

From Anonymous Speaker Seeds to Consistent Characters: Zero-Shot Manga Speaker Attribution

Yanhao Chen¹, Wataru Shimoda², Kota Yamaguchi², Hideki Nakayama¹

¹The University of Tokyo, Tokyo, Japan

chen-yanhao@g.ecc.u-tokyo.ac.jp, nakayama@ci.i.u-tokyo.ac.jp

²CyberAgent, Tokyo, Japan

{wataru_shimoda,yamaguchi_kota}@cyberagent.co.jp

Abstract

Identifying the speaker of each dialogue instance is a key step in manga understanding, supporting applications such as translation, summarization, and voice generation. In zero-shot settings, the task is challenging because each new manga title introduces an unseen cast and provides no speaker annotations, predefined character identities, or per-title model adaptation. Existing methods, including iterative multimodal fusion, still organize speaker reasoning around manga-specific visual evidence such as characters' appearance and local text-character relationships. However, such evidence is highly variable in manga, where speakers may be off-panel, character appearance changes across poses and styles, and local visual relations can be misleading. We propose Speaker-Seeded Identity Consolidation (SSIC), which reframes zero-shot manga speaker attribution as book-level identity consolidation driven by recurring dialogue consistency rather than local visual-region propagation. SSIC induces local anonymous speaker hypotheses from ordered dialogue context, aligns them into cross-window speaker seeds, constructs a global speaker identity registry, and performs registry-constrained attribution. In the main controlled evaluation on Manga109Dialog, SSIC achieves 74.62% speaker attribution accuracy, improving over iterative multimodal fusion by 7.02 points. The gains are largest on hard cases where local text-character relationships are unreliable, and ablations confirm that cross-window speaker seeds are central to registry construction.

1 Introduction

Manga is a widely read form of visual storytelling that combines sequential drawings, page layout, and written dialogue. Understanding who speaks each dialogue instance is a fundamental step toward character-aware manga understanding, since downstream tasks such as translation, summariza-

tion, and voice generation require dialogue to be grounded in speaker identity.

The problem is especially difficult in zero-shot settings. Each new manga title introduces an unseen cast, visual style, naming convention, and narrative structure, while providing no speaker annotations, predefined character identities, or per-title model adaptation. Thus, the system must not only assign a label to each dialogue instance, but also induce a stable identity space for the target story. This is difficult because manga often violates the simple local assumption that speakers can be identified from nearby visual evidence. For example, a speech balloon may appear while its speaker is outside the panel, leaving no valid nearby character region to match. Conversely, multiple visually plausible characters may be present, while recurring dialogue patterns across later pages provide stronger evidence for the actual speaker. This setting is related to speaker and quotation attribution in NLP, where systems link extracted utterances to speakers using rich narrative context and character mentions (He et al., 2013; Muzny et al., 2017; Zhong et al., 2024). In manga, the extracted text is much sparser than in novels, so speaker cues often come from page layout, visual context, and recurring dialogue patterns rather than explicit textual narration.

Existing methods have made important progress by associating speech balloons with character regions through geometric cues, scene-graph-based text-character relations, or iterative pseudo-label propagation between speaker predictions and character-region labels. Despite these advances, their speaker reasoning remains organized around local visual regions. As a result, the speaker identity space is derived indirectly from associations of the local visual region rather than constructed as a book-level object. Such region-centric mechanisms are effective when the correct character is visible and the local relation is reliable, but they are

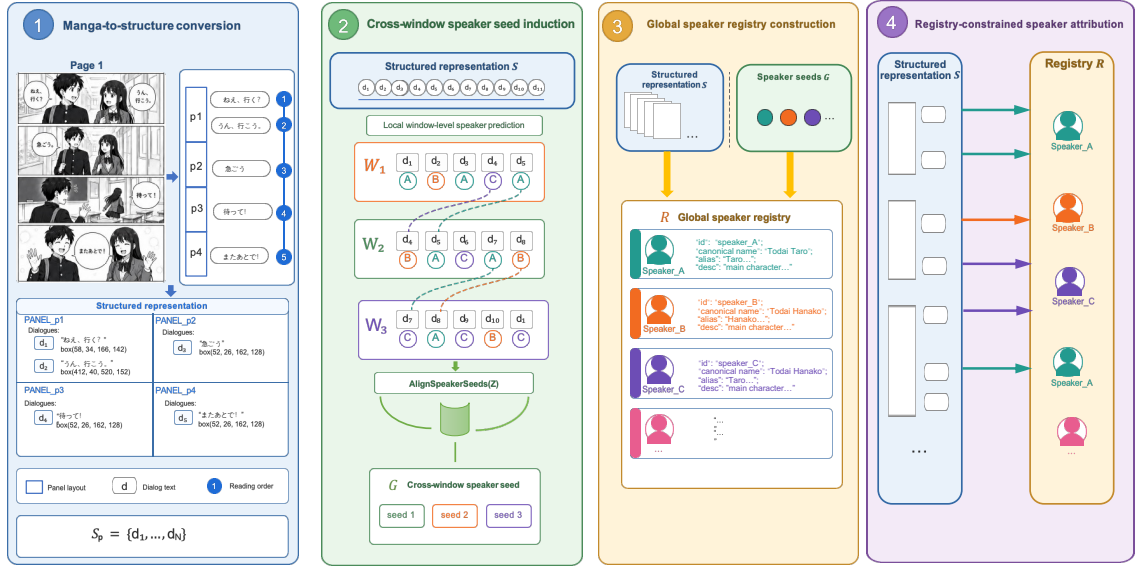


Figure 1: Overview of Speaker-Seeded Identity Consolidation (SSIC).

less suited to cases where the speaker is far from text, visually ambiguous, or recoverable only from recurring dialogue evidence.

Our central idea is to construct the speaker identity space from recurring dialogue consistency rather than from local visual-region propagation. We propose Speaker-Seeded Identity Consolidation (SSIC), a framework that builds a reusable speaker identity space for each target book before assigning individual dialogues. SSIC first groups dialogues within ordered windows into local anonymous speaker hypotheses, then aligns these hypotheses into cross-window speaker seeds. These seeds serve as weak anchors for constructing a global speaker identity registry instead of final labels. Final attribution is performed within this global speaker identity registry, rather than as independent local text-character matching. Figure 1 summarizes this four-stage registry-centric pipeline.

We evaluate SSIC under two input settings. For fair comparison with prior baselines, the main controlled evaluation uses Manga109Dialog benchmark dialogue regions and transcripts as input, isolating identity consolidation from text detection and OCR errors. We also report an end-to-end image-input setting, where this representation is produced by the visual front-end.

Our contributions are twofold. First, we reformulate zero-shot manga speaker attribution as book-

level speaker identity consolidation driven by recurring dialogue consistency, rather than local visual-region propagation. This formulation addresses a key limitation of region-centric methods, whose identity space is derived from manga-specific visual regions and local text-character relations that are unstable under off-panel speech, visual style variation, and unreliable local visibility. Second, we introduce Speaker-Seeded Identity Consolidation (SSIC), which operationalizes this formulation by inducing local anonymous speaker hypotheses from ordered dialogue context, aligning them into cross-window speaker seeds, constructing a reusable speaker identity registry, and performing registry-constrained attribution.

