CANVAS: A Multimodal Dataset of Chinese Textbook Images for Bias and Representation Analysis

Haotian Zhu[1], **Kefan Yu**[2], **Min Li**[2]
[1]Department of Linguistics, University of Washington, Seattle, WA, U.S.A.
[2]Department of Education, University of Washington, Seattle, WA, U.S.A.
{haz060, ky285, minli}@uw.edu

Abstract

Social biases in educational materials can subtly shape students' perceptions of social roles and participation. However, most existing bias benchmarks for Chinese language models focus on text or isolated images, overlooking the multimodal scenes commonly found in educational textbooks. To address this gap, we introduce **CANVAS** (**C**hinese **AN**notated **V**isual **A**nd **S**ocial scenes), a multimodal dataset constructed from Chinese elementary science textbooks and annotated across multiple social dimensions. **CANVAS** provides fine-grained labels for each depicted character's demographics, social roles, interactions, and power-related attributes within visual scenes. The dataset is created using a semi-automated pipeline in which a vision-language model generates preliminary structured annotations that are subsequently verified and refined by human annotators. The current release focuses on the Grade 6 science subset and serves as an initial annotated version of the dataset. Using this subset, we present an illustrative case study demonstrating how scene-level and interactional annotations in **CANVAS** can be used to analyze gender representation in textbook images. By extending bias analysis to full educational scenes, **CANVAS** provides a new resource for studying representation and fairness in multimodal educational materials and supports future research in NLP, computer vision, and education.

Keywords: bias analysis, multimodal dataset, educational texts, Chinese textbooks, fairness in AI

1. Introduction

Textbooks and their images do more than teach academic content; they subtly communicate ideas about social roles, power dynamics, and who belongs in certain spaces. Research on Chinese elementary school materials shows that male characters are often more visible than female characters and are frequently depicted as active problem-solvers, while female characters are more likely to appear as observers or supporting participants (Sang, 2023). Such portrayals may influence students' perceptions of what is "normal," potentially shaping aspirations and self-efficacy.

Over the past decade, the NLP and computer vision communities have become increasingly aware that biased training data can propagate harmful stereotypes (Bolukbasi et al., 2016). Several benchmarks have been proposed to study social biases in Chinese language models, including CHBias (Zhao et al., 2023) and CBBQ (Huang and Xiong, 2024). However, these resources primarily evaluate textual model behavior or chatbot responses rather than multimodal educational materials. Other datasets address biases in different domains: McBE (Lan et al., 2025) provides multiple-choice scenarios for moral and bias evaluation, and vision-oriented datasets such as FairFace (Kärkkäinen and Joo, 2021) and IllusFace 1.0/CBFeatures 1.0 (Szasz et al., 2022) collect images with demographic labels. While useful for studying representation in visual data, these datasets focus on isolated faces or short visual snippets rather than full classroom scenes. The VISOGENDER benchmark (Hall et al., 2023) extends this line of work by evaluating gender balance in pronoun resolution tasks using occupational images, yet it still does not capture the complex social interactions present in educational illustrations.

To our knowledge, existing resources do not capture the interplay of demographics, activities, and interaction dynamics that occur in classroom scenes. Many bias benchmarks analyze text-based prompts or isolated images, but they do not represent the multimodal content students encounter daily in educational materials, such as who initiates discussions, who controls access to learning materials, or who becomes the focus of attention within a scene.

To address this gap, we introduce **CANVAS** (**C**hinese **AN**notated **V**isual **A**nd **S**ocial scenes), a multimodal dataset constructed from Chinese elementary science textbooks. **CANVAS** systematically annotates textbook illustrations across multiple dimensions, including character demographics, social roles, activities, interactions, and power-related attributes. Each depicted character is labeled with attributes such as age, gender, and role, along with interactional features such as conversational initiation, access to artifacts, and visual attention patterns within the scene.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9226–9239
11–16 May 2026.

Scope of the current release. The current version of **CANVAS** focuses on the Grade 6 science textbook subset and serves as an initial annotated release of the dataset. The empirical analyses presented in this paper are therefore limited to this subset and are intended to illustrate the types of scene-level and interactional analyses enabled by the dataset rather than to draw definitive conclusions about all Chinese primary-school textbooks.

Resource contribution. The primary contribution of this work is the **CANVAS** dataset and its annotation framework. By providing structured labels for demographics, social roles, interactions, and power dynamics in textbook illustrations, **CANVAS** enables analyses of representation that are difficult to conduct using existing bias benchmarks. The case study presented in this paper serves as an illustrative example of how the dataset can support structured investigations of representation in multimodal educational materials.

2. Related Work

Bias in educational materials can reinforce stereotypes and shape learners' perceptions of social roles. We review prior work on bias benchmarks for Chinese language models, fairness datasets for vision and multimodal tasks, and studies examining representation in educational materials.

2.1. Chinese bias benchmarks for language models

Early efforts to evaluate social biases in Chinese language models focused on text-based benchmarks such as CHBias (Zhao et al., 2023), CBBQ (Huang and Xiong, 2024), and McBE (Lan et al., 2025). These resources evaluate bias across multiple social dimensions using question-based or dialogue-based tasks. While valuable for studying textual bias in Chinese LLMs, they focus on language model outputs rather than multimodal educational materials.

