



HAL
open science

Towards Inclusive Education: Multimodal Classification of Textbook Images for Accessibility

Saumya Yadav, Elise Lincker, Caroline Huron, Stéphanie Martin, Camille Guinaudeau, Shin'Ichi Satoh, Jainendra Shukla

► To cite this version:

Saumya Yadav, Elise Lincker, Caroline Huron, Stéphanie Martin, Camille Guinaudeau, et al.. Towards Inclusive Education: Multimodal Classification of Textbook Images for Accessibility. *Multimedia Modelling* 2025, Jan 2025, Nara, Japan. hal-04860245

HAL Id: hal-04860245

<https://hal.science/hal-04860245v1>

Submitted on 31 Dec 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Towards Inclusive Education: Multimodal Classification of Textbook Images for Accessibility

Saumya Yadav¹[0000-0003-3747-961], Élise Lincker²[0009-0005-1104-1785],
Caroline Huron³[0000-0002-3890-6110], Stéphanie Martin⁴, Camille
Guinaudeau^{5,6}[0000-0001-7249-8715], Shin’ichi Satoh⁶[0000-0001-6995-6447], and
Jainendra Shukla¹[0000-0002-6526-0087]

¹ HMI Lab, IIIT-Delhi, India

² Cedric, CNAM, Paris, France

³ SEED, Inserm, Université Paris Cité / Learning Planet Institute, Paris, France

⁴ Le Cartable Fantastique, Paris, France

⁵ Japanese French Laboratory for Informatics, CNRS, Tokyo, Japan

⁶ National Institute of Informatics, Tokyo, Japan

Abstract. To foster inclusive education, accessible educational materials must be tailored to meet the diverse needs of students, especially students with disabilities. Images in educational materials play an important role in understanding textbook exercises but can also distract such students and make them more challenging. Therefore, we need an adaptive system that identifies non-essential images to reduce cognitive load and distractions for students with disabilities. Accordingly, this work proposes a computational framework to categorize textbook exercise images, facilitating their inclusion in accessible textbooks and enhancing learning for visually impaired and neurodevelopmental disorders students. Using three French textbook exercise datasets of 652 (text, image) pairs, we compared monomodal (text-only) and multimodal (text and image) classification approaches. We found that text-based models, particularly CamemBERT, excel in classifying images, achieving an accuracy rate of 85.25% on French text data. We also use Local Interpretable Model-agnostic Explanation for model interpretability and conduct qualitative analyses to deepen insights into model performance. This work is only a very first step towards an automatic translation of inclusive textbooks. Moreover, this paper revealed that this first step is already very challenging. We hope to draw more researchers’ attention to this problem.

Keywords: Educational Technology · Textbook Image Classification · Inclusive Education · Adaptive Learning · Multimodal Classification.

1 Introduction

As per the United Nations Convention on the Rights of Persons with Disabilities, inclusive education asserts that it is the right of every student, not merely a privilege [21, 13], underscoring the importance of ensuring that educational opportunities are accessible to all individuals, irrespective of their abilities or

disabilities [19]. The right to education is universal, transcending limitations imposed by physical or cognitive disabilities. However, conventional educational resources, mainly textbooks, are not inherently designed to cater to the diverse needs of all learners, especially those with disabilities [19, 33]. Inclusive education addresses this disparity by ensuring that educational materials are accessible and accommodating to all students, regardless of their disabilities.

Students benefiting from inclusive education include those with visual impairments and Neuro-Developmental Disorders (NDDs). Visual impairments range from difficulties with images, blurry vision, and partial sight to complete blindness, while NDDs affect brain function, impacting physical, social, academic, and occupational functioning [22]. Common NDDs in childhood include Attention Deficit and Hyperactivity Disorder (ADHD), Autism Spectrum Disorders (ASD), and Developmental Coordination Disorder (DCD) [5, 6]. Children with visual impairment and NDDs often face academic challenges that do not reflect their true abilities due to visual and fine motor skill difficulties [18, 2]. Similarly, visually impaired students encounter obstacles in accessing irrelevant visual materials and materials without suitable adaptations. Therefore, classrooms should consider these students' visual, reading, and gaze coordination difficulties when utilizing textbooks to promote inclusive education.

Recognizing their needs lays the foundation for building a computational system that automatically adapts textbooks, addressing reading, motor coordination, and attention difficulties to foster inclusivity in classrooms by tailoring educational materials for children with disabilities. Some non-profit organizations, such as *Le Cartable Fantastique* [16], have started creating adapted digital textbooks for children with DCD. However, manually transforming materials poses challenges due to diverse textbooks, frequent updates, and resource-intensive processes. The optimal solution would be for publishers to create inherently tailored textbooks for pupils with disabilities, which is currently unattainable. Therefore, automating textbook adaptation using Artificial Intelligence (AI) and Machine Learning (ML) is essential for improving accessibility.

