CoSMo: A Multimodal Transformer for Page Stream Segmentation in Comic Books

Marc Serra Ortega

30 de juny de 2025

Resum– Aquest article presenta **CoSMo**, un nou Transformer multimodal per a la Segmentació de Fluxos de Pàgines (PSS) en còmics, una tasca crucial per a la comprensió automàtica de continguts, ja que és una etapa inicial necessària per a moltes tasques posteriors com l'anàlisi de personatges, la indexació de la història o l'enriquiment de metadades. Formalitzem la PSS per a aquest mitjà únic i creem un nou conjunt de dades anotat de 20.800 pàgines. CoSMo, desenvolupat en variants només visuals i multimodals, supera constantment els models base tradicionals i models de llenguatge visual de propòsit general significativament més grans, en F1-Macro, Qualitat Panòptica i mètriques a nivell de flux. Els nostres resultats destaquen el domini de les característiques visuals per a la macroestructura PSS en còmics, tot demostrant els beneficis multimodals en la resolució d'ambigüitats complexes. CoSMo estableix un nou estat de l'art, obrint el camí per a l'anàlisi escalable de còmics.

Paraules clau– Segmentació de Fluxos de Pàgines, Còmics, Transformer Multimodal, Aprenentatge Profund, Anàlisi de Documents, Visió per Computador, CoSMo

Abstract–This paper introduces **CoSMo**, a novel multimodal Transformer for Page Stream Segmentation (PSS) in comic books, a critical task for automated content understanding, as it is a necessary first stage for many downstream tasks like character analysis, story indexing, or metadata enrichment. We formalize PSS for this unique medium and curate a new 20,800-page annotated dataset. CoSMo, developed in vision-only and multimodal variants, consistently outperforms traditional baselines and significantly larger general-purpose vision-language models across F1-Macro, Panoptic Quality, and stream-level metrics. Our findings highlight the dominance of visual features for comic PSS macro-structure, yet demonstrate multimodal benefits in resolving challenging ambiguities. CoSMo establishes a new state-of-the-art, paving the way for scalable comic book analysis.

Keywords– Page Stream Segmentation, Comic Books, Multimodal Transformer, Deep Learning, Document Analysis, Computer Vision, CoSMo

1 INTRODUCTION

Comics are a globally appreciated medium, enjoyed by readers of all ages—from Italian children collecting issues of *Topolino*, to fans of American classics like *Plastic Man*, to global audiences captivated by Japanese manga such as *Naruto*. Beyond their inherent entertainment, comics serve as invaluable cultural artifacts, reflecting and chronicling societal values, historical moments, and evolving norms across different eras and geographies. Often published weekly or monthly, these books appear in anthology formats: short, self-contained stories from different authors and genres co-existing in a single issue, offering narrative continuity for ongoing plots while exposing readers to new styles and creators. These volumes are later recompiled into collected editions, but the original anthology structure remains a key publishing unit across decades and cultures.

Much of this heritage has fortunately not been lost. Tens of thousands of historical comic books have been digitized and are now preserved in public repositories such as the *Digital Comics Museum*[1] and *Comic Book Plus*[2]. Crucially, these scanned books are enriched through manual annotati-

• E-mail de contacte: Marc.SerraO@autonoma.cat
• Menció realitzada: Computació
• Treball tutoritzat per: Dimosthenis Karatzas and Emanuele Vivoli
• Curs 2024/25

[1] https://digitalcomicmuseum.com/
[2] https://comicbookplus.com/

Fig. 1: Automated Page Stream Segmentation in Comics. Our model accurately identifies semantic page types in comic book streams for robust content analysis.

on pipelines. Volunteers link each book to *comics.org*[3], the most comprehensive structured metadata archive in comics, which catalogs story titles, authors, page spans, characters, and editorial credits. These connections are made manually: a mediator retrieves metadata, identifies story boundaries, and tags each page accordingly. While this human effort has enabled large-scale archival access, it is unsustainable at scale. Automating this process is vital for preserving cultural heritage and enabling intelligent search and retrieval across massive comic corpora.