2.2. Fairness datasets for vision and multimodal tasks

Several datasets examine bias in visual and vision-language systems. FairFace (Kärkkäinen and Joo, 2021) provides a large-scale dataset of facial images with demographic labels to study fairness in face recognition. IllusFace 1.0 and CBFeatures 1.0 (Szasz et al., 2022) focus on illustrated characters in children's books and provide demographic labels for faces in illustrations. VISOGENDER (Hall et al., 2023) introduces a benchmark for gender bias in image–text pronoun resolution tasks. Although these datasets address fairness in visual or multimodal contexts, they typically focus on isolated faces or occupational scenes rather than full educational interactions.

2.3. Representation studies in children's books and textbooks

Recent work has examined representation in children's literature and educational materials. For example, Adukia et al. (2023) analyzed over 1,000 picture books using computer vision methods to detect and classify characters, revealing systematic underrepresentation of women and darker-skinned characters. Studies of Chinese elementary textbooks similarly report gender imbalances in character visibility and activity roles (Sang, 2023). These findings highlight the importance of analyzing both visual representation and contextual roles in educational materials.

2.4. Gaps in prior literature

Despite these advances, existing studies typically emphasize textual content or isolated visual elements rather than multimodal educational scenes. For example, Islam and Asadullah (2018) analyzed gender stereotypes in textbooks primarily through text-based indicators such as names and pronouns, with limited attention to visual interactions. Similarly, Susanti et al. (2021) found gender imbalances in Indonesian EFL textbooks but focused mainly on frequency counts across text and images.

However, simple frequency counts may not fully capture representational bias. Contextual factors such as who initiates discussions, who controls learning materials, or who becomes the center of attention in a scene can also shape perceptions of authority and participation. These interactional dimensions remain largely underexplored in existing datasets. Our work addresses this gap by introducing a multimodal dataset that systematically annotates characters, activities, interactions, and power-related attributes within textbook scenes.

3. Resource Description

In this section, we describe the data sources, dataset statistics, and annotation schema used to construct **CANVAS**.

3.1. Data source

In China's elementary education system, different regions use textbooks from various publishers. We obtained Mandarin Chinese science textbooks in PDF format from the official websites of several major publishers and downloaded and

preprocessed the files for annotation. The publishers include: PEP (People's Education Press)[1], Jiangsu (Jiangsu Education Press)[2], ESE (Education Science Electron Publishing House)[3], Daxiang (Daxiang Press)[4], Guangdong (Guangdong Education Press), Hebei (Hebei Education Press)[5], Hunan (Hunan Science Technology Publishing House)[6], and Qingdao63 (Qingdao Press Edition 63)[7]. These sources provide broad coverage of science curricula used across multiple regions in China.

3.2. Resource details

Our collection consists of 96 science textbooks spanning Grades 1–6 from the eight publishers listed above. The current release of **CANVAS** includes annotations for the Grade 6 textbooks only. In Grade 6, there are 16 textbooks totaling 1,213 pages. Of those pages, 539 contain images with human figures, and we identified 3,359 individual people in these images (see Table 1 for a summary). Future releases will extend the annotations to additional grade levels.

# of total grades	6
# of publishers	8
# of total textbooks	96
# of total pages in Grade 6	1,213
# of pages in Grade 6 that contain people	539
# of people identified in Grade 6 images	3,359

Table 1: Basic statistics of the dataset. The top portion gives overall counts across the collected textbooks, while the bottom portion corresponds to the annotated Grade 6 subset.

3.3. Annotation schema

Our annotation schema is informed by sociocultural theories of learning and interaction (Vygotsky, 1978; Bruner, 1986; Rogoff, 2003), which emphasize that learning activities are embedded in social interactions and shared environments.

Guided by the sociocultural perspective, our coding schema includes five main modules that correspond to: (1) **demographics of actors**; (2) **social setting and activity**; (3) **interaction format**; (4) **artifacts and access**; and (5) **power dynamic pattern**.

Module 1: The coding module of demographics (i.e., Module 1) is intended to identify and offer detailed descriptive labels for each actor appearing in an image, including gender, age, ethnicity, disability status, etc., along with a unique ID that can be referenced by and linked to the codes assigned in other modules.

Module 2 and 3: The next two modules (i.e., Modules 2 and 3), social setting and activity as well as interaction format, are designed to describe the nature of activity an actor engages with and portrait the setting or surrounding environment where the activity takes place (e.g., using a lawn mower to mow the front yard of a house, making musical instruments, eating the breakfast with other family members, brainstorming ideas with peers in the classroom, running a survey with residents in a neighborhood). Within those two modules, actors' social identity will be labeled as well based on the nature of activities and events that they engage with; for example, a teenage actor can be labeled as child in an image that shows they are shopping at a grocery store with their parents or a student in another image when a group of students are modeling the locations of the Sun, Earth, and Moon.

Module 4 and 5: The coding questions in the last two modules are adapted and phrased based on coding decisions from Modules 2 and 3 because artifacts, interaction norms, and group dynamics are dependent on the nature and setting of events and activities (e.g., (Wang et al., 2019)). The definitions and examples of coding categories under those modules are informed by the relevant literature and then further refined and elaborated during the training, tryout, and calibration of the coders when implementing the coding procedures. The initial training and tryout phases involved five coders, among whom three continued with the calibration phase and inter-coder reliability studies.