A significant challenge in creating tailored textbooks is the manual annotation of images for diverse students. Professional educators and annotators currently spend considerable time and effort to classify images in textbooks, a process that is both labour-intensive and prone to errors [16]. This highlights the importance of automating image classification to ensure consistency and efficiency. Textbook exercise image classification is crucial to address this challenge, especially for the visually impaired and NDD students. By accurately identifying and categorizing images into different classes, we can remove those that are not required to understand the text. This approach ensures that essential visual information is preserved while managing non-essential images to reduce cognitive load and distractions. To address this, we propose a novel image classification framework for textbook exercise images. This framework categorizes images into *Essential*, *Informative*, and *Decorative* classes with expert help. It aids textbook adaptation by determining image inclusion in the user interface and allowing for customization based on user needs and preferences. Images can remain unaltered, be placed at the end of exercises, or be substituted with alternative

text. Excluding *Decorative* images ensures document clarity and accessibility [28], particularly for students with distraction disabilities like NDD.

The main contributions of this work are: (1) the development of a novel computational framework for classifying textbook illustrations into distinct categories; (2) an empirical evaluation of monomodal and multimodal approaches for classifying (text, image) pairs; and (3) a thorough interpretability analysis using Local Interpretable Model-agnostic Explanations (LIME) [27].

2 Related Work

The work presented in this article is related to different fields: Natural Language Processing (NLP) applied to textbooks and text-image similarity and interaction. Limited research exists on NLP applied to textbooks, with some studies focusing on question generation and interactive content integration [3, 7, 1]. Interest in adaptive textbooks is increasing, yet manual concept indexing remains challenging. Recent advancements introduce ML methods like FACE [4] for automatic concept extraction, enhancing adaptive textbook technologies. The proliferation of MOOCs presents challenges with unstructured data, but ML frameworks for concept extraction show promise in addressing these issues [11]. Moreover, the research explores ML methods to automate the labour-intensive process of labelling educational data, particularly in Japanese schools [32]. While educational image classification primarily focuses on image type identification, recent work like [20] introduces datasets for illustration classification and [31] summarizes progress in chart classification. Additionally, [8] proposes a novel semi-supervised image classification method using curriculum learning, MMCL, outperforming five state-of-the-art methods on eight image datasets.

Similar to our goal of adapting textbooks for children with disabilities, recent studies have focused on modelling and extracting content from textbooks [16] or classifying exercises based on their adaptation for children with DCD [15]. However, these works focus on layout and textual content, not the textbook’s images. To our knowledge, there is a gap in research regarding image classification for textbook adaptation to promote inclusivity. Two recent papers explore the relationship between text and image in image-text retrieval and classification [25, 23]. In [23], authors present a classification framework analyzing semantic relationships between images and textual descriptions, defining eight classes based on cross-modal mutual information, semantic correlation, and status metrics. While they utilize publicly available datasets, they mainly comprise single-sentence captions or labels, differing from the context-rich text in our exercises. Several efforts, such as the Web Form Accessibility Framework for the Visually Impaired (WAFI) [9], have been made to improve website accessibility for visually impaired individuals. *W3C Web Accessibility Initiative*¹ provides guidelines for replacing images with alternative text on web pages. Although not designed for educational contexts, these guidelines can be adapted to ensure essential visual information is retained, enhancing comprehension for all.

¹<https://www.w3.org/WAI/tutorials/images/>

No prior research explores the relationship between images and text in educational materials. While strides have been made in adapting textbooks for children with disabilities, a gap remains in understanding how images contribute to inclusive learning. Existing studies focus on layout, textual content, and semantic relationships between images and text but often overlook the nuanced interaction between images and more extensive textual content in educational exercises. To address these challenges, we propose a computational framework for multimodal classification of textbook illustrations for adaptive learning environments.

3 Problem and Data Challenges

Images in textbooks serve diverse roles, making the complex task of annotating them essential for automated adaptation to meet diverse needs. This process involves classifying images into categories such as *Essential*, *Informative*, or *Decorative*, based on their educational roles. Differentiating between *Essential* and *Informative* images involves understanding their educational value and context. An *Informative* image is not crucial for completing an activity but serves an informative purpose, such as providing clues for solving an exercise or depicting a concept unfamiliar to students. The *Essential* and *Informative* categories aim to provide subtle adaptations for diverse visual impairments, ensuring optimal accessibility to educational materials.