Comics, however, are particularly challenging to analyze computationally. Unlike text-dominant documents, they combine richly illustrated content with stylized typography and non-linear panel layouts. Narrative elements span modalities—text, image, spatial layout—and vary dramatically across eras and cultures. Speech balloons overlap with characters; onomatopoeia can dominate the composition; panel layouts often exhibit non-linear reading paths, diverging from standard conventions such as the top-to-bottom, left-to-right flow of American and European comics, or the right-to-left sequence common in Japanese manga. As a result, comics defy standard document parsing pipelines and instead require models capable of robust multimodal reasoning.

In this paper, we address the task of Page Stream Segmentation (PSS) for comics, as shown in Fig. 1, the automatic division of scanned books into semantically meaningful sequences of pages, such as stories, advertisements, or text-only inserts. PSS is foundational—without it, downstream tasks like character analysis, story indexing, or metadata enrichment operate on misaligned page groups. While prior work has explored PSS for traditional documents, its adaptation to comics has remained unexplored due to the unique structure and multimodality of the medium.

To this end, we introduce **CoSMo**, a Transformer-based encoder architecture tailored for PSS in comics. We construct a dataset of 430 annotated comic books sourced from public archives. Recognizing that publicly available metadata from *comics.org*) often contains inconsistencies or misalignments with scanned content, we performed extensive manual validation and annotation to ensure high-quality labels for our task. Through extensive experimentation, including ablation studies on multimodal inputs and benchmarking against various machine learning and deep learning approaches, CoSMo consistently demonstrated superior performance. Our lightweight models notably outperformed all baselines across all metrics, even against zero-shot evaluations of Multimodal Large Language Models (MLLMs) an order of magnitude larger.

Our main contributions are:

- We formalized the PSS task for comic books and created a high-quality annotated dataset, meticulously aligned with *comics.org* metadata.
- We develop CoSMo, a modular Transformer encoder available in a lightweight vision-only configuration and a robust multimodal variant.
- We perform extensive experiments and ablations to rigorously evaluate the impact of diverse representations and benchmark against a wide array of baselines, showcasing CoSMo's superior performance compared to traditional methods and zero-shot multimodal LLMs.

All code, data, and annotations are available in GitHub[4] to empower the comics community to automatically generate accurate annotations for their extensive corpus of comic books, a task our method performs with an impressive error rate of approximately 1%.

2 Objectives

This section defines the Page Stream Segmentation (PSS) task for comic books and outlines our core research objectives.

2.1 Problem Definition

Page Stream Segmentation (PSS) for comic books is fundamentally defined as the task of identifying semantic boundaries to form coherent and meaningful groups of subsequent pages within a sequential stream of scanned or digitized comic pages. In this work, we reframe this as

[3] https://www.comics.org/

[4] https://github.com/mserra0/CoSMo-ComicsPSS

a single-page classification problem, which involves assigning a semantic label to each page. Given a document stream represented as an ordered sequence of n pages: $\mathcal{S} = \{p_1, p_2, \ldots, p_n\}$, the goal is to learn a segmentation function $f_\theta : \mathcal{S} \rightarrow \{y_1, y_2, \ldots, y_n\}$, where each $y_i \in \mathcal{Y}$ is the class label associated with page p_i, and $\mathcal{Y}$ is the set of predefined semantic categories: $\mathcal{Y} = \{\text{Cover}, \text{Advertisement}, \text{Story}, \text{Text Story}\}$.

This problem is cast as a multiclass sequence labeling task, where each prediction may depend not only on the visual and textual content of the current page but also on the broader narrative context across the page stream. The model must integrate multimodal cues—visual layout, stylistic features, and text—across the sequence to produce accurate and coherent segmentations. This formulation grounds PSS in a structured learning framework and motivates the specific research objectives described below.

2.2 Research Objectives

To guide our investigation, we define the following goals:

- **Establish Baselines for Comic-Book PSS:** Implement simple reference models to characterize the task's complexity and assess their limitations in the absence of prior benchmarks.
- **Curate Dataset:** Introduce a high-quality, human-checked, labeled dataset for PSS in Comics.
- **Develop an End-to-End Model (CoSMo):** Design and implement a Transformer-based architecture that processes full page streams to leverage long-range narrative dependencies for segmentation.
- **Explore Comic Book Multimodal Representations:** Investigate how different modeling approaches capture the unique visual, structural, and narrative characteristics of comic books.
- **Assemble a comprehensive evaluation suite:** Define and apply metrics at the page, document, and stream levels to holistically evaluate segmentation quality and real-world utility.

