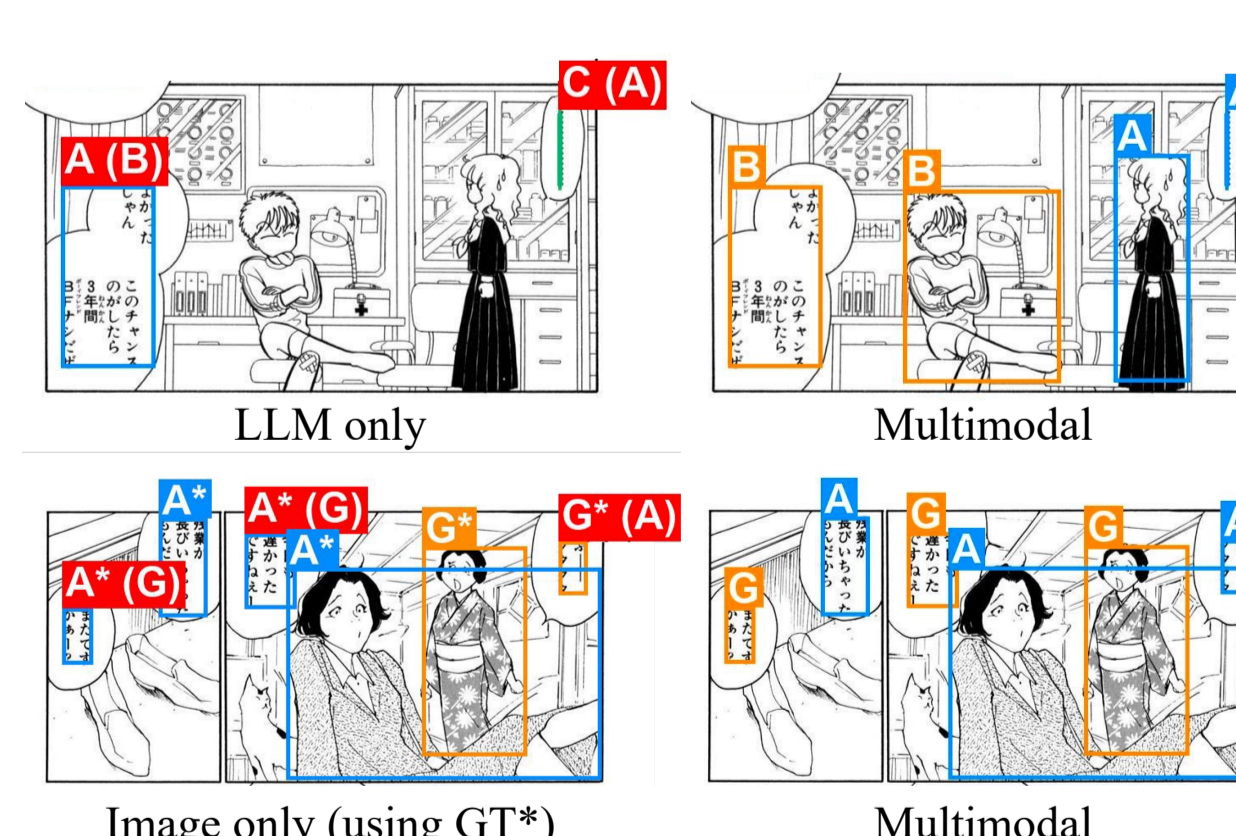
Zero-shot results

	iter	Speaker pred.	Character id.
LLM only	0	34.1	-
Multimodal	1	37.7	35.6
	2	38.7	35.0
	3	37.9	33.8
Upper bound	-	67.3	63.9