2 Related Work

Region-centric speaker association. Early comics speaker attribution methods formulate the task as associating a text region with a character region, often using geometric cues such as distances between speech balloons and nearby characters or anchor points derived from balloon shapes (Rigaud et al., 2015). Subsequent work studies speech-balloon detection, segmentation, and multi-task comic image analysis as part of the broader pipeline for extracting dialogue and visual entities from comic pages (Dubray and Laubrock, 2019; Nguyen et al., 2019). Manga109Dialog

extends this region-association paradigm with large-scale speaker-to-text annotations and scene graph generation based text-character relationship prediction, further using frame reading order to improve speaker detection (Li et al., 2024a). More recent work combines visual relations with natural-language processing to improve comic speaker prediction (Zhang and Hotta, 2025). These methods provide useful grounding when visual character evidence and text-character relations are reliable. However, their speaker reasoning remains coupled to visual character regions or local text-character associations. SSIC differs by inducing a book-level speaker identity space before assigning individual dialogue regions.

Visual character identity in comics. Character identity in comics has also been studied from the visual side, including supervised cartoon-face recognition and unsupervised manga character clustering or re-identification (Zheng et al., 2020; Tsubota et al., 2018; Zhang et al., 2022; Zhang and Chu, 2023). These methods group or classify character regions by appearance, but manga visual language makes this signal unstable: the same character can vary across poses, expressions, panels, and drawing styles, while different characters may share similar visual conventions. Thus, visual identity is informative but insufficient as the primary basis for speaker attribution. SSIC instead induces a reusable speaker identity space from recurring dialogue evidence across the story.

Manga parsing and transcription systems. Recent manga understanding systems have moved from local speaker association toward page- and chapter-level parsing. Fully automated translation pipelines combine text detection, OCR, layout analysis, and translation for cross-lingual manga access (Hinami et al., 2021). The Magi series extends this direction by detecting panels, text boxes, and character boxes, ordering regions, clustering characters, associating dialogue with speakers, and producing structured transcripts (Sachdeva and Zisserman, 2024; Sachdeva et al., 2024; Sachdeva and Zisserman, 2025). Recent benchmarks have further broadened comic understanding beyond individual parsing components. CoMix evaluates multiple tasks including speaker identification, character re-identification, and reading-order prediction (Vivoli et al., 2024), while MangaUB evaluates large multimodal models on both single-panel and multi-panel manga understanding (Ikuta et al.,

2025). While these systems provide strong parsing pipelines, their speaker association remains tied to page-level visual matching and external character banks. SSIC instead builds a reusable book-level speaker identity space from recurring dialogue evidence.

Zero-shot manga speaker prediction. Recent work moves beyond fixed local association by formulating zero-shot character identification and speaker prediction for unseen comics (Li et al., 2024b). The iterative multimodal fusion framework combines LLM-based speaker prediction with image-based character identification and propagates pseudo labels through text-character relationship scores. This establishes an important title-level zero-shot protocol and shows that language reasoning can complement visual character evidence. However, its identity space is still mediated by detected character regions: speaker predictions are converted into pseudo labels for character crops, and updated character labels are propagated back to text regions through local relationship scores. SSIC follows the same title-level zero-shot motivation, but changes the basis of identity induction. Rather than refining speaker labels through visual character regions, SSIC induces anonymous speaker hypotheses from ordered dialogue context, aligns them into cross-window speaker seeds, and constructs a reusable speaker identity registry.

3 Method

SSIC takes manga page images $\mathcal{I}$ as input and predicts a speaker label for each dialogue region. It first converts the pages into an ordered text-layout representation $\mathcal{S}$, which records dialogue regions, text content, page positions, and reading order. SSIC then performs speaker identity consolidation over $\mathcal{S}$ by inducing local anonymous speaker hypotheses, aligning them across overlapping windows into cross-window speaker seeds, using these seeds to construct a global speaker identity registry, and finally assigning each dialogue to a registry entry. The speaker identity space is therefore constructed from recurring dialogue evidence rather than local character-region associations.

3.1 Problem Setting

A test instance is a complete manga book with page images $\mathcal{I}$. SSIC first converts $\mathcal{I}$ into an ordered text-layout sequence $\mathcal{S}$. Each dialogue region d_i in $\mathcal{S}$ is associated with a page index, text-box location,

Algorithm 1 SSIC framework overview

Input: page images $\mathcal{I}$; LLM $\mathcal{M}$
Output: speaker predictions $\hat{\mathcal{Y}}$
1: **Text-layout representation construction**
 $\mathcal{S} \leftarrow \text{PARSEPAGES}(\mathcal{I})$
2: **Speaker seed induction**
 $\mathcal{W} \leftarrow \text{WINDOWS}(\mathcal{S}, w, s)$
 $\mathcal{G} \leftarrow \text{ALIGNSEEDS}(\{\text{LOCALSEEDS}(\mathcal{M}, W_k)\}_{W_k \in \mathcal{W}})$
3: **Global registry construction**
 $\mathcal{R} \leftarrow \text{BUILDREGISTRY}(\mathcal{M}, \mathcal{S}, \mathcal{G})$
4: **Registry-constrained attribution**
 $\hat{\mathcal{Y}} \leftarrow \text{ASSIGNWITHREGISTRY}(\mathcal{M}, \mathcal{S}, \mathcal{R})$
5: **return** $\hat{\mathcal{Y}}$

text content x_i , and reading-order position. No title-specific speaker labels, predefined character identities, or per-title adaptation are available.

The goal is to predict one speaker label $\hat{y}_i$ for each dialogue region d_i . Because the speaker label space is not known in advance, SSIC induces a book-level registry $\mathcal{R}$ and constrains each prediction to refer to one reusable registry entry, so that $\hat{y}_i \in \mathcal{R}$. Registry entries may be explicit names when the story provides them, or stable descriptive placeholders for recurring unnamed speakers. Ground-truth speaker IDs are used only after inference for evaluation.

In the main experiments, $\mathcal{S}$ is constructed from benchmark dialogue regions and transcripts. The end-to-end image-input setting, where $\mathcal{S}$ is produced by a visual front-end, is reported in Appendix A. We use $\mathcal{G}$ for cross-window speaker seeds and $\hat{\mathcal{Y}}$ for the final speaker predictions.

3.2 Framework Overview

Algorithm 1 summarizes the full image-input SSIC pipeline. The first stage constructs the ordered text-layout representation $\mathcal{S}$ from manga pages. Given $\mathcal{S}$, SSIC induces cross-window speaker seeds, builds a book-level speaker identity registry, and performs registry-constrained attribution. Overlapping windows are used for seed induction, where shared dialogue regions enable label alignment.