3 State of the Art

Comic Books. The fast-growing literature on computational comics understanding has recently been condensed by Vivoli et al.[21], introducing the *"Layer of Comics Understanding (LoCU)"* framework to situate vision–and–language tasks along the axes of input/output modality and cognitive complexity. Their survey exposes three systemic bottlenecks—limited open datasets, weak reproducibility, and an over-reliance on single-page, low-level vision settings. In response, they released two complementary resources: the *Comics Dataset Framework (CDF)* [19], which unifies annotation schemas to facilitate reproducibility, and *CoMix* [18], a multitask benchmark that probes previously unexplored single-page capabilities such as character naming and dialogue generation. Building on this foundation, the *Magi* series pushes analysis into higher *"LoCU layers"*, advancing detection, character re-identification, and narrative generation for manga pages [16, 14, 15]. Beyond the single-page regime, multimodal reasoning across panel sequences has been explored via the text-cloze task [7, 20]. Yet, to date, no method ingests an entire volume—crucial for book-level understanding, metadata extraction, or coherent page grouping. Closing this gap is the central objective of the present work.

Page Stream Segmentation. PSS is a fundamental task in document analysis, aimed at segmenting a continuous sequence of pages into coherent document units (e.g., individual issues, stories, or sections). This task has been extensively explored in domains such as banking, business, and legal workflows. Early methods heavily relied on OCR to extract text-based features, which were then passed to traditional machine learning models [2, 22, 9, 4]. Over time, these approaches evolved into multimodal systems combining CNNs for visual input and RNNs for textual input, enabling the models to learn richer representations of page structure and semantics [2]. More recently, Transformer-based architectures have been applied in PSS too, revealing that decoder-only large language models (LLMs), fine-tuned with parameter-efficient techniques, can leverage textual features alone to surpass smaller, dedicated multimodal PSS approaches [5]. Standardisation has followed suit. OpenPSS [6] unifies earlier task formulations—spanning Tab This Folder [9] and the panoptic-segmentation literature [8]—into a single benchmark with harmonised document- and stream-level metrics, enabling apples-to-apples comparison across domains. Despite this progress, comics remain conspicuously absent from the dedicated PSS corpus.

4 Methodology

In this section, we introduce **CoSMo** (Comic Stream Modeling), our novel transformer-encoder model for robust Page Stream Segmentation (PSS) in comic books. CoSMo's design explicitly tackles the unique multimodal challenges inherent in comic book documents by integrating both visual and textual information within a sequential context. To offer both comprehensive and cost-effective solutions, CoSMo is developed in two primary variants: a **multimodal model** that leverages visual and textual cues, and a **vision-only model** optimized for scenarios with textual processing limitations. Both variants share a common architectural inspiration from the successful encoder-only Transformer design.

4.1 *CoSMo*

4.1.1 Multimodal Architecture

The *CoSMo* **Multimodal** architecture, as shown in Figure 2, integrates visual features with contextualized textual embeddings, feeding them into a Transformer encoder for sequence processing.

Visual Feature Extraction: Visual features are obtained from each page using a frozen **SigLIP** backbone. SigLIP is chosen for its strong performance in general image representation, providing robust visual embeddings.

Textual Feature Extraction: We extracted rich, contextualized OCR from Qwen2.5-VL-32B, which provided a comprehensive textual understanding by identifying both