Module 5 focuses on identifying power dynamics in gender interactions, which are determined through participants' positioning, roles, and dialogues during group activities. These indicators can reveal an imbalance of participation or the absence of empowerment for a particular gender group. For example, Figure 1 presents four images of boys and girls participating in a handcraft project. Across these images, boys appear to take on the central crafting tasks such as gluing materials, cutting components, and assembling structures; while girls are more frequently positioned in peripheral roles such as holding materials, observing, or helping others.

[1] https://www.pep.com.cn/
[2] https://www.1088.com.cn/about.html
[3] https://www.esph.com.cn/index.htm
[4] https://www.daxiang.cn/
[5] https://www.hbep.com/Home/Index
[6] https://hnstp.com/jiaocai.html
[7] https://www.qingstudy.cn/

This distribution of roles suggests a pattern consistent with traditional gendered divisions of labor, where male students are positioned as the primary actors responsible for technical or physically demanding tasks. In contrast, female students appear less engaged in decision-making or construction activities. Such visual representations can indicate unequal participation and implicit power hierarchies within the group.

Figure 1: Screenshot of textbook image from the Jiangsu Press.

Power dynamics can also be shown in verbal interactions and positions. For instance, Figure 2 shows a group discussion about how to design a raincoat. During the discussion, the girl at the center of the group proposes a new idea for improving the design. Her active contribution demonstrates that power relations are not fixed and may shift depending on the interaction context. In this case, the girl takes an initiative role by introducing suggestions that influence the group's decision-making process. This example illustrates how dialogue can serve as an important indicator of empowerment within group activities.

4. Creation Methodology

This section describes the pipeline used to construct **CANVAS**, including automated preprocessing, human verification, and manual annotation procedures.

4.1. Annotation pipeline

The dataset construction pipeline consists of four stages: (1) data preprocessing, (2) automated annotation, (3) human correction, and (4) manual labeling, as illustrated in Figure 3.

Figure 2: Screenshot of textbook image from the Daxiang Press.

4.1.1. Data preprocessing

After downloading the original textbook PDFs from publisher websites, each page is converted into an image. These images are used as the input for automated annotation and subsequent human verification steps.

4.1.2. Automated annotation

The automated stage generates preliminary structured annotations using a vision-language model (GPT-5) (OpenAI, 2025). Each textbook page is converted to a high-resolution image and resized ($\leq$1024px) before being encoded and processed by the model using a structured JSON prompt.

The model performs three primary tasks:

- segmenting each page into visual panels (ordered top-to-bottom and left-to-right),
- identifying panels that contain human figures, and
- generating candidate annotations for each detected character.

The generated schema includes panel identifiers, a short scene description, the number of people in the image, and per-person attributes such as gender, approximate age group, inferred role, clothing, gesture, and activity. The output is saved as a structured JSON file and converted to spreadsheet format for downstream verification.

These machine-generated annotations serve only as **preliminary labels** that assist the human annotation process.

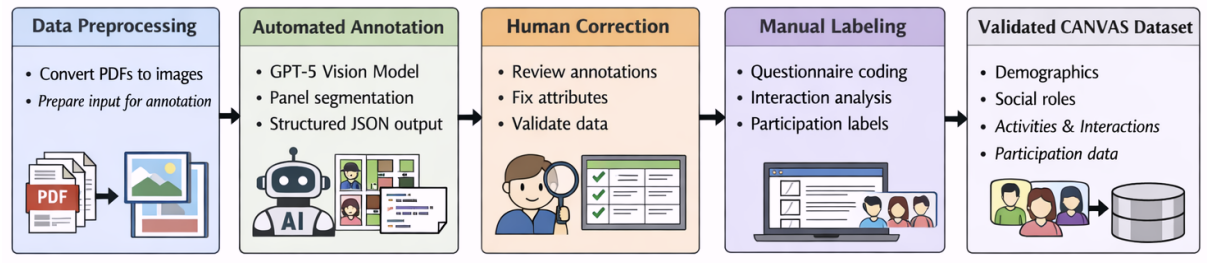

Figure 3: Overview of the CANVAS dataset creation pipeline. Textbook PDFs are converted to page images. A vision-language model generates preliminary annotations, which are verified by human annotators and augmented with manual interaction labels to produce the final dataset.

4.1.3. Human correction

In the human correction stage, trained annotators review the machine-generated annotations alongside the original textbook images. Annotators verify panel segmentation, confirm the number of people present, and correct any incorrectly inferred attributes such as gender, age group, or social role.

When discrepancies occur, annotators directly edit the structured annotations to ensure that all labels accurately reflect the visual content. This step ensures that the dataset does not rely solely on model outputs and that all annotations are validated by human reviewers.

4.1.4. Manual labeling

After verification of the preliminary annotations, additional manual labels are collected through a web-based annotation interface. The interface is implemented as a structured questionnaire with branching logic that guides annotators through a series of questions regarding scene context and interactions.

The questionnaire contains 35 questions organized according to the five modules described in Section 3.3. These questions capture information about actor roles, social settings, interactions among participants, access to artifacts, and indicators of participation or influence within group activities.