3.3 Ordered Text-Layout Representation

SSIC uses an ordered text-layout representation $\mathcal{S}$ as the interface between manga page images and speaker identity consolidation. This design follows recent work on layout-aware representation and instruction tuning, where spatial layout and textual content are represented together for downstream reasoning (Feng et al., 2023; Luo et al., 2024). Each dialogue region is serialized with a

dialogue ID, page index, text-box location, text content, and reading-order position. This representation preserves the dialogue order and layout cues needed for cross-page speaker reasoning. In the full image-input pipeline, we use a fine-tuned Qwen3-VL-8B model (Qwen Team, 2025) to detect dialogue text regions and recognize the text in each region. The detected regions are then ordered according to estimated human reading order and converted into the structured representation $\mathcal{S}$. All downstream SSIC stages operate on $\mathcal{S}$.

This representation is used throughout SSIC for speaker seed induction, registry construction, and registry-constrained attribution. It does not include character boxes, character descriptions, or explicit text-to-character spatial links. This design keeps the identity space from being tied to local character-region associations while preserving the dialogue order and layout cues needed for cross-page speaker reasoning.

3.4 Cross-Window Speaker Seed Induction

This stage induces weak identity anchors from local dialogue context. We split each book into overlapping page windows with window size w and stride s . For each window W_k , the LLM receives the ordered text-layout representation and groups the dialogue regions into anonymous speaker hypotheses. The input includes dialogue content, page positions, and reading order, which allow the model to infer local speaker consistency from nearby utterances. These labels are window-local: a label such as speaker 1 in one window is not directly comparable to speaker 1 in another window.

We align local labels using dialogue regions shared by overlapping windows. If two windows W_a and W_b overlap, some dialogue regions appear in both LLM calls. For each shared dialogue region, we record the pair of local labels assigned by the two windows. Repeated co-occurrence of a label pair suggests that the two local labels refer to the same recurring speaker. Intuitively, if two local speaker labels are repeatedly assigned to the same shared dialogue regions in the overlap, they are likely to denote the same speaker. We summarize these co-occurrences with a contingency table $C^{a,b}$:

$$C_{u,v}^{a,b} = \sum_{d \in W_a \cap W_b} \mathbf{1}[z_a(d) = u \wedge z_b(d) = v], \quad (1)$$

where $z_a(d)$ and $z_b(d)$ are the local speaker labels assigned to dialogue region d in windows W_a and

W_b .

We filter candidate links whose co-occurrence count is below a minimum threshold τ and solve a maximum-weight bipartite matching over the remaining links. The matching is partial and does not force every local label to have a correspondence. Matched local labels are linked with a union-find structure, while unmatched labels are retained as separate components rather than discarded. Thus, weak or unsupported links do not force potentially different speakers to be merged, and unmatched speaker hypotheses can still be consolidated later during global registry construction. Repeating this process for all overlapping window pairs yields transitive cross-window clusters of local speaker labels, denoted as speaker seeds $\mathcal{G}$. These seeds are not ground-truth labels or final predictions. They serve as weak identity anchors for registry construction, helping the model consolidate recurring speakers across windows.

3.5 Global Speaker Identity Registry

Speaker seeds connect local speaker hypotheses across windows, but they remain anonymous and cannot serve directly as the final prediction space. As shown in Figure 2, the registry construction stage turns these seed clusters, together with their supporting dialogue evidence, into reusable book-level speaker identity entries.

Each registry entry contains a stable registry ID, an optional explicit name, aliases or descriptive mentions, role and speaking-style cues, and representative utterance IDs. Given the structured text-layout representation $\mathcal{S}$ and the aligned speaker seeds $\mathcal{G}$, the LLM groups recurring dialogue evidence that appears to refer to the same speaker and produces a compact set of reusable speaker identity entries. A registry entry may use an explicit character name when the dialogue provides one, or a stable descriptive placeholder when a recurring speaker remains unnamed.

Registry construction does not use character boxes, character descriptions, or explicit text-to-character links. Thus, the registry is not defined by local character-region associations; it is induced primarily from cross-window speaker seeds and their recurring dialogue evidence.

3.6 Registry-Constrained Attribution

The final stage assigns each dialogue region to an existing registry entry. Given the complete ordered text-layout representation $\mathcal{S}$ and the global registry

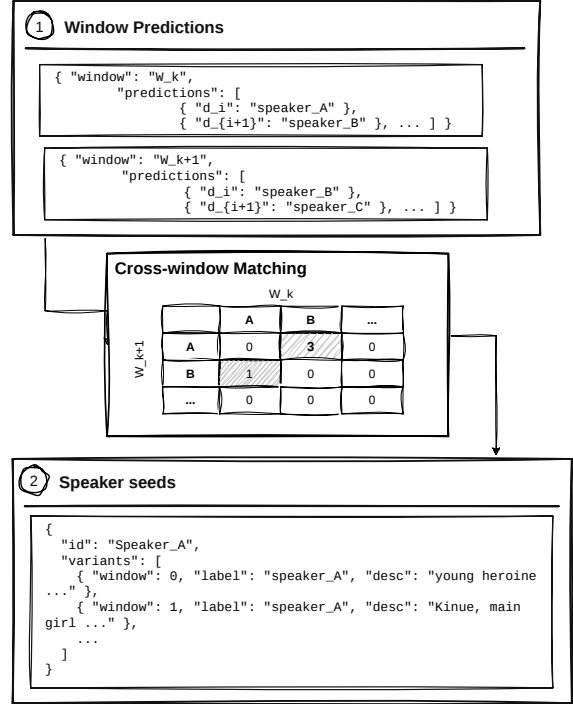


Figure 2: Cross-window matching turns local anonymous speaker hypotheses into speaker seeds, which guide construction of a book-level speaker identity registry from recurring dialogue evidence.

$\mathcal{R}$, the model predicts a speaker for every dialogue region in a constrained fixed registry label space. The input includes dialogue IDs, recognized text, page indices, text-box locations, and reading order, together with the same global registry $\mathcal{R}$.

For each dialogue region, the LLM must select one registry ID as the predicted speaker, and free-form speaker names are not allowed. Because all predictions are made against the same registry, speaker labels remain consistent across the book. The output is the final set of speaker assignments $\hat{\mathcal{Y}}$.

3.7 Zero-Shot Protocol

The entire pipeline is zero-shot with respect to speaker annotations. Ground truth character IDs, body IDs, and evaluation mappings are never used in prompts, registry construction, seed induction, or registry-constrained attribution. The only book-specific identity signal is self-generated, as speaker seeds are produced from the same test book and then used as weak evidence by the registry builder. Ground truth is used only after inference to evaluate the predicted speaker labels.

4 Experiments

We evaluate whether constructing a book-level registry improves zero-shot manga speaker attribution over region-centric and multimodal baselines. The main controlled evaluation uses benchmark dialogue regions and transcripts to isolate identity consolidation from text detection and OCR errors. We use ablation studies and qualitative analysis to examine when and why SSIC improves attribution. An additional end-to-end image-input setting is reported in Appendix A.

4.1 Experimental Setup

Dataset. We evaluate on Manga109Dialog, a speaker-to-text annotation dataset built on Manga109(Aizawa et al., 2020; Li et al., 2024a). Manga109Dialog links dialogue text regions to their speakers, enabling dialogue-level speaker attribution evaluation on Japanese manga. Following the prior zero-shot speaker-prediction protocol, we use 23 volumes for testing. The test set contains 23,896 dialogue regions, with 15 Easy books containing 15,300 regions and 8 Hard books containing 8,596 regions. The Easy/Hard split follows the prior protocol of SGG score and reflects the reliability of local text-character relationship signals.