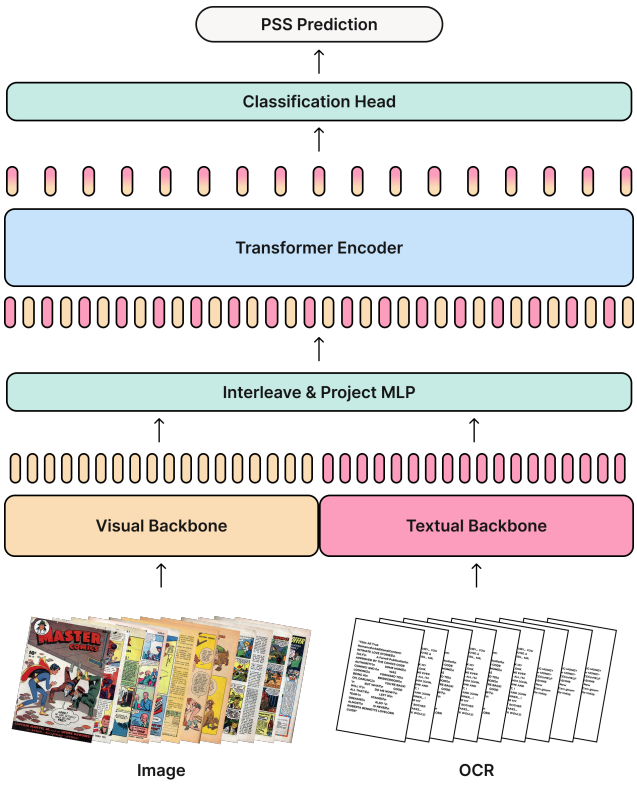

Fig. 2: Multimodal CoSMo model architecture.

the reading order and distinct text types, including *Titles*, *Panels*, and *Content Text*, particularly relevant for pages like *Text-Stories* and *Advertisements*. This structured OCR output was then embedded using Qwen3Embedding-0.6B [5][24]. This embedding model, specifically tailored for textual representation, effectively leverages the advanced semantic understanding capabilities of large language models.

Feature Projection and Interleaving: Before applying the multi-token arrangement, both visual and textual feature vectors are independently projected to a shared dimensionality of 768 using a three-layer MLP. This projection module incorporates GELU activations, dropout (rate = 0.4), and layer normalization to ensure stable and expressive feature transformations. After projection, the resulting embeddings are normalized to maintain consistent feature scaling across modalities. For interleaving, we construct a dual-token representation per page—one for the visual modality and one for the textual modality—allowing the Transformer encoder to jointly attend over both representations in sequence. This structure enables the model to capture complementary signals and cross-modal dependencies effectively.

Transformer Encoder: The page-level multimodal representations are processed by a encoder-only Transformer, which models contextual dependencies across the entire comic book sequence. The encoder consists of 4 Transformer layers with 4 self-attention heads per layer. It uses a hidden size of 256, an input embedding size of 768, and an intermediate feedforward dimension of 3072 (i.e., 4× the hidden size). To regularize training, a dropout rate of 0.4 is applied throughout the network. We adopt absolute positional encoding to preserve the ordering of pages within the book, which is essential for learning narrative flow and structural transitions. **Classification Head:** From the dual-token sequence, only the second token corresponding to each page is forwarded to the classification head. This head consists of a three-layer MLP that includes GELU activations, layer normalization, and a dropout rate of 0.4. It maps the contextualized representation of each page to one of the predefined semantic categories (e.g., story, cover, advertisement), enabling fine-grained classification across the comic stream.

4.1.2 Vision-Only Architecture

The *CoSMo* **Vision-Only** model offers a computationally efficient alternative to its multimodal counterpart by relying exclusively on visual information. This variant retains the same core Transformer encoder and classification head but omits the OCR extraction, textual embedding, and the interleaving components—substantially reducing model complexity and the need for OCR. The motivation behind this design stems from two key observations: first, high-quality OCR in comics is often computationally expensive due to complex layouts and stylized fonts; second, as shown in our experiments, visual features alone already capture much of the structural and semantic information required for accurate segmentation. Consequently, the Vision-Only CoSMo model presents a highly competitive and cost-effective solution for Page Stream Segmentation, achieving strong results with minimal trade-off in performance.

5 Dataset

Our dataset consists of 430 *classic* comic books sourced from the Digital Comic Museum (DCM)[6], a public-domain archive of *Golden Age comics*. We manually annotated these books and performed thorough quality checks to ensure accuracy. This collection contains over 20,800 pages, ensuring stylistic and structural diversity while supporting deep model training.

Each page is labeled with one of five semantic classes: *Cover*, *Advertisement*, *Text Story*, *Story*, or *First-Page*—the latter being a derived label marking the first page in a narrative block. These classes reflect common structural components of vintage comic books.

Comics typically begin with a *Cover*, followed by a mix of *Advertisements*, *Text Stories*, and one or more *Story* segments. Transitional pages often separate story blocks, but this structure varies: some books feature continuous narratives, others have only one or two categories.