Annotators examine each image and answer questions regarding scene composition, participant identities, activity types, and interaction patterns. In scenes containing multiple participants, additional questions capture interactional features such as conversational initiation, eye-contact focus, or access to learning materials.

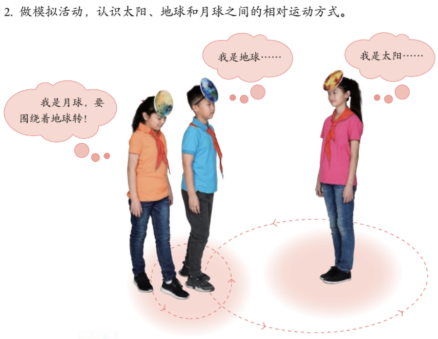

Figure 4: Example textbook illustration used during manual annotation (Section 4.1.4). Students role-play the relative motion of the Sun, Earth, and Moon in a classroom activity. Such scenes allow annotators to label participant roles, activities, and interaction patterns.

Figure 5: Example of a more complex classroom scene containing multiple interacting participants and dialogue bubbles. Such images require annotators to identify participants, interpret interaction structure, and label conversational roles.

4.2. Inter-annotator agreement

To assess annotation reliability, we conducted an inter-annotator agreement study. One instructional unit was randomly selected from each chapter of the Grade 6 textbooks across all publishers. In total, 37 pages were independently annotated by two coders.

Agreement was measured using Cohen's κ, yielding a score of 0.833. This indicates a high level of consistency in annotation decisions despite the complexity of some interaction categories.

4.3. Quality control

Several procedures were implemented to ensure annotation quality. Units from different publishers and textbook volumes were randomly sampled to reduce potential ordering effects during annotation. Annotators were assigned units in a balanced manner to avoid repeated exposure to visually similar scenes.

Randomized assignment and calibration sessions were used to maintain consistency across annotators. Disagreements during pilot annotation were discussed and used to refine annotation guidelines.

4.4. Ethical considerations

The purpose of **CANVAS** is to support research on representation and bias in educational materials. However, datasets that document social attributes may themselves reproduce stereotypes if used without caution.

To mitigate this risk, annotators were trained to avoid assumptions about identity attributes when visual evidence was unclear and to select "cannot determine" where appropriate. The dataset also preserves multiple interaction attributes per character so that participation patterns can be analyzed without imposing rigid categorical interpretations.

Researchers using the dataset should interpret demographic and interactional annotations as observational descriptions of textbook imagery rather than normative judgments about individuals or groups.

5. Applications and Case Study

This section outlines potential applications of **CANVAS** in both NLP research and educational studies. We also present an illustrative case study demonstrating how the dataset can support analyses of gender representation in textbook images.

5.1. Applications in NLP

Bias evaluation and fairness analysis: Existing bias evaluation datasets for Chinese language models, such as CHBias (Zhao et al., 2023), highlight the importance of studying multiple social dimensions including age and appearance. **CANVAS** complements these resources by providing structured annotations of demographic attributes and interaction patterns in visual scenes. Researchers can use the dataset to evaluate how vision–language models describe textbook illustrations. For example, generated captions can be compared against the annotated attributes to examine whether models systematically omit certain participants or emphasize particular roles. In this way, **CANVAS** can serve as a benchmark for evaluating bias and fairness in multimodal systems.

Multimodal language generation and translation: Because the dataset links visual scenes with structured metadata, it can support research on image captioning and multimodal machine translation in Mandarin Chinese. Models trained on the dataset may learn to generate captions that capture activities, gestures, and social roles rather than generic scene descriptions. The presence of role annotations such as *teacher* or *student* also enables controlled caption generation, where models are prompted to produce descriptions focusing on specific participants or interactions.

Social role identification and dialogue analysis: The dataset includes annotations of interactional features such as conversation initiation, eye-contact centers, and speaking roles. These annotations enable new tasks in multimodal social role recognition. For instance, models could be trained to predict which character initiates a discussion or holds the attention of other participants. Combining the visual annotations with dialogue text from speech bubbles further enables discourse-level analyses, such as aligning utterances with speakers or measuring participation patterns within group interactions.

5.2. Applications in education

Educational materials play an important role in shaping students' understanding of social roles and participation in learning activities. Because **CANVAS** provides structured annotations of textbook imagery, it can support several education-related applications.

Curriculum material analysis: The dataset can be used to train models that automatically analyze visual representation patterns in learning materials. By leveraging annotated examples of characters, activities, and interaction roles, models can help identify patterns in how different groups are portrayed in textbooks and other educational resources.

Learning material generation: Generative AI has increasingly been used to assist with course design and educational content creation (Meron and Araci, 2023). However, generative models may reproduce biases present in their training data (Bettayeb et al., 2024). A curated dataset such as **CANVAS** can therefore support the development of tools that generate or evaluate textbook images while taking representation patterns into account.

Assessment item development: Visual content also appears in educational assessments. Datasets that capture representation patterns can help researchers examine whether images used in test items unintentionally emphasize certain groups

or roles. Such analyses may support efforts to improve the fairness and inclusiveness of educational assessments.