Task setting. The main controlled evaluation uses benchmark dialogue regions and transcripts from Manga109Dialog to construct the ordered text-layout representation $\mathcal{S}$. This setting follows prior zero-shot speaker-prediction protocols and isolates the central problem of assigning speakers from text detection and OCR errors. No ground-truth speaker identities, character IDs, body IDs, or evaluation mappings are used during inference. We report the stricter end-to-end image-input setting, where benchmark dialogue regions and transcripts are replaced by predicted text boxes and OCR outputs, in Appendix A.

Metric. Predicted speaker IDs are generated independently for each book and need not match the ground-truth character IDs. For each book, we compute a one-to-one Hungarian matching between predicted labels and ground-truth speaker IDs, and report micro-averaged accuracy over dialogue regions. Unless otherwise stated, all scores are speaker Hungarian micro accuracy in percentage points. Ground-truth speaker labels are used only for post-hoc evaluation.

Baselines. We compare with the released baselines from prior zero-shot character identification and speaker-prediction work (Li et al., 2024b). K-means + Distance and K-means + SGG cluster character-region crops within each title and propagate character labels to dialogue regions using either geometric distance or SGG relation scores. Following the released implementation, these clustering baselines use an optimistic post-hoc mapping from visual clusters to ground-truth character labels before speaker labels are propagated to dialogue regions. Iterative multimodal fusion alternates between LLM-based speaker prediction and image-based character identification through text-character relation scores. These baselines are region-centric because they rely on character body regions and local text-character relations, whereas SSIC constructs a registry from ordered dialogue evidence. We also let LLM predict speakers directly from ground-truth dialogue annotations in Manga109.

Implementation. We use GPT-5.4, version 2026-03-05, as the LLM backend for all LLM-based stages of SSIC. We keep the model version and decoding parameters fixed across all experiments, and do not tune sampling settings on the test set. For a fair comparison, we use the same LLM to conduct the LLM direct baseline and we reproduce the iterative multimodal fusion baseline with the same LLM backend, the same benchmark dialogue regions and transcripts, and the same zero-shot restriction. For SSIC, speaker seed induction uses overlapping windows with $w = 10$ pages, stride $s = 5$, and minimum co-occurrence threshold $\tau = 3$. Final registry-constrained attribution is performed over the complete book representation using the same global registry. All prompts and output schemas follow the same zero-shot protocol and do not access ground-truth speaker labels. Full prompt templates and decoding settings are provided in Appendix B.

4.2 Main Results

Table 1 shows that SSIC improves over all baselines. It achieves 74.62% overall speaker accuracy, outperforming iterative multimodal fusion by 7.02 points. The gain is modest on Easy books, where SSIC improves from 68.16% to 72.65%, but much larger on Hard books, where it improves from 66.61% to 78.13%. This 11.52-point gain on Hard books provides the strongest evidence for the

Method	Paradigm	Easy	Hard	Overall
K-means + Distance	Region-centric	41.84	33.83	38.96
K-means + SGG	Region-centric	45.86	35.98	42.31
Iterative multimodal fusion	Region-centric fusion	68.16	66.61	67.60
LLM direct	Implicit registry	64.75	68.54	66.12
SSIC	Registry-centric	72.65	78.13	74.62

Table 1: Main results on the 23-book Manga109Dialog test set. Scores are accuracy (%).

ID	Variant	Acc.	Δ
E0	SSIC	74.62	0.00
E1	Seed induction without layout cues	72.43	-2.19
E2	Registry without speaker seeds	70.10	-4.52

Table 2: Ablation studies on Manga109Dialog. Scores are accuracy (%), and Δ is measured against SSIC.

registry-centric formulation, showing that improvements concentrate where local visual associations are least reliable. Figure 3 shows that the improvement is especially clear on Hard books, where local text-character relations are less reliable. Easy and Hard are defined by the reliability of local text-character relationship signals in the prior protocol. Region-centric baselines depend on geometry, SGG relation scores, character-region clustering, or pseudo-label propagation between text and character regions. When these local signals are unreliable, their errors can directly affect speaker labels and can be reinforced during refinement. SSIC is less dependent on any single local relation because it first consolidates recurring speaker evidence into a registry-defined identity space and then assigns dialogues within that space.

The comparison also suggests that local visual character information is not always beneficial for speaker prediction. Iterative multimodal fusion uses visual character evidence and relation scores, but SSIC improves on it by relying on dialogue continuity and registry construction as the primary identity mechanism. This does not imply that visual information is useless. Rather, noisy local character associations can distract attribution when text-character relations are unreliable.

4.3 Ablation Studies

We ablate the two design choices most directly tied to SSIC: cross-window speaker seeds and layout-aware seed induction. All variants in Table 2 use benchmark dialogue regions and transcripts, as in the main evaluation, so the comparison isolates identity consolidation from text detection and OCR errors.

Removing speaker seeds reduces accuracy from

74.62% to 70.10%, a 4.52-point drop. This shows that cross-window seeds provide useful anchors for registry construction. Without these anchors, the registry builder must infer the book-level identity space directly from dialogue evidence, making it more difficult to maintain consistent recurring identities.

Removing layout cues during seed induction reduces accuracy by 2.19 points. This suggests that ordered dialogue provides a strong identity signal, while page layout and text-box locations further help resolve local ambiguity during seed induction.

Together, these ablations show that both cross-window speaker seeds and layout-aware seed induction contribute to SSIC. Speaker seeds stabilize registry construction, while layout cues improve the quality of the local hypotheses that are later aligned across windows.

Hyperparameter sensitivity. We further examine the sensitivity of SSIC to the window configuration and the cross-window alignment threshold. For the window configuration, we test $(w, s) = (10, 5)$, $(16, 8)$, and $(20, 10)$, all with 50% overlap. Accuracy varies within 1.20 points across these settings. For the alignment threshold, $\tau \in \{1, 2, 3\}$ changes accuracy by at most 0.58 points. These results indicate that SSIC is relatively stable across reasonable window scales and alignment thresholds. Full results are reported in Appendix C.

4.4 Analysis

Qualitative case study. Figure 4 illustrates the failure mode behind the Hard-split gains. The example is from *YasasiiAkuma*, a Hard title where local text-character relations are often ambiguous. The iterative multimodal fusion baseline correctly assigns only 8 of the 20 dialogue regions on this page. Several errors follow visually plausible nearby body regions selected by the SGG relation signal, but these local links are inconsistent with recurring speakers in the surrounding story context.