The challenge also lies in the variability of content and layouts across textbooks, which makes standardization difficult. Data challenges include dealing with inconsistent image quality and varying levels of detail, which complicate automated classification. Accurate manual annotation requires significant expertise for consistent classification but is time-consuming and resource-intensive, especially for large datasets. This process presents scalability issues and cannot efficiently handle high volumes of images. Variability in image types and content further complicates manual annotation, necessitating adaptable methods. Ensuring consistent quality across datasets is challenging due to potential errors and inconsistencies. Automated systems can address these issues by providing initial classifications that experts can review, thereby enhancing efficiency and consistency. Automation is crucial for managing large datasets and improving accessibility, as ML algorithms can classify images based on visual and contextual features, reducing manual effort and supporting the creation of inclusive educational materials.

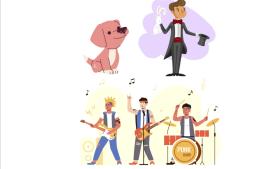


4 Dataset

Our dataset consists of exercises with images from three elementary-grade language-learning French textbooks in PDF format. For the extraction process, each PDF is parsed to an XML file using pdfalto² and MuPDF³ tools. The extracted words are then grouped into text segments, which are grouped into activity blocks based

²<https://github.com/kermitt2/pdfalto>

³<https://github.com/ArtifexSoftware/mupdf>

Table 1: Categorization of Images and Text in the Adaptation Process

Class	Essential	Informative	Decorative
Images			
Text	Write the sound common to the three words represented by the drawings.	Text: During prehistoric times, people painted cave paintings on the walls of their caves. Q: Find the verb. What tense is it conjugated to?	Copy the sentences if you recognize the verb "to go". (a) I leave at the same time every morning. (b) Saturday I weeded the garden path.

on layout, font style, and spacing features. Images are associated with blocks according to their position on the page. Then, two experts from *Le Cartable Fantastique* defined three classes and performed the manual annotation for each image associated with its respective text. The classes are as follows:

- **Essential Images:** These are compulsory for understanding or resolving an activity. They will be incorporated into the adaptation.
- **Informative Images:** They contribute to understanding the text and provide supplementary information, such as clues for solving an exercise or imparting knowledge. Although not essential, these images can be placed after the exercise.
- **Decorative Images:** They are unrelated to the overall exercise and irrelevant to the text. They may be removed when adapting the activity to streamline the interface.

Some contend that streamlining presentation by omitting decorative images could be beneficial [28]; while others claim that all users should be offered the same experience, including the option to receive descriptions of decorative elements. For visually impaired students, varying detail in image descriptions based on their relevance is crucial. Excluding *Decorative* images ensures only relevant content is retained, boosting accessibility alongside materials like braille and screen readers. Classifying illustrations helps NDD students by allowing educators to include images that support educational objectives while excluding those categorized as *Decorative*. An example of the categorization of images with their respective text is shown in Table 1. In the *Essential* class, the image is mandatory to solve the exercise, while the purpose of the image in the *Informative* class is to give additional information to the student, who may not know exactly what is a cave painting. Finally, for the *Decorative* class, the image associated with the text is not required to solve the exercise and only has a decorative purpose.

Our study used three elementary-grade French language textbooks, two from the same editor for training and validation with an 80:20 split, and one from a different editor as the test set to ensure unbiased evaluation. The test set includes data from unseen textbooks to assess model performance on diverse publishers.

Table 2: Distribution of unique labels of train+validation set and test set

	Essential	Informative	Decorative
Test Set	131	75	38
Train+Validation Set	257	93	58

Table 3: Most frequent words in each class

Test			Train+Validation		
Essential	Informative	Decorative	Essential	Informative	Decorative
name: 29	word: 77	text: 14	write: 160	word: 93	verb: 21
drawing: 27	text: 46	verb: 11	drawing: 103	text: 47	text: 20
write: 24	verb: 41	sentence: 10	word: 101	verb: 42	sentence: 16
find: 23	observe: 34	recopy: 8	name: 89	observe: 24	word: 15
give: 17	sentence: 25	combine: 7	use: 75	write: 23	write: 15
associate: 17	red: 20	complete: 6	find: 66	name: 21	complete: 13
word: 16	name: 20	name: 6	represent: 66	remove: 17	find: 13
describe: 16	letter: 18	word: 6	sound: 64	other: 14	recopy: 9
complete: 15	read: 15	write: 5	letter: 46	sound: 14	name: 8
sentence: 14	c: 15	personal: 5	sentence: 41	pink: 14	remove: 8

After removing blank entries from the PDFs, the processed data’s final size is shown in Table 2. Although the distribution appears nearly uniform, we acknowledge the impact of topics and image characteristics on model effectiveness and ensured consistency by evaluating with a separate textbook. Table 3 highlights the most frequent words, revealing distinctive patterns across *Essential*, *Informative*, and *Decorative* categories. Notably, ‘drawing’ and ‘write’, are most prevalent in *Essential*, ‘word’, and ‘text’ in *Informative*, and ‘text’, and ‘verb’ in *Decorative*, offering insights into textual elements characterizing each class.