5.3. Case Study

To illustrate how **CANVAS** can support empirical analyses of representation patterns, we conduct a preliminary case study examining gender distributions in Grade 6 science textbooks. Previous studies of Chinese textbooks have reported uneven gender representation and stereotypical activity assignments (Sang, 2023; Zhu et al., 2024). Here, using the structured annotations in **CANVAS**, we adopt the gender bias taxonomy and analysis methods developed by Zhu et al. (2024) to examine two indicators: (1) the overall distribution of gendered characters and (2) the distribution of conversation initiators within scenes.

5.3.1. Distributional patterns

A basic indicator of representation is the frequency of male and female characters appearing in textbook illustrations. Using **CANVAS**, we aggregated the number of characters by gender for each publisher's Grade 6 science textbooks. Table 2 summarizes these counts.

Publisher	Gender		
	Male	Female	Unknown
PEP	49	48	46
Jiangsu	136*	82	116
ESE	84	80	91
Daxiang	343*	251	88
Guangdong	163*	147	88
Hebei	136*	80	270
Hunan	164*	114	81
Qingdao63	341	309	52
Total	1,389	1,111	832

Table 2: Distribution of gendered characters in Grade 6 textbooks. $*$ indicates statistical significance for a one-sided binomial test; p-values are adjusted via False Discovery Rate.

These counts provide an overview of how frequently male and female characters appear across publishers. Such descriptive statistics offer a starting point for examining representation patterns in textbook illustrations.

5.3.2. Interactional patterns

Beyond raw counts, the dataset allows analysis of interactional features within scenes. One example is identifying which character initiates a conversation in scenes containing dialogue. Table 3 summarizes the distribution of conversation initiators by gender.

Publisher	Conversation Initiator		
	Male	Female	Unknown
PEP	19	18	20
Jiangsu	63*	23	70
ESE	23	15	24
Daxiang	102*	69	15
Guangdong	69	67	20
Hebei	32	19	167
Hunan	80*	30	23
Qingdao63	103	90	17
Total	491	331	356

Table 3: Distribution of characters identified as conversation initiators in Grade 6 textbook scenes. $*$ indicates statistical significance for a one-sided binomial test; p-values are adjusted via False Discovery Rate.

Because conversation initiation often corresponds to who begins a discussion or proposes ideas within a scene, this indicator can be used to explore participation patterns in group interactions. However, as with other descriptive statistics, these results should be interpreted as exploratory observations rather than definitive conclusions about bias in all textbooks.

Overall, this case study demonstrates how the structured annotations in **CANVAS** enable analyses of representation and interaction patterns in educational imagery.

6. Limitations and Future Work

While **CANVAS** provides a new resource for studying representation in educational materials, several limitations should be acknowledged.

Data scope: The current release of **CANVAS** covers only the Grade 6 science textbooks. Although the broader collection includes textbooks from Grades 1–6 across multiple publishers, only the Grade 6 subset has been fully annotated. Future work will extend the annotations to additional grade levels and subjects (e.g., mathematics) to enable more comprehensive analyses across the elementary curriculum.

Cultural and linguistic context: CANVAS is derived from Chinese educational materials and therefore reflects the cultural and curricular context of Chinese primary education. While this focus provides valuable insights into representation within that context, the findings may not generalize directly to textbooks used in other countries or educational systems. Future work may involve constructing comparable datasets from other languages and regions to support cross-cultural analyses of representation in educational materials.

Annotation subjectivity: Some annotation categories, particularly those related to interaction pat-

terns or conversational roles, involve subjective interpretation. Although we followed detailed annotation guidelines and obtained strong inter-annotator agreement, certain scene attributes (e.g., identifying a conversation initiator or determining interaction centrality) may still be interpreted differently by annotators. Future work will refine the annotation schema and explore incorporating additional contextual signals, such as textual dialogue, to improve reliability.

Automation and model bias: The automated preprocessing stage uses a vision-language model to generate preliminary annotations that are subsequently verified and corrected by human annotators. While human review mitigates many errors, biases present in the underlying model could influence the initial candidate annotations. Future work will examine the differences between raw model outputs and human-corrected labels and explore alternative or ensemble approaches for automated preprocessing.

7. Conclusion

We introduced **CANVAS**, a multimodal dataset designed to capture the social and visual contexts of Chinese elementary science textbook illustrations. The dataset provides structured annotations of characters' demographics, social roles, interactions, and participation patterns within educational scenes. By extending bias analysis beyond isolated text or facial datasets to full classroom illustrations, **CANVAS** enables researchers to study representation and interaction patterns in multimodal educational materials.

The dataset was constructed using a semi-automated pipeline in which a vision-language model generates preliminary annotations that are subsequently verified and refined by human annotators. The resulting annotations provide a foundation for studying how characters, activities, and interactions are portrayed in textbook imagery. An illustrative case study demonstrates how the dataset can be used to analyze representation patterns such as gender distributions and conversational participation.

For the NLP and computer vision communities, **CANVAS** offers a resource for developing and evaluating multimodal systems on tasks such as caption generation, role recognition, and interaction analysis. For education research, the dataset provides structured evidence for examining representation patterns in learning materials and supports future studies on curriculum design and educational equity.

More broadly, **CANVAS** highlights the importance of studying visual context in fairness research. Educational images convey information about participation, authority, and collaboration that may not be captured by text alone. By focusing on Chinese educational materials, the dataset also contributes a perspective that complements existing resources developed primarily for Western contexts. We hope **CANVAS** will support future work at the intersection of NLP, computer vision, and education research.