SSIC correctly attributes all 20 dialogue regions on the same page. Before final attribution, SSIC

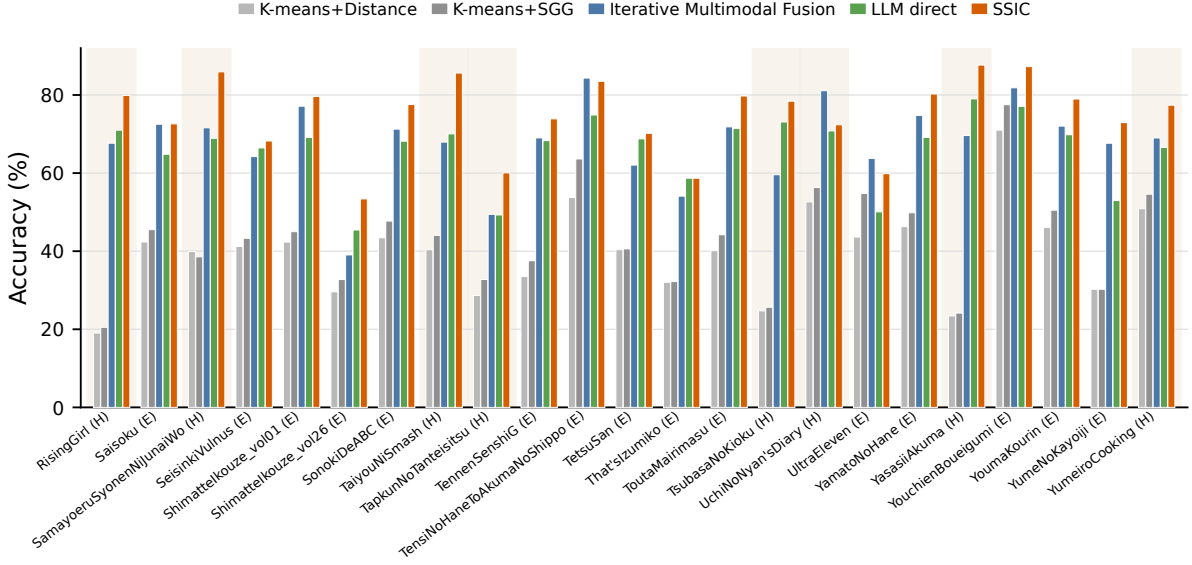


Figure 3: Per-book speaker accuracy on Manga109Dialog. Background shading indicates the Easy/Hard split.

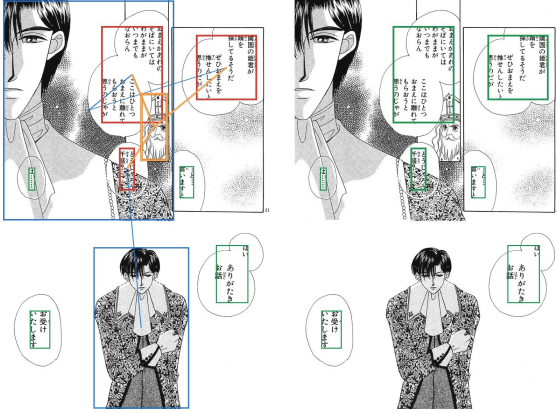


Figure 4: Qualitative example from a Hard book. The left page shows the results of iterative multimodal fusion baseline. The right page shows the SSIC result. Green and red boxes denote correct and incorrect attributions. Blue boxes and lines show the top text-character relation selected by the baseline. Orange boxes and lines show the ground-truth speaker links for the errors.

has already consolidated recurring dialogue evidence into a book-level registry. Each local decision is therefore made within a stable identity space rather than under an unconstrained local relation. This example shows that SSIC does not simply choose a better local region. Instead, it uses accumulated dialogue evidence to override locally plausible but globally inconsistent speaker links.

Effect of text-layout representation on seed quality. We analyze whether structured text-layout information reduces redundant speaker seeds while preserving seed quality. For each Manga109 test ti-

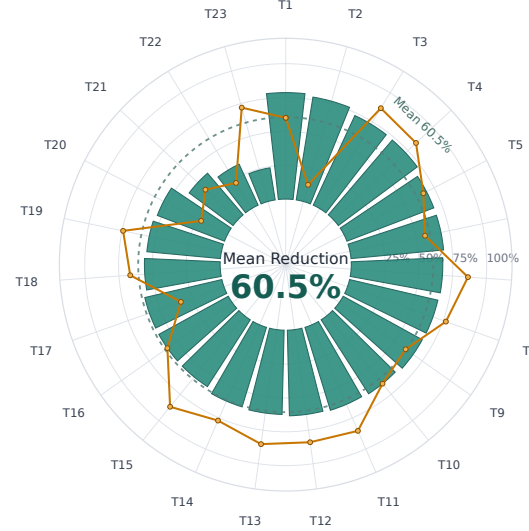


Figure 5: Per-book reduction in local speaker groups after using the structured text-layout representation for seed induction. Bars show the relative reduction for each test title, and the center value reports the mean reduction across the 23 Manga109Dialog test books. The orange line indicates ground-truth characters in the same comparison.

tle, we compare the number of speaker clusters produced by ordered text and the structured text-layout representation. Figure 5 shows that the percentage of local speaker groups is reduced for all 23 test books. The structured representation reduces the number of speaker clusters for all 23 titles. On average, the predicted speaker count decreases from 95.4 to 34.1 per book, corresponding to a mean reduction of 60.5% and a median reduction of 63.2%.

This compression does not come at the cost of

seed quality. The accuracy of Seed-level speaker attribution improves from 0.519 to 0.560, and the accuracy improves on 18 of 23 titles. These results indicate that hierarchical text-layout cues produce fewer redundant local speakers, improving alignment to ground-truth character identities.

Registry errors. Because final attribution is constrained to the induced registry, registry construction can also introduce systematic errors. We consider two failure modes. Merge error occurs when one predicted registry entry contains dialogue evidence from multiple ground-truth characters. While split error occurs when one ground-truth character is represented by multiple registry entries. Across the 23 books, 36 of 323 predicted registry entries (11.15%) contain merge errors, while 38 of 610 ground-truth characters (6.23%) are split across multiple entries. Merge errors impose a lower bound of 5.77% on dialogue-attribution errors, while split fragmentation affects 5.92% of dialogue instances. These results show that registry construction remains an important source of error even after cross-window consolidation. Together with the 4.52-point drop observed when speaker seeds are removed, this analysis highlights both the importance of constructing a stable identity space and the remaining difficulty of consolidating sparse or ambiguous speaker evidence.

4.5 Backend and Input Robustness

Open-weight LLM backend. The main experiments use GPT-5.4, raising the question of whether SSIC is tied to a proprietary LLM. We therefore evaluate SSIC with the open-weight DeepSeek-V3.2 model under the same zero-shot protocol. As shown in Table 3, DeepSeek-V3.2 achieves 67.26% speaker attribution accuracy, compared with 74.62% for GPT-5.4. This result shows that SSIC can operate with an open-weight backend, although its absolute performance remains dependent on the capability of the underlying LLM.

Model	Type	Acc.
DeepSeek-V3.2	Open-weight	67.26
GPT-5.4	Proprietary	74.62

Table 3: Speaker attribution accuracy (%) of SSIC with different LLM backends.

