DATASET CHALLENGES

We identified several key challenges within each modality of our dataset, which are outlined below.

A. Challenges in Cover Image

The complexity of book cover images poses challenges in accurately identifying book genres.

a) Background: Book cover image background plays a crucial role in setting the tone or mood, sparking interest, and assisting readers in determining whether the book aligns with their tastes and preferences. For instance, *sci-fi & fantasy* book covers frequently showcase detailed background elements, featuring enchanting landscapes and vivid colors, designed to captivate readers with a sense of wonder and adventure (Fig. 1:*viii*). Based on the visual cues available in the background of the cover image, we can group it as follows:

b) Limited Visual Cues: Sometimes, the cover page provides minimal or no visual information, containing only text in standard fonts, making genre identification challenging (Fig.s 1:*i-ii*).

c) Moderate Information: The cover page often includes moderate visual features to represent some of its genres but may fail to capture others due to the multi-label nature of genres (Fig.s 1:*v-viii*).

d) Complex Background: In some instances, the background of a book cover image becomes convoluted due to the presence of extensive or composite visual effects or elements (Fig.s 1:*iii-iv*).

e) Foreground: The foreground of a book cover image plays a vital role in capturing reader attention. However, foreground images on book covers pose challenges for genre identification due to ambiguity, cross-genre similarities, abstract designs, stylistic variations, cultural influences, evolving trends, and marketing-driven misrepresentations.

f) Living/ Non-living Element: The foreground image can effectively communicate the book theme or atmosphere by strategically placing engaging foreground elements. Some non-living elements (e.g., mountain, sea, etc.) are typically associated with genres like *travel*. However, in some cases (Fig. 1:*x*), such elements are used to represent genres like *literature*, *mystery & thriller & suspense & horror & adventure*, *romance*, and *sci-fi & fantasy*, which creates a disconnect and makes genre identification challenging. Similarly, living objects contribute to this ambiguity. For example, while an animal on the cover often implies a connection to genres like *animals & wildlife & pets*, there are instances (Fig. 1:*xi-xii*) where this association does not hold, making genre identification more challenging.

g) Scene Image: Scene-based images on book covers further complicate genre prediction due to their intricate visual details, clutter, and lack of a cohesive composition (Fig.s 1:*xiii-xvi*). Unlike single-object covers, scene-based images often contain multiple elements—characters, landscapes, objects, or action sequences—that may belong to different genres. This complexity increases ambiguity, as the dominant visual theme may not be immediately apparent. Additionally, certain scenes may be common across multiple genres, making classification more challenging. Variations in artistic styles, lighting, and color palettes further contribute to genre ambiguity.

h) Inter-Variation: Sometimes, books from different genres may contain similar visual information (Fig.s 1:*xvii-xviii*, and Fig.s 1:*xix-xx*). This may be intentional to challenge reader expectations, create intrigue, or highlight genre blends. For example, cover image of Fig. 1:*xvii* comprises genres *childrens' book*, *comics & graphics*, *sci-fi & fantasy*, *teen & young adult*, but another book's cover image (Fig 1:*xviii*) with similar visual cues belongs to *arts & photography*.

i) Intra-Variation: The visual styles and design elements used in cover images within a specific genre can vary widely. This variation is often influenced by the author/artist's unique vision and the intended mood of the book, resulting in diverse interpretations of the same genre. Additionally, publishers frequently adhere to specific house styles or branding guidelines, incorporating consistent visual elements or color schemes across their catalog. While this ensures some level of uniformity within a publisher's collection, the broader diversity in design choices makes genre identification more challenging (Fig.s 1:*xxi-xxii*, and Fig.s 1:*xxiii-xxiv*).

j) Collage: In some cases, book covers integrate multiple aforementioned background and foreground elements, forming a collage of numerous smaller images. This visual complexity often leads to information overload, making it challenging to accurately determine the book's genre (Fig.s 1:*xxv-xxviii*).

k) Number/ Letter as Image: Occasionally, book covers feature numbers or letters stylized as visual elements. This artistic choice is often intended to emphasize a theme, establish a distinctive style, or highlight key information related to the book's content. However, such designs pose a unique challenge for genre identification, as the graphical representation of text or numbers can obscure their intended meaning, making interpretation more difficult (Fig.s 1:*xxix-xxx*).

l) Unclear Image: Sometimes, book covers feature images that are blurry, hazy, or of low resolution. Extracting meaningful features from such images becomes difficult, making accurate genre identification challenging (Fig.s 1:*xxxi-xxxii*).

1) Challenges in Cover Text: We extracted text from book cover images using an OCR (Optical Character Recognition) engine. However, these extracted texts present additional challenges, which we summarize below.

a) Limited Text: Some book covers feature only minimal text, such as the author name, or the book title, without supplementary elements like subtitles, or descriptive taglines. These additional details often play a crucial role in genre identification. The absence of such details forces the OCR system to rely solely on sparse information, significantly increasing the risk of misclassification (Fig.s 2:*i-iv*).

b) Linguistic Issues: OCR systems are often designed with a primary focus on specific languages, e.g., English. When book covers feature text in non-English languages, these systems may face difficulties in accurately recognizing and processing the characters. Errors in language detection or character recognition can lead to misinterpretations, further complicating text-based genre classification (Fig. 2:*v-viii*).