5 Approaches

For the textbook illustration classification, we utilized various modalities, including both multimodal approaches, which integrate images and text, and monomodal approaches that consider text or images independently. The workflow for the final approaches used in this work is shown in Fig. 1.

5.1 Multimodal Approaches

For the multimodal approach, we used the CLIP model [26], which excels in bridging the semantic gap between images and textual descriptions through zero-shot transfer, natural language supervision, and multimodal learning. Our text data is extracted from French textbooks, so we translated the text into English using the open-source offline translation library Argos Translate⁴ with Open-

⁴<https://github.com/argosopentech/argos-translate>

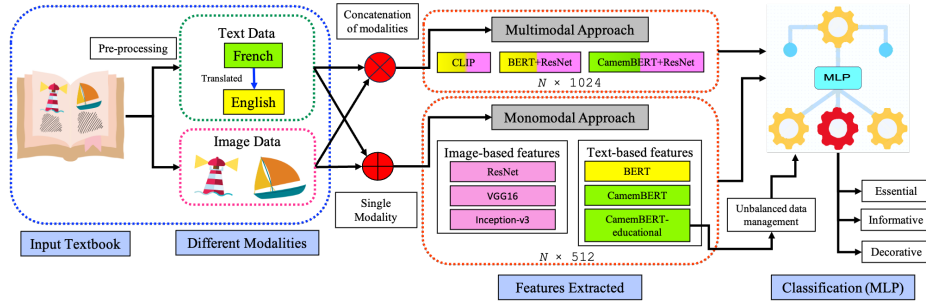


Fig. 1: Workflow for Multimodal Textbook Illustration Classification

NMT [14] to align with CLIP’s requirement for English text input. We chose the RN101 variant in CLIP, which performed better on our dataset. This variant has a 512-dimensional embedding, processes 224×224 images with a backbone of (3, 4, 23, 3) layers, and includes a text transformer with 12 heads, 512 width, and eight layers, enhancing multimodal comprehension.

We utilized the CLIP model to compute the cosine similarity between images and texts, establishing their relationship. For example, the cosine similarity between text and image features in Table 1 is 0.42, 0.48, and 0.41 for the *Essential*, *Informative*, and *Decorative* classes. Since some of our textual data exceeded CLIP’s default token length of 77, we used truncation and segmentation for text data. Truncation shortens texts to the default length, while segmentation divides texts into segments of the default length. We calculated average, maximum, and truncated cosine similarity scores for text-image pairs. Descriptive statistics reveal that the *Essential* class has the highest mean average similarity score (0.42), slightly above *Decorative* (0.419), with *Informative* showing lower central tendencies. For maximum similarity scores, *Informative* leads with a mean of 0.438, followed by *Decorative* (0.431) and *Essential* (0.422). The *Informative* class also has the highest truncated similarity score (0.433). Despite *Informative* showing higher scores, the close mean values across classes indicate that CLIP-based similarity alone may not fully capture distinctions. To improve classification, we used the RN101 variant of CLIP to extract 512-size feature vectors and applied a Multi-Layer Perceptron (MLP) as shown in Fig. 1.

5.2 Monomodal Approaches

Text-based features encoding: We strategically used domain-specific language models for text feature extraction: BERT for English text [12] and CamemBERT for its French counterpart [17]. These state-of-the-art language models provided comprehensive text understanding and nuanced insights into linguistic complexity. We followed the same procedure for classification as we did for our multimodal approach. We extracted features using *bert-base-uncased* for English and *camembert-base* for French. Larger models diminished performance, likely

due to overfitting and increased computational demands. We then tokenized English text using *bert-base-uncased* and French text using *CamembertTokenizer*, processed tokens through all layers of BERT and CamemBERT, and extracted features from the last hidden layer. Adaptive average pooling was used across the sequence length dimension for fixed-size representation. The textual embeddings obtained through these models are then fed to a MLP for classification.