End-to-end input robustness. The controlled experiments above use benchmark dialogue regions and transcripts to isolate identity consolidation

from visual parsing errors. We further compare SSIC with IMF when both methods use the same predicted dialogue boxes and OCR outputs, while IMF retains the scene-graph relations required by its original pipeline. Under these noisy inputs, SSIC achieves 56.72% speaker attribution accuracy over all ground-truth dialogue regions, compared with 54.42% for IMF. Both methods degrade substantially relative to the controlled setting, confirming that visual page parsing remains an important bottleneck. Nevertheless, SSIC retains a 2.30-point advantage under the same front-end predictions. Detailed evaluation protocols and detection statistics are provided in Appendix A.

5 Conclusion

We presented Speaker-Seeded Identity Consolidation (SSIC), a zero-shot framework that treats manga speaker attribution as identity consolidation over structured representations. Rather than relying on local text-character region associations as the main reasoning mechanism, SSIC constructs a book-level speaker registry from cross-window speaker seeds and performs final attribution within the registry-defined identity space.

Experiments on Manga109Dialog show that this registry-centric formulation improves over region-centric and iterative multimodal fusion baselines. SSIC achieves 74.62% speaker attribution accuracy, improving the strongest prior baseline by 7.02 points. The gains are especially clear on books where local text-character relationships are unreliable, and ablations show that cross-window speaker seeds are the key component for constructing a useful registry.

These findings suggest that the bottleneck in zero-shot manga speaker attribution is not only local visual association, but also the construction of a stable identity space for an unseen story. Once pages are serialized into an ordered representation, dialogue semantics, recurrence, and registry constraints can provide useful evidence for speaker identity. This points toward future manga understanding systems that combine robust visual page parsing with explicit book-level identity reasoning, rather than treating speaker attribution as a collection of independent local matching decisions.

Limitations

Our main experiments evaluate speaker attribution after dialogue regions and text content are avail-

able. This controlled setting follows prior zero-shot speaker prediction protocols and isolates identity consolidation from page parsing errors, but it does not remove the need for reliable page parsing. In the end-to-end image-input setting reported in Appendix A, speaker attribution degrades when the ordered text-layout representation is produced by the visual front-end. This gap reflects the additional difficulty of constructing the representation from images, where missed dialogue regions, OCR errors, reading-order errors, and serialization noise can affect downstream identity consolidation.

SSIC also depends on the quality of LLM-based identity consolidation. When a book contains many minor speakers, repeated role names, sparse dialogue, or abrupt scene changes, the registry may merge distinct speakers or split one speaker into multiple entries. Because final attribution is constrained by the registry, such errors can propagate to multiple dialogue instances.

Finally, our evaluation is limited to the 23-book Manga109Dialog test set and to Japanese manga. Broader validation is needed across languages, genres, lettering styles, scan quality, and longer series with larger casts. Future work should also explore how to incorporate visual layout and character appearance without returning to brittle local text-character matching.

Ethical Considerations

This work uses existing manga research benchmarks for speaker attribution. Manga pages, dialogue, and character identities are copyrighted creative works, so any release of code, prompts, intermediate outputs, or qualitative examples should respect the access conditions and licenses of the underlying datasets. We do not redistribute manga images or transcripts beyond what is permitted by the dataset providers.

Automatic speaker attribution can support various downstream tasks like translation or dubbing, but incorrect speaker assignments may misattribute dialogue, alter character consistency, or affect downstream generation systems such as dubbing and derivative content creation. Such systems should be used with human review in public or creative settings.

References

Kiyoharu Aizawa, Azuma Fujimoto, Atsushi Otsubo, Toru Ogawa, Yusuke Matsui, Koki Tsubota, and

Hikaru Ikuta. 2020. [Building a manga dataset “manga109” with annotations for multimedia applications](#). *IEEE MultiMedia*, 27(2):8–18.

David Dubray and Jochen Laubrock. 2019. [Deep cnn-based speech balloon detection and segmentation for comic books](#). In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1237–1243.

Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2023. [LayoutGPT: Compositional visual planning and generation with large language models](#). In *Advances in Neural Information Processing Systems*.

Hua He, Denilson Barbosa, and Grzegorz Kondrak. 2013. [Identification of speakers in novels](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1312–1320, Sofia, Bulgaria. Association for Computational Linguistics.

Ryota Hinami, Shonosuke Ishiwatari, Kazuhiko Yasuda, and Yusuke Matsui. 2021. [Towards fully automated manga translation](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Hikaru Ikuta, Leslie Wöhler, and Kiyoharu Aizawa. 2025. [MangaUB: A manga understanding benchmark for large multimodal models](#). *IEEE MultiMedia*, 32(2):33–43.

Yingxuan Li, Kiyoharu Aizawa, and Yusuke Matsui. 2024a. [Manga109Dialog: A large-scale dialogue dataset for comics speaker detection](#). In *Proceedings of the IEEE International Conference on Multimedia and Expo*.

Yingxuan Li, Ryota Hinami, Kiyoharu Aizawa, and Yusuke Matsui. 2024b. [Zero-shot character identification and speaker prediction in comics via iterative multimodal fusion](#). In *Proceedings of the ACM International Conference on Multimedia*.

Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. 2024. [LayoutLLM: Layout instruction tuning with large language models for document understanding](#). *arXiv preprint arXiv:2404.05225*.

Grace Muzny, Michael Fang, Angel X. Chang, and Dan Jurafsky. 2017. [A two-stage sieve approach for quote attribution](#). In *European Chapter of the Association for Computational Linguistics*.

Nhu-Van Nguyen, Christophe Rigaud, and Jean-Christophe Burie. 2019. [Comic MTL: optimized multi-task learning for comic book image analysis](#). *Int. J. Doc. Anal. Recognit.*

Qwen Team. 2025. [Qwen3-VL technical report](#). *arXiv preprint arXiv:2511.21631*.

Christophe Rigaud, Nam Le Thanh, J.-C. Burie, J.-M. Ogier, Motoi Iwata, Eiki Imazu, and Koichi Kise. 2015. [Speech balloon and speaker association for comics and manga understanding](#). In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 351–355.

Ragav Sachdeva, Gyungin Shin, and Andrew Zisserman. 2024. [Tails tell tales: Chapter-wide manga transcriptions with character names](#). In *Proceedings of the Asian Conference on Computer Vision*.

Ragav Sachdeva and Andrew Zisserman. 2024. [The manga whisperer: Automatically generating transcriptions for comics](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Ragav Sachdeva and Andrew Zisserman. 2025. [From panels to prose: Generating literary narratives from comics](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Koki Tsubota, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2018. [Adaptation of manga face representation for accurate clustering](#). In *ACM SIGGRAPH Asia 2018 Posters*, pages 15:1–15:2.

Emanuele Vivoli, Marco Bertini, and Dimosthenis Karatzas. 2024. [Comix: A comprehensive benchmark for multi-task comic understanding](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 140828–140846.

Ci-Yin Zhang and Wei-Ta Chu. 2023. [Occlusion-aware manga character re-identification with self-paced contrastive learning](#). In *Proceedings of the 5th ACM International Conference on Multimedia in Asia*.