Fig. 1: Examples of some challenging book cover images with mentioned issues

Fig. 2: Examples of some challenging cases involving cover text

c) Low Resolution Cover Page: Low-quality and glossy book covers present significant challenges for accurate text extraction using OCR. Glossy surfaces can create glare and reflections when photographed or scanned, interfering with character recognition. Additionally, low-resolution images often result in blurred text, making it difficult to distinguish individual characters. These factors contribute to errors in text extraction and misinterpretation (Fig.s 2:*ix-xii*).

d) Complex background: Colorful or busy backgrounds on book covers present a significant challenge for OCR systems. These complex backgrounds can interfere with the OCR process in several ways:

e) Visual Clutter: A busy background with multiple colors, patterns, or images can create visual clutter. This clutter makes it difficult for the OCR system to distinguish the text from the background. The presence of various elements can cause the OCR algorithm to misidentify parts of the background as text or fail to recognize the text altogether (Fig.s 2:*xiii-xvi*).

f) Color Contrast: Text on colorful backgrounds might not have sufficient contrast. When the text color closely matches the background colors, the OCR system struggles to differentiate between them. High contrast between text and background is crucial for accurate OCR, and colorful backgrounds often fail to provide this (Fig.s 2:*xiii-xvi*).

B. Challenges in Blurb

The blurb often provides genre-related cues, but accurately identifying genres from it poses several challenges, as outlined below.

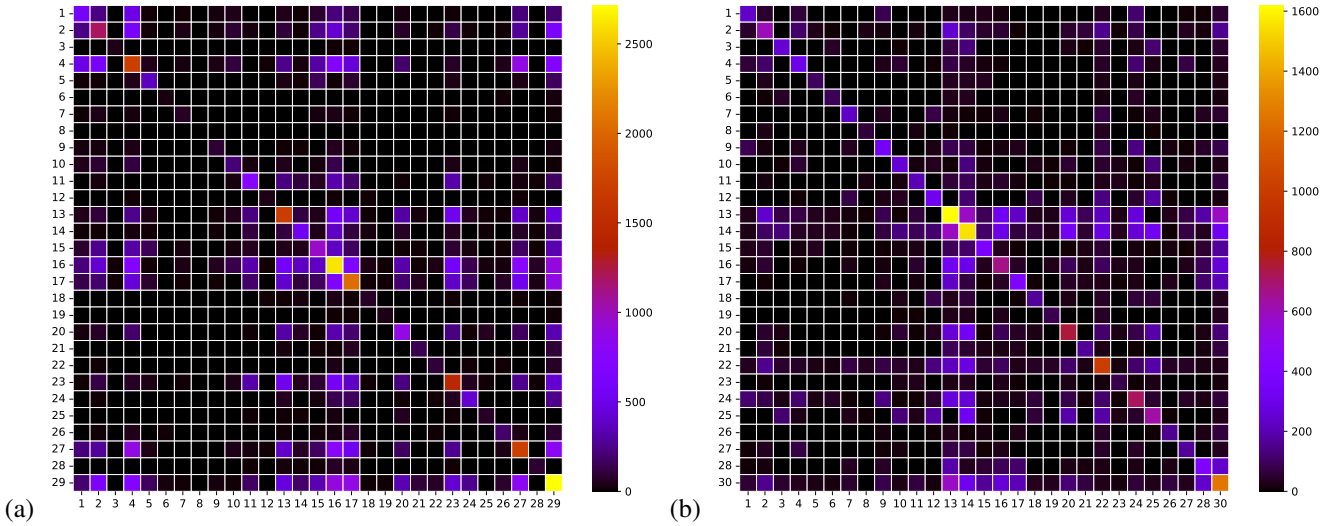

Fig. 3: Co-occurrence matrices for book genres: (a) *fiction*, (b) *non-fiction*

a) Insufficient Information: Several books include blurbs that lack sufficient detail, making it difficult to accurately determine their genres.

b) Irrelevant Information: We encounter many books with blurbs that contain irrelevant information. These blurbs often include vague or generic statements, offering little insight into the genre and making accurate genre identification challenging.

c) Multilinguality: Blurb texts may appear in multiple languages, each with distinct scripts and sentence structures, creating challenges in designing a framework that can uniformly interpret and process multilingual content.

C. Challenges of Metadata

There are several challenges for identifying book genres using book metadata information. We summarized those challenges as follows.

a) Surface-Level Information: Metadata doesn't reveal a user's deeper interests, reading level, or genre preferences. One author writes books on different genres. So, someone who enjoys a particular author might not like all of their books.