8. Acknowledgements

This work was supported by the National Science Foundation (NSF) under Grant #1920512, ECR-HER Core Research Program: *Automatic Profiling of Science Assessment Items to Model Item Parameters: A Natural Language Processing Approach*.

9. Bibliographical References

Anjali Adukia, Alex Eble, Emileigh Harrison, Hakizumwami Birali Runesha, and Teodora Szasz. 2023. What we teach about race and gender: Representation in images and text of children books*. *The Quarterly Journal of Economics*, 138(4):2225–2285.

Anissa M Bettayeb, Manar Abu Talib, Al Zahraa Sobhe Altayasinah, and Fatima Dakalbab. 2024. Exploring the impact of chatgpt: conversational ai in education. In *Frontiers in Education*, volume 9, page 1379796. Frontiers Media SA.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Neural Information Processing Systems*.

Jerome S. Bruner. 1986. *Actual minds, possible worlds*. Harvard University Press, Cambridge, MA.

Siobhan Mackenzie Hall, Fernanda Gonçalves Abrantes, Hanwen Zhu, Grace Sodunke, Aleksandar Shtedritski, and Hannah Rose Kirk. 2023. Visogender: A dataset for benchmarking gender bias in image-text pronoun resolution.

Yufei Huang and Deyi Xiong. 2024. CBBQ: A Chinese bias benchmark dataset curated with human-AI collaboration for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2917–2929, Torino, Italia. ELRA and ICCL.

Kazi Md Mukitul Islam and M Niaz Asadullah. 2018. Gender stereotypes and education: A

comparative content analysis of malaysian, indonesian, pakistani and bangladeshi school textbooks. *PloS one*, 13(1):e0190807.

Kimmo Kärkkäinen and Jungseock Joo. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1547–1557.

Tian Lan, Xiangdong Su, Xu Liu, Ruirui Wang, Ke Chang, Jiang Li, and Guanglai Gao. 2025. McBE: A multi-task Chinese bias evaluation benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6033–6056, Vienna, Austria. Association for Computational Linguistics.

Yaron Meron and Yasemin Tekmen Araci. 2023. Artificial intelligence in design education: evaluating chatgpt as a virtual colleague for postgraduate course development. *Design Science*, 9:e30.

OpenAI. 2025. Gpt-5 api. `https://platform.openai.com/`. Large language model.

Barbara Rogoff. 2003. *The cultural nature of human development*. Oxford University Press, Oxford.

Ruicong Sang. 2023. Analysis of gender images in elementary school mathematics textbooks of jiangsu education press. *Lecture Notes in Education Psychology and Public Media*, 15:88–95.

Luthfi Mahdya Susanti, Nunung Suryati, and Utari Praba Astuti. 2021. *Gender Inequality and Education: A Content Analysis of Indonesian EFL Textbook*. Ph.D. thesis, State University of Malang.

Teodora Szasz, Emileigh Harrison, Ping-Jung Liu, Ping-Chang Lin, Hakizumwami Birali Runesha, and Anjali Adukia. 2022. Measuring representation of race, gender, and age in children's books: Face detection and feature classification in illustrated images. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3371–3380.

Lev S. Vygotsky. 1978. *Mind in society: The development of higher psychological processes*. Harvard University Press, Cambridge, MA.

Na Wang, Min Li, and Duo Dong. 2019. Measuring equitable contextualized items in science assessment. In *Annual Meeting of the American Educational Research Association (AERA)*, Toronto, Canada.

Jiaxu Zhao, Meng Fang, Zijing Shi, Yitong Li, Ling Chen, and Mykola Pechenizkiy. 2023. CHBias: Bias evaluation and mitigation of Chinese conversational language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13538–13556, Toronto, Canada. Association for Computational Linguistics.

Haotian Zhu, Kexin Gao, Fei Xia, and Mari Ostendorf. 2024. Disagreeable, slovenly, honest and un-named women? investigating gender bias in English educational resources by extending existing gender bias taxonomies. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 219–236, Bangkok, Thailand. Association for Computational Linguistics.

A. Annotation Schema Details

The **CANVAS** annotation schema organizes labels into five modules that capture demographic attributes, scene context, interaction structure, artifact access, and participation patterns within textbook illustrations. Table 4 summarizes the main modules and the types of attributes included in each.

These modules are designed to capture both the presence of characters and the roles they play in educational interactions. By combining demographic attributes with interactional and contextual labels, the schema enables analyses of representation patterns, participation structures, and social dynamics within textbook imagery.

B. Example Annotated Scene

Figure 6 presents an example illustration from a Grade 6 science textbook used in the **CANVAS** dataset. The scene depicts a classroom activity in which three students simulate the relative motion of the Sun, Earth, and Moon. The student on the right represents the Sun, while the two students on the left represent the Earth and the Moon. Dialogue bubbles indicate the roles assigned to each student during the activity.

Table B shows a subset of the structured annotations for the characters appearing in this scene.

In this scene, the interaction format is coded as a classroom demonstration activity. The three actors participate in a role-playing exercise to illustrate astronomical motion. The annotations record each participant's demographic attributes, assigned role in the activity, and participation in the group interaction.