Yunlong Zhang and Seiji Hotta. 2025. [Comic speaker prediction based on visual relations and natural language processing tasks](#). *IEEE MultiMedia*, 32(3):50–59.

Zhimin Zhang, Zheng Wang, and Wei Hu. 2022. [Unsupervised manga character re-identification via face-body and spatial-temporal associated clustering](#). *arXiv preprint arXiv:2204.04621*.

Yi Zheng, Yifan Zhao, Mengyuan Ren, He Yan, Xiangju Lu, Junhui Liu, and Jia Li. 2020. [Cartoon face recognition: A benchmark dataset](#). In *Proceedings of the 28th ACM International Conference on Multimedia*.

Wenjie Zhong, Jason Naradowsky, Hiroya Takamura, Ichiro Kobayashi, and Yusuke Miyao. 2024. [Who said what: Formalization and benchmarks for the task of quote attribution](#). In *LREC-COLING*.

A End-to-End Image-Input Evaluation

The main experiments use benchmark dialogue regions and transcripts to construct the ordered text-layout representation $\mathcal{S}$. We additionally evaluate

an end-to-end setting in which a fine-tuned Qwen3-VL-8B predicts dialogue regions and text content from manga page images. The predicted regions are ordered and serialized into $\mathcal{S}$ before speaker attribution.

Detected dialogue regions are matched to benchmark regions on the same page using one-to-one IoU matching with a threshold of 0.5. Unmatched benchmark dialogue regions are counted as incorrect speaker predictions. The visual front-end achieves 78.15% precision, 84.15% recall, and 81.04% F1 for dialogue-region detection.

To examine whether SSIC remains effective under the same front-end noise, we also evaluate iterative multimodal fusion (IMF) using the same predicted dialogue boxes and OCR outputs, while retaining the scene-graph relations required by its original pipeline.

Method	Controlled	End-to-end
IMF	67.60	54.42
SSIC	74.62	56.72

Table 4: Speaker attribution accuracy (%) under controlled benchmark input and end-to-end visual-front-end input.

As shown in Table 4, both methods are substantially affected by front-end errors. SSIC decreases from 74.62% to 56.72%, while IMF decreases from 67.60% to 54.42%. Nevertheless, SSIC remains 2.30 points more accurate than IMF when both methods use the same predicted dialogue regions and OCR outputs. These results show that visual page parsing remains an important bottleneck, while the advantage of SSIC persists under noisy end-to-end inputs.

B Prompt Templates and Decoding Settings

B.1 Decoding Settings and Shared Input Format

We report the prompt templates used by SSIC and the direct LLM baseline, with book-specific input replaced by placeholders. All prompts follow the zero-shot protocol and contain no ground-truth speaker labels, character IDs, body IDs, or evaluation mappings.

All calls use a system message and a user message. We use the same GPT-5.4 deployment for all stages and keep the default sampling parameters. The maximum completion length is 65,536 tokens

for local seed extraction and final attribution, and 32,768 tokens for registry construction.

The shared text-layout representation uses the following fields.

```
{
  "page": "{PAGE_ID}",
  "panel_id": "{PANEL_ID}",
  "reading_order": "{READING_ORDER}",
  "text_id": "{TEXT_ID}",
  "text_bbox": "{TEXT_BBOX}",
  "content": "{TEXT}"
}
```

B.2 Local Speaker Hypothesis and Seed Artifacts

The local speaker hypothesis stage groups dialogue lines within each sliding window into anonymous speaker labels. This stage focuses on recurring main speakers within the window; minor, unclear, narrative, and non-dialogue text is grouped into others to avoid creating unstable seed clusters. The reported seed-based SSIC runs consume the merged seed artifact obtained after seed alignment based on deterministic overlap.

System message.

```
You are an expert in manga dialogue analysis.
The input is a multi-page window of manga text lines
ordered by page and reading order.
For each window, distinguish only the main speakers and
group all other text lines as others.

Rules:
1. Each line has global_order, page, order, id, bbox, and
   content.
2. Use speaker_A to speaker_J only for a small number of
   main speakers in the current window.
   Use at most 10 speaker labels, and only use as many as
   needed.
3. Use others for all non-main speakers, minor characters,
   mobs, short shouts,
   narration, sound effects, onomatopoeia, and unclear
   voices.
   Do not use separate narrator or sfx labels at this
   stage.
4. speaker_A to speaker_J should refer to individual
   speakers whenever possible.
   If a speaker is unnamed but can be tracked from wording,
   context, bbox/location,
   addressee, and reading order, assign that speaker to a
   speaker_X label.
5. Do not create speaker_X labels for mixed voices, teams,
   factions, role categories,
   locations, audiences, crowds, narration, sound effects,
   or unknown voices.
   Put them in others unless they can be tracked as an
   individual speaker.
6. Use the same speaker_X label for the same person within
   the window, even across pages.
7. As a soft guard against assigning nearly all lines to
   others, prefer assigning recurring
   individual speakers to speaker_A to speaker_J when
   there is enough evidence.
   Do not create extra labels only to satisfy this
   preference.

Return JSON only.
{
  "speakers": {
```

```
    "speaker_A": "brief description of the main speaker",
    "others": "all non-main speakers without further
              subdivision"
  },
  "assignments": [
    {
      "id": "{TEXT_ID}",
      "page": "{PAGE_ID}",
      "global_order": "{GLOBAL_ORDER}",
      "speaker": "speaker_A"
    }
  ]
}
```

The speakers object may contain only speaker_A to speaker_J and others. Every input line must appear exactly once in assignments. The id, page, and global_order fields must match the input.

User message template.

```
{
  "task": "Cluster speakers within this sliding window.",
  "input": "{WINDOW_TEXT_LINES_IN_PAGE_AND_READING_ORDER}",

  "line_format": {
    "page": "{PAGE_ID}",
    "panel_id": "{PANEL_ID}",
    "reading_order": "{READING_ORDER}",
    "text_id": "{TEXT_ID}",
    "text_bbox": "{TEXT_BBOX}",
    "content": "{TEXT}"
  },
  "output": "Return speakers and assignments for every
            input line."
}
```

Merged seed artifact. Seed merging itself is deterministic and does not use an additional LLM prompt. The merged artifact has the following structure.

```
{
  "book": "{BOOK}",
  "window_index": "{WINDOW_INDEX}",
  "pages": ["{PAGE_ID}", "..."],
  "dialogues": [
    {
      "page": "{PAGE_ID}",
      "panel_id": "{PANEL_ID}",
      "reading_order": "{READING_ORDER}",
      "text_id": "{TEXT_ID}",
      "text_bbox": "{TEXT_BBOX}",
      "content": "{TEXT}"
    }
  ],
  "speakers": {
    "local_speaker_label": "brief local speaker
                           description"
  },
  "assignments": [
    {
      "id": "{TEXT_ID}",
      "page": "{PAGE_ID}",
      "global_order": "{GLOBAL_ORDER}",
      "speaker": "local_speaker_label"
    }
  ]
}
```

B.3 Speaker Identity Registry Construction Prompt

The registry construction prompt receives the structured text-layout evidence and the aligned speaker seeds. Its goal is to build a compact book-level speaker identity registry. Speaker seeds are used only as weak cross-window evidence and are not treated as ground-truth identities.