To further enhance our exercise representation, we follow the work of [15] and fine-tune CamemBERT’s language model on educational data: lessons and activities from four textbooks (two textbooks from the collection used for training, excluding the exercises used to build our dataset, and two other unseen textbooks), 1293 Fantastiques Exercices provided by the organization *Le Cartable Fantastique*, and the 79 original reading texts from *Alector*.

Image-based features encoding: For Image feature extraction, we used CNN-based models, specifically ResNet [10], VGG16 [29], and Inception-v3 [30], renowned for their proficiency in image recognition tasks. We used pre-trained ResNet-50, VGG16, and Inception-v3 models, adjusting their final layers to produce 512-dimensional feature vectors that matched CLIP’s size. All layers, except the modified final one, are frozen to preserve the knowledge encoded in earlier layers. The images were pre-processed by resizing to ensure compatibility with the ResNet model’s input expectations.

We extracted the features using these models, generating the feature vectors of dimensions $N \times 512$, with N representing the total number of data instances. Following the same method we used for the CLIP model, we extracted data using these models, ensuring uniformity in our approach. Subsequently, we used MLP to train the extracted data, enhancing the analytical capabilities of our research.

5.3 Unbalanced Data Management

The initially processed data is highly imbalanced, as shown in Table 2, which can detrimentally affect model performance by favouring the majority class. To mitigate this issue, we used two strategies, either jointly or separately:

- **Class_weight strategy:** We applied the *class_weight* strategy using scikit-learn [24] to prioritize the minority class. The ‘balanced’ strategy dynamically adjusts class weights based on their distribution in the training data, giving higher weights to underrepresented classes.
- **Data generation:** We augment the initial training set with 150 instances from the *Decorative* class, consisting of text-only data and random images from textbooks. We then merge this augmented data with the original set, extract features, and apply the same procedure as previous models, passing it through the MLP.

6 Result and Discussions

6.1 Experimental Setups and Ablation Study

Our MLP has an input layer of size 512, two hidden layers with 256 and 128 neurons, respectively, and an output layer for 3 classes, with ReLU activations

in between. It was trained using CrossEntropyLoss and Adam optimizer, with a batch size of 32 over 30 epochs. Based on validation performance, early stopping with patience of 5 epochs was applied. As part of our ablation study, we experimented with different numbers of hidden layers in the MLP: 1, 2, 3, 4, and 6 hidden layers. The results indicated that the MLP with 2 hidden layers achieved the best performance.

We also conducted an ablation study on fusion methods to evaluate their impact on model performance, using both Early fusion and Late fusion of text and image modalities. In Early fusion, we combined modalities by computing the maximum, minimum, and concatenation of two features. In Late fusion, we applied the same MLP used previously to the extracted data of both modalities, averaging and taking the weighted average of both outputs to predict the result. The best accuracy result obtained from these methods on the test data was 57.14%, which was not as good as the performance of the Early fusion approach. Therefore, we generated all results using the Early fusion concatenation method. We chose concatenation because it preserves information from both modalities, enriching the feature space and capturing complementary data. The concatenated features, sized 1024, were then input into an MLP.

6.2 Results

The results section presents predictions on test data from a third textbook by a distinct editor. Table 4 shows that the text-based models outperform the image-based models. The image-based classification (ResNet, VGG16, and Inception-v3) gives lower results than the majority class classifier, showing that image data alone is insufficient for classifying images as *Essential*, *Informative*, or *Decorative* in the context of an exercise, indicating the need for additional data or features to enhance classification accuracy. The French language model CamemBERT achieved the highest accuracy in text-based classification without fine-tuning, likely due to the original data being in French, highlighting the importance of linguistic compatibility in model performance. While CamemBERT-educational was expected to perform better, its lower performance suggests that semantic features of text exercises may not be crucial in *image classification*.

The bottom part of Table 4 shows that early fusion of image and textual data does not enhance performance relative to text-based classification. The best multimodal accuracy is achieved with the CLIP model, slightly surpassing the concatenation of CamemBERT and ResNet features. As previously shown in Fig. 1, similarities values between text and image computed with the CLIP model have nearly similar values for all three classes, contrary to our intuition that the *Decorative* images had a lower semantic cosine similarity with the image (redundancy) where the necessary images had a higher semantic cosine similarity (complementarity). Further, Table 4 also indicates that CLIP’s higher accuracy stems from effectively integrating textual and image features, whereas CamemBERT+ResNet achieves a higher F1 score due to its superior precision and recall in capturing text-image relationships.