The dataset poses several challenges:

Intra-class diversity: Pages in the same class (e.g., *Advertisement*) differ greatly in layout and style as illustrates in Figure 3.

Inter-class similarity: *Text Stories* and text-heavy *Advertisements* are often visually similar. Likewise, *First-Page* and *Story* pages can be hard to distinguish without narrative context.

Class imbalance: *Story* pages dominate the dataset (71%), while other classes are underrepresented: 2.4% *Cover*, 8.8% *Advertisement*, 13.4% *First-page* and 4.2% *Text-story*.

[5] https://hf.co/Qwen/Qwen3-Embedding-0.6B

[6] https://digitalcomicmuseum.com/

Fig. 3: Examples of annotated comic book page types used for multiclass page stream segmentation. *First-Page* is a derived label marking story segment beginnings.

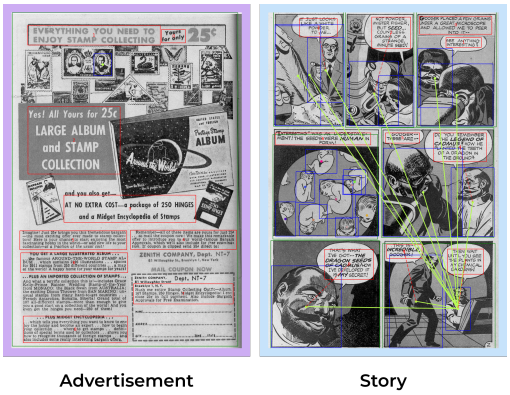

Fig. 4: Detection examples from the MAGI model.

6 Experiments

This section outlines the experimental framework used to evaluate CoSMo and several baselines for the PSS task. We describe both single-page and multi-page modeling paradigms, including variations in input modalities and backbone architectures, to systematically explore the challenges of segmenting comic book pages.

6.1 Baselines

Handcrafted Detection Features with Traditional ML. We began by extracting semantic layout elements—characters, faces, panels, text blocks—from each page using the object detection module from MAGI, an example of the detected elements is shown in Fig 4.

Building upon these detections, 20 handcrafted features were extracted for each page. These features, designed to characterize content density, spatial layout, and element distributions, were crucial for discerning structural transitions within the sequence. We then utilized these features to train an XGBoost model.

Pretrained Visual Backbones To assess visual representation quality, we tested CLIP and SigLIP under two regimes:

(i) Zero-shot classification via prompt-based inference, and **(ii)** Linear probing using a lightweight MLP trained on frozen visual embeddings.

These experiments aimed to evaluate both the transferability of vision–language models to the comic domain and their capacity for fine-grained semantic separation. Prompt

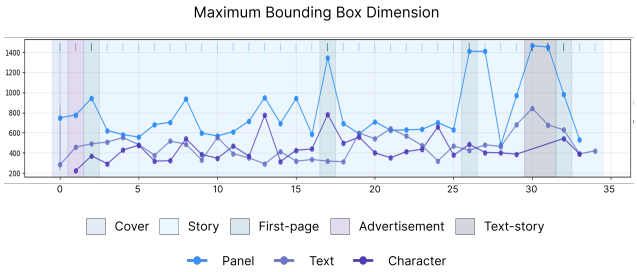

Fig. 5: Evolution of handcrafted features across pages in a comic book.

tuning was used to improve zero-shot alignment with class labels. The linear probe setup allowed direct comparison of learned representations without full fine-tuning.

Zero-Shot LLM Experiments To evaluate large-scale instruction-tuned models in a zero-shot setting, we tested Qwen2.5-VL-32B [7][1] using two prompt formats: (i) a CLS-focused prompt focused on providing a short caption followed by a direct classification of the image, and (ii) an OCR-focused prompt that added the OCR extraction, encouraging text interpretation before classification. This helped explore how prompt design affects segmentation accuracy and whether massive models can generalize to comic PSS without task-specific fine-tuning.

6.2 Multi-page CoSMo

Our base model, CoSMo, operates on frozen visual embeddings from backbone models and serves as the foundation for all multi-page experiments.

Visual Backbone Ablation Study. We evaluated four pretrained visual backbones—CLIP, SigLIP, SigLIP2, and DINOv2—within the CoSMo architecture. These models differ in training paradigms and input resolution.