Module	Description and Example Labels
Demographics of actors	Describes attributes of each character appearing in the scene. Example labels include gender, approximate age group, ethnicity etc. Each character is assigned a unique identifier that links annotations across modules.
Social setting and activity	Captures the environment and activity type in which the actors participate. Examples include classroom experiment, group discussion, home activity, or outdoor observation.
Interaction format	Describes how actors interact within a scene, such as whether the activity is individual, collaborative, or discussion-based. This module also records interaction features such as conversational structure or group participation patterns.
Artifacts and access	Identifies the learning materials or objects present in the scene and which actors interact with them. Examples include laboratory equipment, textbooks, writing materials, or experimental tools.
Participation patterns	Captures indicators of interactional influence or participation in group activities. Example labels include conversation initiator, interaction center (e.g., eye-contact focus), speaking participation, and access to shared resources.

Table 4: Overview of the CANVAS annotation schema modules and example labels.

ID	Gender	Age	Identity	Activity
P1	Female	Child	Student	Simulating the Moon's orbit
P2	Male	Child	Student	Representing the Earth
P3	Female	Child	Student	Representing the Sun

Table 5: Example subset of annotations corresponding to Figure 6.

C. Automated Annotation Prompt and Output Schema

To generate preliminary annotations, each textbook page image is processed by a vision–language model. The model receives a page image together with an instruction prompt that asks it to segment the page into visual panels and return structured JSON output. The resulting annotations are used only as preliminary labels and are subsequently reviewed and corrected by human annotators.

Figure 6: Example annotated illustration from a Chinese elementary science textbook. Students simulate the relative motion of the Sun, Earth, and Moon.

Actor ID	English Translation of Dialogue
P1	"I am the Moon. I revolve around the Earth."
P2	"I am the Earth."
P3	"I am the Sun."

Table 6: English translation of dialogue bubbles appearing in the example illustration shown in Figure 6.

C.1. Prompt Template

The following prompt excerpt illustrates the instructions used in the automated annotation stage:

> You are a vision-language assistant helping analyze textbook pages.
>
> This image is a textbook page and may contain multiple independent visual panels or illustrations.
>
> 1. Identify and describe each independent image (if more than one), and give each a unique ID like `{image_id}_A`, `{image_id}_B`, etc., following the order top-to-bottom and left-to-right.
> 2. For each image that contains any people (real or fictional), provide:
>   - a short image description,
>   - the number of people,
>   - a list of people with:
>   - person ID,
>   - inferred gender,
>   - inferred age group,
>   - social identity,
>   - outfit description,
>   - gesture,
>   - activity,
>   - chatbox text (if any),
>   - conversation initiator,

- central role,
- supporting role.

Respond in valid JSON format only. Do not include explanations or Markdown.

C.2. Output Schema

The automated stage returns a JSON list of annotated panels. A simplified version of the schema is shown below:

```
[
  {
    "image_id": "...",
    "description": "...",
    "people_count": 3,
    "people": [
      {
        "id": "P1",
        "gender": "...",
        "age": "...",
        "identity": "...",
        "outfit": "...",
        "gesture": "...",
        "activity": "...",
        "chat": "...",
        "conversation_initiator":
            "...",
        "central_role": "...",
        "supporting_role": "..."
      }
    ]
  }
]
```

C.3. Use in the annotation pipeline

The model output is treated as a candidate annotation only. In the next stage of the pipeline, human annotators review the original page image together with the generated structured output, correct segmentation errors, revise incorrect person attributes, and add or refine interaction labels. This hybrid workflow allows the automated stage to improve efficiency while preserving human validation of the final dataset.

D. Annotation Guidelines

This appendix provides a condensed version of the annotation guidelines used during the manual labeling stage of the **CANVAS** creation process. The goal of these guidelines is to ensure that annotators apply consistent criteria when identifying demographic attributes, interaction patterns, and social roles depicted in textbook illustrations.

The full annotation schema is organized into five modules that capture complementary aspects of the visual scenes:

- **Module 1: Demographics** records descriptive attributes of actors appearing in an image, including inferred gender, age group, and clothing information.
- **Modules 2 and 3: Social Settings and Activities** describe the context of the depicted activity, including the social identities of participants, the setting in which the activity takes place, and the type of learning activity being performed.
- **Module 4: Group Interactions** captures observable interaction structures among actors, such as eye-level positioning, access to shared materials, and visible engagement through gestures or body language.
- **Module 5: Power Dynamics** focuses on conversational and social dynamics that may signal relative influence or participation within a group activity.

Each module contains a set of coding questions accompanied by operational rules designed to guide annotators when interpreting ambiguous visual cues. These rules were refined during annotator training and calibration to improve consistency across coders.

Tables 7–11 summarize the items and coding rules associated with each module. The tables are intended as a reference for readers and for future researchers who may wish to reproduce or extend the annotation process.