Finally, Table 4 presents the results obtained with the different strategies for dealing with our data unbalanced issues. The best results are obtained when

Table 4: Monomodal and Multimodal Classification Results with CamemBERT, Using Unbalanced Data Strategies (CW: Class Weight, DA: Data Augmentation)

Models	Modality	Accuracy	F1-Score
Majority Class	-	0.727	
BERT	text	0.816	0.818
CamemBERT	text	<u>0.836</u>	<u>0.831</u>
CamemBERT-educational	text	0.80	0.798
ResNet	image	<u>0.525</u>	<u>0.431</u>
Inception-v3	image	0.504	0.421
VGG16	image	0.50	0.421
BERT+ResNet	text+image	0.754	0.755
CamemBERT+ResNet	text+image	0.803	<u>0.796</u>
CLIP	text+image	<u>0.807</u>	0.79
CamemBERT-CW ⁻ -DA ⁻		0.836	0.831
CamemBERT-CW ⁺ -DA ⁻		0.816	0.83
CamemBERT-CW ⁻ -DA ⁺		0.828	0.815
CamemBERT-CW ⁺ -DA ⁺		0.853	0.849

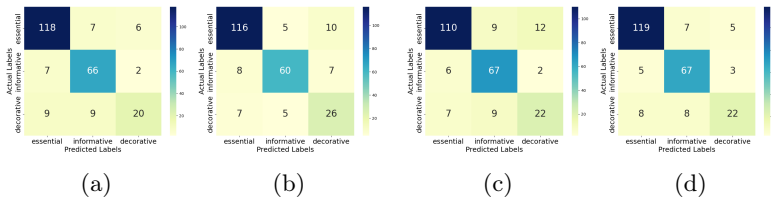





Fig. 2: Confusion matrices for the CamemBERT model only (a), with data augmentation (b), with class weight strategy (c) and both data augmentation and class weight strategies (d).

both class weight and data augmentation strategies are used, achieving an accuracy of 85.25% when applied on our best model (CamemBERT only)⁵. From a qualitative point of view, as pictured in the confusion matrices in Fig. 2, the data augmentation tends to classify more examples in the *Decorative* class, both correctly and incorrectly (Fig. 2b), while the class weight strategy tends to improve the number of correctly classified instances of the under-represented classes (*Decorative* and *Informative*) at the expense of the *Essential* class (Fig. 2c). It’s worth noting that utilizing either the class weight or data augmentation strategy alone decreases performance compared to the initial result. Finally, combining both strategies improves the number of correctly classified instances for all classes.

Furthermore, these findings are significant for individuals with visual impairments and NDD students. By removing *Decorative* images, our approach ensures that only essential visual information is retained, enhancing accessibility and catering to the specific needs of students with NDD. This deliberate cura-

⁵Similar tendencies are found when applied on the other models.

Table 5: Comparison of results from text-based, image-based, and multimodal models. ✓(✗) indicates correct (incorrect) labelling of the (text, image) pair.

	Exercise 1	Exercise 2	Exercise 3
Text	Decipher this puzzle.	Choose the correct adjectives to describe the princess.	What type of art is this?
Images			
Text-based	✗	✓	✓
Image-based	✓	✓	✗
Multimodal	✗	✗	✗
Cosine similarity	0.403	0.456	0.422

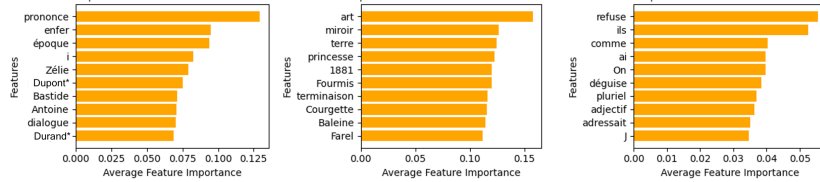


Fig. 3: LIME explanation of top features for each class (a) Essential, (b) Informative, and (c) Decorative, based on average importance. *: anonymized

tion of visual data fosters inclusivity and improves learning outcomes for these marginalized groups.

6.3 Qualitative Analysis and Explainability

Table 5 compares results obtained with text-based, image-based, or multimodal models on three random (text, image) pairs. For the given exercises, the text-based model incorrectly labels Exercise 1, related to *image description*, while the image-based model performs well in both Exercise 1 and Exercise 2, i.e., in *image description and image reading* but fails to classify Exercise 3. Notably, images are required for all exercises. The cosine similarities for Exercises 1, 2, and 3 are 0.403, 0.456, and 0.422, respectively, reflecting the varying degrees of alignment between text and image features. The text-based model’s misclassification of Exercise 1 highlights its limitation in understanding the role of visual content in specific tasks. The multimodal model struggles with all exercises, reflecting difficulties in effectively integrating text and image modalities.