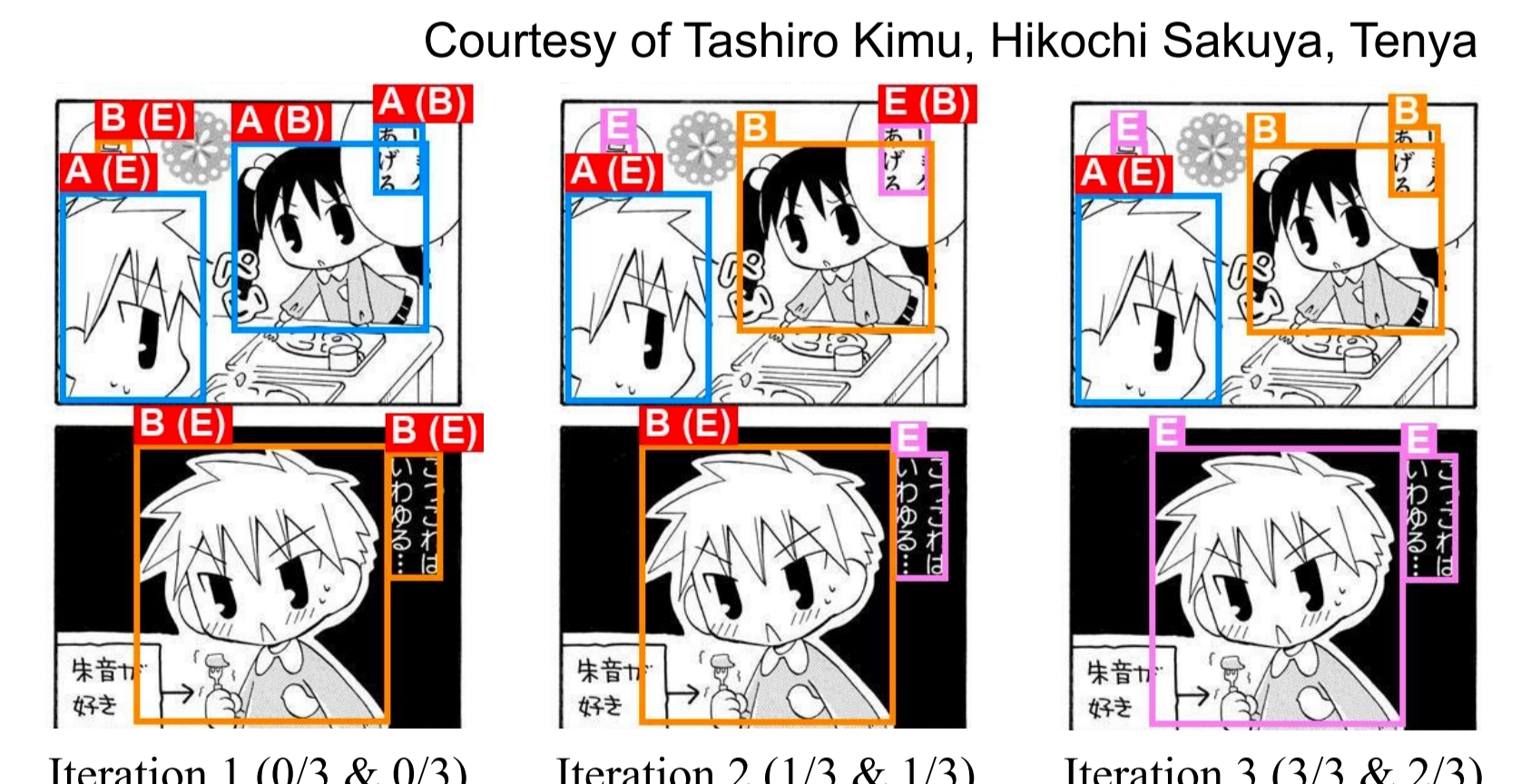
Correct prediction: The region was detected with an IoU > 0.5 and was correctly labeled
Upper bound: Accuracy under ideal conditions (when all labels of extracted names are perfectly predicted)

Qualitative results

Unimodal vs. Multimodal



One-step vs. Iterative



Reference

- Manga109Dialog: A large-scale dialogue dataset for comics speaker detection. Li et al. ICME 2024.
- Cartoon face recognition: A benchmark dataset. Zheng et al. ACM MM 2020.
- Building a manga dataset "manga109" with annotations for multimedia applications. Aizawa et al. IEEE MultiMedia 2020.

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 both text and image information for character identification and speaker prediction

Explore our project page for more details!



Paper on OpenReview



Project page