D. Challenges of Ground-Truthing

Establishing reliable ground-truth genre labels presented significant challenges due to the inherent subjectivity and complexity of the task.

a) Partial Understanding by Human Annotators: Since it is often impractical for annotators to read an entire book, the ground-truthing process primarily relied on linguistic experts who assigned genre labels based on limited content—such as publisher blurbs, user reviews (e.g., from Goodreads), and selected excerpts. This partial exposure may lead to a superficial or incomplete understanding of the book's thematic nuances, thereby impacting the accuracy of genre labeling.

b) Subjectivity and Inconsistency in Genre Interpretation: Genre classification is often influenced by subjective interpretation. Annotators might interpret the same book differently based on the emphasis placed on particular elements. For instance, a book that blends psychological drama and crime might be labeled as *thriller* by one expert, and as *mystery* or *drama* by another, depending on their reading perspective.

c) Fiction vs. Non-fiction Ambiguities: Differentiating between fiction and non-fiction genres was particularly challenging for works related to well-known fictional universes. For example, while a "*Harry Potter*" novel is clearly fiction, a companion book like "*Harry Potter – Page to Screen: The Complete Filmmaking Journey*" is non-fiction, despite sharing the same universe and characters. Such genre-conflicting edge cases complicated the labeling process, requiring careful contextual analysis.

d) Multi-Label Overlaps and Unclear Boundaries: Many books span multiple genres, making it difficult to determine which labels are most appropriate and at what level of the hierarchy.

e) Lack of Standardized Taxonomy: Genre taxonomies vary significantly across publishers, retailers, and literary databases, resulting in inconsistent ground-truth references. Aligning expert annotations with a unified genre hierarchy was non-trivial and often required manual reconciliation.

E. Challenges of Overall Dataset

In preparing a high-quality dataset for hierarchical multi-label book genre classification, several overarching challenges emerged, particularly due to the multi-label nature of the task and the complexity of real-world genre distributions.

a) Class Imbalance in Multi-Label Context: The dataset exhibits significant class imbalance, with certain genres being highly represented, while others appear far less frequently. However, unlike single-label classification, this imbalance cannot be easily corrected through traditional resampling techniques because books often belong to multiple genres

simultaneously. Balancing one under-represented genre could unintentionally disrupt the co-occurrence patterns with more frequent genre.

b) Constraints in Data Augmentation: Augmentation techniques commonly used to balance datasets pose unique risks in the multi-label setting. When augmenting samples from a genre with few examples, the co-occurring genres in the original sample, many of which may already be overrepresented, also get replicated. This reinforces existing imbalances and biases, making it difficult to selectively boost specific genre classes without inadvertently inflating others. Fig.s 3: (a), (b) visually depict the relationships and co-occurrence patterns among book *fiction* and *non-fiction* genres.

c) Sparse Label Combinations: The multi-label setting also leads to a large number of unique label combinations, many of which appear very infrequently. This sparsity in the label space challenges the model's ability to generalize well to unseen or rare combinations and limits the effectiveness of frequency-based heuristics.

d) Dependency Between Genre Labels: Usually, genre labels are not mutually independent; the presence of one genre may often influence the likelihood of another. Modeling these dependencies becomes increasingly difficult as the number of labels grows, especially under imbalance and sparsity, and it further complicates synthetic balancing or sampling strategies.

Data Augmentation

We observed that some specific genres within both fiction and non-fiction categories were underrepresented in our dataset. To address this imbalance, we applied data augmentation techniques. Each book sample in our dataset consists of a cover image, cover text, blurb, metadata, and genre labels. For augmentation purposes, we focused on the cover image, cover text, and blurb, while keeping metadata and genre labels unchanged. We now discuss our data augmentation strategies.

1) Visual Data Augmentation: We augmented the cover images using SDEdit [1], a diffusion-based generative model. SDEdit introduces noise to an input image and then refines it to produce high-quality synthetic variations, preserving the semantic structure while diversifying the visual representation.

2) Textual Data Augmentation: For cover text and blurb augmentation, we leveraged the Gemini large language model [2]. By applying prompt engineering, we guided the model to generate alternative versions of the text that retained the contextual integrity and relevance of the original content.

All augmented samples were manually reviewed and validated by human annotators. As a matter of fact, these augmented data instances were used solely during the model training phase to enhance generalization and robustness.

Statistical Information

Table I presents a statistical overview of each input modality, both before and after augmentation. The reported metrics include the minimum, maximum, mean, median, and standard deviation (SD). The measured attributes are area of the cover page image (in pixel^2), word counts from the blurb and cover text, and the number of books authored or published as derived from the metadata.

TABLE I: Multimodal data statistics

Modality	Attribute	Minimum	Maximum	Mean	Median	SD
Cover image	Image area before augmentation	20460	16301712	435282.41	146775	1125224.73
	Image area after augmentation	20460	16301712	325205.84	135850	980565.34
Blurb	#Words before augmentation	0	1786	120.52	98	102.37
	#Words after augmentation	0	1953	110.01	96	80.89
Cover text	#Words	1	944	20.67	15	27.69
Metadata	#Books written by an author	1	43	1.57	1	1.91
	#Books published by a publisher	1	268	5.80	1	16.64

References

[1] C. Meng *et al.*, "SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations," *arXiv:2108.01073*, 2021.

[2] R. Anil *et al.*, "Gemini: A family of highly capable multimodal models," *arXiv:2312.11805*, 2023.