We use LIME [27] to enhance model interpretability for text classification. It provides localized explanations for the predictions, improving transparency. In Table 4, CamemBERT, incorporating CW and DA, achieved the best results, prompting us to perform LIME analysis on our French text data. Fig. 3 illustrates the top 10 words for each class, with their importance scores, where the Y-axis displays the top 10 contributing words and the X-axis shows their weights.

For the *Essential* class, Family names and firstnames like "Zélie", "Bastide", and "Antoine" appear in the top features. Besides, the feature "i", which is the number 1 in Roman letters, indicates that these texts are closely related to key educational elements like multiple-choice questions. Interestingly, for the *Informative* class, all top features—"art" (art), "miroir" (mirror), "terre" (earth), "princesse" (princess), "Fourmis" (ants), "terminaison" (ending), "Courgette" (zucchini), "Baleine" (whale), and "Farel" (a proper noun, likely a name) are nouns, with the exception of "1881", which is a number. Unlike the other two classes, no verbs are present in this class. Finally, it seems that the *Decorative* class contains more grammatical words like "ils" (they), "comme" (like), "ai" (have), and "on" (we) compared to the other two classes and also words related to grammar exercises ("adjectif" (adjective), "pluriel" (plural)).

Overall, the LIME analysis on CamemBERT reveals how the model differentiates between content in children’s textbooks. Higher weights for the *Informative* class indicate the model’s focus on contextually significant terms, while lower weights for the *Decorative* highlight its ability to identify less critical features. These insights are specific to the classification task, and domain expertise remains essential to fully understand the model’s decision-making process.

In addition to the promising results, it is essential to acknowledge the limitations of our study. We recognize the dataset’s limited size, which may affect the findings. However, given the novel nature of this work, we focused on assessing the feasibility of our methods. Encouraged by the positive outcomes, we plan to expand and validate our approach using larger datasets in future research. Additionally, due to the presence of private data in our dataset, we cannot share it with the community, hindering the reproducibility of our experiments. We are working on annotating publicly available data to share with the community.

7 Conclusion

Our study proposes an automatic system to classify textbook images into *Essential*, *Informative*, and *Decorative* classes, aiding inclusive education for visually impaired and NDD students. Using a French textbook dataset of 652 (text, image) pairs, we found that text-based models, particularly CamemBERT on French text data, outperformed multimodal methods. Further, the LIME analysis revealed that for *Essential* class, the family names and first names "Zélie" and "Bastide" appear in top features, as well as "i", likely denoting the numeral for multiple-choice questions. Interestingly, for the *Informative* class, we can see that only nouns are accounted as top features. Finally, it seems that the *Decorative* class contains more grammatical words and those related to grammar exercises. This approach ensures that only images with educational content are retained. In the future, we plan to expand the dataset, incorporating additional sources such as [23] and exploring alternative similarity measures. We aim to explore alternative text generation to enhance textbook adaptation for visually impaired and NDD students. Additionally, future work could introduce a new dataset of images paired with contextual descriptions for the visually impaired, which may be shared with the community.