System message.

You build a zero-shot global speaker identity registry for manga speaker attribution.

The input is a structured text-layout representation serialized from manga pages: page ids, panel ids when available, reading order, dialogue text ids, text bounding boxes, text content, and explicit name or role mentions when available.

The input does not include character boxes, visual character descriptions, ground-truth speaker labels, character IDs, body IDs, evaluation mappings, or text-character spatial links.

Speaker seeds are noisy weak evidence from local window-level speaker inference and overlap-based merging. They are not ground-truth identities.

A seed may mix multiple speakers or split one speaker into several local descriptions.

Use seeds as soft anchors only when supported by dialogue evidence.

Create stable, reusable speaker identity ids for registry-constrained speaker assignment.

Prefer explicit names when clear; otherwise use stable descriptive placeholders.

Do not create regular registry entries for narrator, sfx, crowd, generic groups, or purely unknown voices; use special labels for them.

Return JSON only.

User message template.

```
{
  "task": "Build a book-level speaker identity registry.",
  "book": "{BOOK}",
  "evidence_mode": "text_layout",
  "name_mentions": "{NAME_MENTION_COUNTS}",
  "speaker_seeds": "{SEED_SPEAKERS}",
  "text_layout_evidence": "{TEXT_LAYOUT_EVIDENCE}",
  "output_schema": {
    "format": "global_speaker_identity_registry_v1",
    "book": "{BOOK}",
    "speakers": [
      {
        "id": "spk_001",
        "canonical_name": "
explicit_name_or_stable_placeholder",
        "aliases": ["optional"],
        "description": "concise evidence-grounded identity
cues",
        "speaking_style": "dialogue or voice clues",
        "representative_lines": ["short text snippets"]
      }
    ],
    "special_ids": ["narrator", "sfx", "crowd", "others",
"unknown"]
  },
  "limits": {
    "max_speakers": 80,
    "id_format": "spk_001, spk_002, ..."
  }
}
```

```
}
```

B.4 Registry-Constrained Attribution Prompt

The final attribution prompt receives the speaker identity registry and the structured text-layout scene. The model must assign each dialogue text to one existing registry ID or one allowed special ID. It is not allowed to create new speaker IDs during final attribution.

System message.

You are doing zero-shot manga speaker attribution with a fixed speaker identity registry.

For every dialogue text, assign exactly one speaker id from the registry or one allowed special id.

Use reading order, text content, page-panel grouping, text-box geometry, direct address, names, speech style, and narrative context.

If a speaker is not represented in the registry, use others or unknown.

Do not create new speaker ids during final attribution.

Every input text id must appear exactly once in assignments.

Return compact JSON only.

Do not output confidence, evidence, character_refs, new_speakers, notes, pages, or speaker lists.

User message template.

```
{
  "task": "Assign every text line to one registry id or
special id.",
  "book": "{BOOK}",
  "registry": "{GLOBAL_SPEAKER_IDENTITY_REGISTRY}",
  "scene": "{TEXT_LAYOUT_SCENE}",
  "output_schema": {
    "assignments": [
      {
        "t": "t00000000",
        "speaker": "spk_001_or_special_id"
      }
    ]
  },
  "constraints": {
    "cover_every_input_text_once": true,
    "speaker_space": "registry ids plus special ids only",
    "special_ids": ["crowd", "narrator", "others", "sfx",
"unknown"],
    "return_json_only": true
  }
}
```

B.5 Direct LLM Baseline Prompt

The direct LLM baseline uses the same Manga109Dialog benchmark dialogue regions and transcripts as the main controlled evaluation. Unlike SSIC, it does not use cross-window speaker seeds, a speaker identity registry, or registry-constrained attribution. Instead, it directly asks the LLM to infer anonymous book-level speakers and assign every dialogue text to one speaker.

System message.

You are doing zero-shot manga speaker attribution from one complete book represented as text boxes.

The input contains page order, text ids, text boxes, and text content.

No images, character boxes, visual descriptions, SSIC registry, speaker seeds, ground-truth speaker labels, character IDs, body IDs, evaluation mappings, previous predictions, or repaired reading order are provided.

Infer a stable anonymous speaker list for the book and assign one speaker to every dialogue text.

Use the same speaker id for the same recurring speaker across the whole book.

Use special ids for narration, sound effects, crowds, minor unclear voices, and unknown speakers.

Return compact JSON only.

User message template.

```
{
  "task": "Infer anonymous book-level speakers and assign every text line.",
  "book": "{BOOK}",
  "text_layout": "{FULL_BOOK_TEXT_LAYOUT}",
  "input_fields": ["page", "order", "id", "bbox", "text"],
  "special_ids": ["crowd", "narrator", "others", "sfx", "unknown"],
  "output_schema": {
    "speakers": {
      "speaker_001": "brief stable description inferred from the whole book",
      "others": "minor unclear individual voices"
    },
    "assignments": [
      {
        "page": "{PAGE_ID}",
        "id": "{TEXT_ID}",
        "speaker": "speaker_001_or_special_id"
      }
    ]
  },
  "constraints": {
    "cover_every_input_text_once": true,
    "speaker_space": "speaker_001, speaker_002, ... plus special ids",
    "return_json_only": true
  }
}
```

C Hyperparameter Sensitivity

We evaluate the sensitivity of SSIC to two parameters used in cross-window speaker seed induction: the window configuration (w, s) and the minimum co-occurrence threshold τ . The results show that SSIC remains stable across the tested configurations.

C.1 Window Size and Stride

SSIC divides each book into overlapping windows during local speaker hypothesis induction. We evaluate three window configurations, $(w, s) = (10, 5)$, $(16, 8)$, and $(20, 10)$, all with 50% overlap between consecutive windows.

As shown in Table 5, the three configurations

achieve 74.62%, 75.82%, and 75.31% accuracy, respectively. The maximum difference is only 1.20 percentage points. This limited variation indicates that SSIC is relatively robust to the choice of window scale within the tested range.

Window w	Stride s	Acc.
10	5	74.62
16	8	75.82
20	10	75.31

Table 5: Sensitivity to window size and stride. Scores are speaker attribution accuracy (%). All configurations use 50% overlap.

C.2 Cross-Window Alignment Threshold

We also evaluate the sensitivity of SSIC to the minimum co-occurrence threshold τ used for cross-window speaker alignment. The threshold filters weak candidate links between local speaker labels before bipartite matching.

Table 6 reports results for $\tau \in \{1, 2, 3\}$. Overall accuracy ranges from 74.04% to 74.62%, with a maximum variation of only 0.58 percentage points. Performance on both the Easy and Hard subsets also remains stable across the tested values. These results indicate that SSIC is relatively insensitive to the precise choice of the alignment threshold.

τ	Easy	Hard	Overall
1	72.79	76.61	74.16
2	71.99	77.68	74.04
3	72.65	78.13	74.62

Table 6: Sensitivity to the minimum co-occurrence threshold τ . Scores are speaker attribution accuracy (%).