Detection Feature Fusion. To incorporate structural layout cues, we fused handcrafted features into the best visual-only CoSMo model (SigLIP backbone). Features were projected, normalized, concatenated with visual embeddings, and passed through an MLP before being processed by the Transformer encoder. This tested whether explicitly injected layout signals improve segmentation of ambiguous classes.

Multi-Backbone Fusion. We tested two strategies for combining the backbones in compared in the Visual Backbone Study: (1) *Fusion*, where projected embeddings were concatenated and passed through a two-layer MLP; (2) *Multi-token*, where each embedding was treated as a separate input token. These experiments assessed whether visual diversity leads to complementary representations and performance gains.

Text-Only Modeling. In our Text-Only Modeling approach, textual features, obtained via the strategy detailed in Section 4.1.1, constituted the sole input to CoSMo. This configuration allowed us to assess the efficacy of OCR-based representations in distinguishing fine-grained page roles, especially under conditions where visual cues are ambiguous.

Multimodal CoSMo Finally, the full multimodal CoSMo variant was evaluated by integrating SigLIP visual em-

[7] https://hf.co/Qwen/Qwen2.5-VL-32B-Instruct

TABLE 1: F1-MACRO AND ACCURACY FOR MULTICLASS SINGLE-PAGE CLASSIFICATION UNDER DIFFERENT TRAINING SETUPS.

Model	Modality	Setup	F1-Macro ↑	Accuracy ↑
XGBoost	Detection	End-to-End	83.50	90.76
CLIP	Vision + Prompt	Zero-Shot	61.90	80.30
SigLIP	Vision + Prompt	Zero-Shot	62.70	84.90
CLIP + MLP	Vision	Linear Probe	80.64	90.59
SigLIP + MLP	Vision	Linear Probe	89.92	95.81
Qwen (CLS Prompt)	Vision + Prompt	Zero-Shot	87.12	92.25
Qwen (OCR Prompt)	Vision + Prompt	Zero-Shot	88.26	93.84
CoSMo	Vision	End-to-End	93.46	96.16
CoSMo	Multimodal	End-to-End	97.82	98.41

beddings with Qwen-derived OCR embeddings, these modalities were combined using two distinct strategies—fused and multitoken—before being encoded through the same Transformer backbone.

6.3 Evaluation Metrics

To rigorously evaluate the performance of our Page Stream Segmentation (PSS) models, we employ a combination of page-level and document-level metrics tailored to both single-page and multi-page modeling scenarios, as well as providing standardized metrics in the PSS field [6]. These metrics assess not only classification accuracy but also the quality of sequential predictions, which are crucial in structured documents such as comic books.

6.3.1 Single-Page Level Evaluation

For single-page modeling, the classification task is framed as a multi-class problem with significant class imbalance. To address this, we adopt the **Macro-averaged F1 score (F1-Macro)** as our primary evaluation metric. This metric is consistently used across all experiments as our primary metric.

The F1-Macro score is defined as the harmonic mean of precision and recall, calculated for each class and then averaged across all classes. The formula for F1-Macro is:

$$F1_{\text{Macro}} = \frac{1}{N} \sum_{i=1}^{N} \frac{2\text{P}_i\text{R}_i}{\text{P}_i + \text{R}_i} \tag{1}$$

where: N is the number of classes, P_i and R_i are the precision and recall for class i, respectively. This formulation ensures fair evaluation across rare and frequent classes. Additionally, we analyze the confusion matrix to gain insights into class-specific prediction strengths and weaknesses.

6.3.2 Multi-Page Level Evaluation

The evaluation of multi-page models requires sequence-aware metrics that assess not only the correctness of individual page classifications but also the coherence and accuracy of entire document structures.

Document-Level Metrics: We implement **Document-Level F1**, as proposed in prior works [6, 9], which computes F1 scores at the level of contiguous semantic segments (e.g., stories or advertisement blocks), rather than individual pages. Furthermore, inspired by the Panoptic Segmentation task in computer vision [8], we adopt a unified metric called **Panoptic Quality (PQ)**, which measures both recognition quality and segmentation quality:

$$\text{PQ} = \text{DocF1} \times \text{SQ} \tag{2}$$

Where: **DocF1** is the harmonic mean of segment-level precision and recall, and **Segmentation Quality (SQ)** is defined as:

$$\text{SQ} = \frac{1}{|TP|} \sum_{(p,g)\in TP} \frac{|p \cap g|}{|p \cup g|} \tag{3}$$

Segments are considered matched when their Intersection-over-Union (IoU) exceeds a certain threshold, 0.5 in our case. Based on these matched segments, we compute standard document-level metrics, including Precision, Recall, and F1, derived from true positives (TP), false positives (FP), and false negatives (FN).