References

1. Alpizar-Chacon, I., van der Hart, M., Wiersma, Z.S., Theunissen, L.S., Sosnovsky, S., Brusilovsky, P., Baraniuk, R., Lan, A.: Transformation of PDF textbooks into intelligent educational resources. In: Proceedings of the 2nd International Workshop on Intelligent Textbooks, 21st International Conference on Artificial Intelligence in Education (2020)
2. Babij, S., James, M.E., Veldhuizen, S., Rodriguez, C., Price, D., Kwan, M., Cairney, J.: Cumulative prenatal risk factors and developmental coordination disorder in young children. *Maternal and Child Health Journal* pp. 1–7 (2023)
3. Ch, D.R., Saha, S.K.: Generation of multiple-choice questions from textbook contents of school-level subjects. *IEEE Transactions on Learning Technologies* (2022)
4. Chau, H., Labutov, I., Thaker, K., He, D., Brusilovsky, P.: Automatic concept extraction for domain and student modeling in adaptive textbooks. *International Journal of Artificial Intelligence in Education* **31**, 820–846 (2021)
5. Craig, F., Savino, R., Trabacca, A.: A systematic review of comorbidity between cerebral palsy, autism spectrum disorders and attention deficit hyperactivity disorder. *European Journal of Paediatric Neurology* **23**(1), 31–42 (2019)
6. Delgado-Lobete, L., Santos-del Riego, S., Pértega-Díaz, S., Montes-Montes, R.: Prevalence of suspected developmental coordination disorder and associated factors in spanish classrooms. *Research in Developmental Disabilities* (2019)
7. Gerald, T., Ettayeb, S., Le, H.Q., Vilnat, A., Paroubek, P., Illouz, G.: An annotated corpus for abstractive question generation and extractive answer for education. In: *Conférence sur le Traitement Automatique des Langues Naturelles* (2022)
8. Gong, C., Tao, D., Maybank, S.J., Liu, W., Kang, G., Yang, J.: Multi-modal curriculum learning for semi-supervised image classification. *IEEE Transactions on Image Processing* **25**(7), 3249–3260 (2016)
9. Hakami, W.A.S., Al-Aama, A.Y.: A framework to improve web form accessibility for the visually impaired. *IEEE Access* (2023)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *IEEE conference on computer vision and pattern recognition* (2016)
11. Jiang, Z., Zhang, Y., Li, X.: Moocon: a framework for semi-supervised concept extraction from mooc content. In: *Database Systems for Advanced Applications: DASFAA 2017 International Workshops: BDMS, BDQM, SeCoP, and DMMOOC*, Suzhou, China, March 27-30, 2017, Proceedings 22. pp. 303–315. Springer (2017)
12. Kenton, J.D.M.W.C., Toutanova, L.K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of NAACL-HLT*. pp. 4171–4186 (2019)
13. Kielblock, S., Woodcock, S.: Who’s included and who’s not? an analysis of instruments that measure teachers’ attitudes towards inclusive education. *Teaching and Teacher Education* **122**, 103922 (2023)
14. Klein, G., Kim, Y., Deng, Y., et al.: OpenNMT: Open-source toolkit for neural machine translation. In: *Association for Computational Linguistics - System Demonstrations* (2017)
15. Lincker, É., Guinaudeau, C., Pons, O., et al.: Noisy and unbalanced multimodal document classification: Textbook exercises as a use case. In: *20th International Conference on Content-based Multimedia Indexing* (2023)
16. Lincker, E., Pons, O., Guinaudeau, C., et al.: Layout-and activity-based textbook modeling for automatic pdf textbook extraction. In: *Intelligent Textbooks 2023*

17. Martin, L., Muller, B., Suárez, P.J.O., et al.: Camembert: a tasty french language model. In: Annual Meeting of the Association for Computational Linguistics (2020)
18. Missiuna, C., Rivard, L., Pollock, N.: They're bright but can't write: Developmental coordination disorder in school aged children. *Teaching Exceptional Children Plus* **1**(1), n1 (2004)
19. Miyauchi, H.: A systematic review on inclusive education of students with visual impairment. *Education sciences* **10**(11), 346 (2020)
20. Morris, D., Müller-Budack, E., Ewerth, R.: Slideimages: a dataset for educational image classification. In: *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II* 42. pp. 289–296. Springer (2020)
21. Mugambi, M.M.: Approaches to inclusive education and implications for curriculum theory and practice. *International Journal of Humanities Social Sciences and Education* **10**(4), 92–106 (2017)
22. Nowell, K.P., Bodner, K.E., Mohrland, M.D., Kanne, S.M.: Neurodevelopmental disorders. (2019)
23. Otto, C., Springstein, M., Anand, A., Ewerth, R.: Characterization and classification of semantic image-text relations. *International Journal of Multimedia Information Retrieval* (2020)
24. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. *the Journal of machine Learning research* **12**, 2825–2830 (2011)
25. Qu, L., Liu, M., Wu, J., Gao, Z., Nie, L.: Dynamic modality interaction modeling for image-text retrieval. In: *44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2021)
26. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning* (2021)
27. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: *International conference on knowledge discovery and data mining* (2016)
28. Sanchez, C.A., Wiley, J.: An examination of the seductive details effect in terms of working memory capacity. *Memory & cognition* **34**, 344–355 (2006)
29. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society (2015)
30. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: *IEEE conference on computer vision and pattern recognition* (2016)
31. Thiyam, J., Singh, S.R., Bora, P.K.: Chart classification: a survey and benchmarking of different state-of-the-art methods. *International Journal on Document Analysis and Recognition (IJDAR)* **27**(1), 19–44 (2024)
32. Tian, Z., Flanagan, B., Dai, Y., Ogata, H.: Automated matching of exercises with knowledge components. In: *30th International Conference on Computers in Education Conference Proceedings*. pp. 24–32 (2022)
33. Wambaria, M.W.: Accessible digital textbook for learners with disabilities: Opportunities and challenges. *The Educational Review, USA* **3**(11), 164–174 (2019)