Module 1: Demographics	**Rules**
Actors present in the entry?	Code "no" if images don't include any people, parts of people (lower 75% body, such as hands) without heads.
Dressing of actor or actors	a. Copy the descriptions of dressing of each actor, starting from actor 1. b. Use a bullet for each actor. c. Correct the excel text if the ID or description is wrong.
AI-generated descriptions for age or gender	a. Students should wear red scarf. Don't rely on the height of entities to infer if they are students or grown-ups. b. Make corrections in the excel file if needed. c. Leave the responses blank when the GPT descriptions haven't been generated or GPT failed to capture the image. Also include a note about this at the end.
Does any actor have a "cannot tell" code for gender?	
Reason for "cannot tell" code for gender	

Table 7: Items and Coding Rules for Module 1

Module 2&3: Social Settings and Activities	**Rules**
What is the social identity of the child (or children)?	Choose non-student if the image doesn't have any context or background. In addition, don't assign codes for "setting of activity" or "type of activity."
Social identities of all actors 1. Student 2. Teacher 3. Members in a service-learning site as experts who students learn from (e.g., museum, public library, car testing center, factory, observatory) 4. Participants or subjects in a community as the latter are studied or surveyed by the students (... for data collection) 5. Other (include notes at the end)	a. Assign 4 when students conduct a survey or questionnaire with their peers. b. Okay to refer to texts as clues when figuring out the identity of actors
Setting of activities 1. elementary classroom/schools (including school playground, school backyard, field trip) 2. daily household (for family activities) 3. a site for school service-based learning 4. other (specify below)	Okay to reference the text when determining if the professionals are STEM related or not.
Type of activities 1. group discussion 2. investigation, exploration, or experiment (mainly data collection activities) 3. summarizing /communicating/presenting findings 4. displaying or showing case products 5. service learning	a. After coding with 1, likely coders will need to assign codes for group dynamics. b. We decided not to add a category of "teaching" because if the teacher is present with students AND the teacher is the only one who talks or lectures, we should be able to figure out this situation based on codes from other coding questions. So for the parsimony purpose, this category is not needed here. For situations like this, assign 1. c. Assign 4 if actual products constructed by students are displayed. d. Assign 2 if students are collecting data or conducting investigations in a service site. Only go with 5 if the other 4 codes are not applicable.

Table 8: Items and Coding Rules for Module 2 and 3

Module 4: Group Interactions	**Rules**
Does the image include more than one actor?	
Is anyone above the eye level of the rest of others?	a. Typically we code eye level only if individuals interact with each other (otherwise assign the code N/A). b. If all are sitting or kneeling down, then consider none is above the eye level (code NO). c. Notice that above the eye level isn't necessarily an indicator of exerting power. In the setting where others engage with entertainment or eating food, above the eye level is at a lower level of the power hierarchy.
Who is the center of eye contact for some group members?	
Who appears to access learning/lab materials for the shared activity?	
Who demonstrates a greater amount of energy/enthusiasm, engagement, or confidence via their gestures?	a. Here we ask who appears to be more engaging, motivated, or confident as shown in their gestures (nonverbal cues). b. Gestures or body language often include hands higher than chest. Exceptions include students pointing to the Powerpoints or poster content, students being in particular postures or activities (moving, or dancing). c. Sometimes the gesture may also involve an additional code for the previous coding questions for accessing or manipulating learning materials.

Table 9: Items and Coding Rules for Module 4

Module 4: Group Interactions	**Rules**
Does the image include more than one actor? Is anyone above the eye level of the rest of others?	a. Typically we code eye level only if individuals interact with each other (otherwise assign the code N/A). b. If all are sitting or kneeling down, then consider none is above the eye level (code NO). c. Notice that above the eye level isn't necessarily an indicator of exerting power. In the setting where others engage with entertainment or eating food, above the eye level is at a lower level of the power hierarchy.
Who is the center of eye contact for some group members?	
Who appears to access learning/lab materials for the shared activity?	
Who demonstrates a greater amount of energy/enthusiasm, engagement, or confidence via their gestures?	a. Here we ask who appears to be more engaging, motivated, or confident as shown in their gestures (nonverbal cues). b. Gestures or body language often include hands higher than chest. Exceptions include students pointing to the PPT or poster content, students being in particular postures or activities (moving, or dancing). c. Sometimes the gesture may also involve an additional code for the previous coding questions for accessing or manipulating learning materials.

Table 10: Items and Coding Rules for Module 4

Module 5: Power Dynamics	**Rules**
Is the idea, question, or experience being investigated asked by a student?	
Whose input or comment is being valued or privileged by mainstream STEM?	a. Here we ask whose remarks are consistent with how scientists or mathematicians think or comment on. b. If only one entity speaks, then default this individual's idea is being valued. This will be handled in the data analysis.
Who dominates the group discussion in terms of airtime?	a. here we ask who offers more input or comments than others? E.g., all the others have one sentence or one chat bubble, but one entity has two chat bubbles or two sentences in an image. b. An entry may include more than one image. The evaluation of airtime is based on each image.
For those who contribute to the group discussion, what ideas do they offer?	For this coding question, we are interested in whether individuals express disagreement or a new/alternative idea
Do any individuals exhibit a peripheral role?	Here the role (and activities) likely implies a lower status for the hierarchy of social class.
Based on physical positioning or gestures of entities, who are disadvantaged compared to other entities?	a. Physical positioning or gestures of entities can indicate an advantage over other entities. e.g., dad and child sitting in the center of the sofa together, which may indicate a closer or more equal relationship at this moment, especially in terms of shared activity (watching TV), whereas the mom is holding the tea and standing next to the sofa.

Table 11: Items and Coding Rules for Module 5