Stream-Level Metrics: To evaluate the model's utility in real-world annotation workflows, we implement the user-centric metric Minimum Number of Drags and Drops (MnDD) presented in [9]. This metric measures the minimal number of operations needed to transform the predicted segmentation into the ground truth using a graphical interface. It is defined as:

$$\text{MnDD} = N - \sum_{i,j} \max_{i} |G_i \cap P_j| \tag{4}$$

where G_i and P_j are ground truth and predicted segments, respectively.

6.4 Training Protocol

CoSMo models are trained using a cost-sensitive Cross-Entropy loss function, where each class's contribution is weighted inversely to its frequency to mitigate class imbalance, optimizing for multiclass sequence labeling. Training is performed with a learning rate of $1 \cdot 10^{-6}$ and employs Early Stopping to prevent overfitting, with monitoring conducted on an L40s GPU.

7 RESULTS

This section presents the empirical evaluation of our proposed CoSMo models on the PSS task. Subsequently, we

TABLE 2: PERFORMANCE SUMMARY OF COSMO VARIANTS ACROSS MODALITIES AND INTEGRATION STRATEGIES.

	Vision	Textual	Detection	Fusion	F1 Macro ↑	Acc ↑	PQ ↑	MnDD ↓
Vision-Only	✓	✗	✗	✗	97.30	98.46	94.50	0.632
Text-Only	✗	✓	✗	✗	87.92	88.90	70.30	6.322
Vision+Detection	✓	✗	✓	✓	96.12	97.63	92.68	1.069
Multi-Vision	*	✗	✗	✗	95.22	96.72	91.12	1.253
Multi-Vision	*	✗	✗	✓	94.53	97.06	91.80	0.690
Multimodal	✓	✓	✗	✗	**98.10**	**98.65**	**95.08**	**0.437**
Multimodal	✓	✓	✗	✓	96.82	98.21	94.75	0.667

TABLE 3: MULTI-PAGE MODELING RESULTS FOR COSMO BASE WITH DIFFERENT VISUAL BACKBONES.

Backbone	F1-Macro ↑	Acc ↑	PQ ↑	MnDD ↓
SigLIP	**97.30**	**98.46**	**94.50**	**0.632**
SigLIP2	96.27	97.77	93.66	0.7931
CLIP	91.44	91.46	76.72	4.6552
DINOv2	88.62	91.07	75.13	4.8391

detail results from our ablation studies, examining the influence of backbones, fusion strategies, and input modalities. Finally, we present qualitative results and discuss key insights.

7.1 Single-Page Baselines

To understand the capacity of isolated page-level information for inferring semantic roles, we first evaluated various single-page classification models, using F1-Macro as the primary metric due to dataset imbalance. The full results are provided in Table 1.

Experiments on single-page classification revealed several key insights. The **XGBoost classifier**, using handcrafted features, achieved reasonable overall performance but struggled with fine-grained distinctions—particularly for the critical *First Page* class—due to its reliance on layout-only cues.

Zero-shot classification using **CLIP** and **SigLIP** showed strong performance on broad categories like *Cover*, *Advertisement*, and *Story*. However, both models failed to distinguish nuanced classes such as *First Page* and *Text Story*, indicating a lack of domain-specific sensitivity in purely contrastive vision–language embeddings.

Linear probing on frozen visual embeddings led to significant improvements, particularly for challenging categories. Notably, SigLIP surpassed CLIP in this setting, achieving high accuracy without the need for end-to-end fine-tuning—highlighting the quality and structure of its learned representation space.

Zero-shot evaluation with Qwen2.5-VL-32B delivered strong results, especially when using OCR-style prompting. This approach improved performance on context-dependent classes by encouraging the model to "read"the page. However, despite its capabilities, Qwen2.5-VL-32B lagged behind the lighter SigLIP Linear Probe in accuracy and demanded far greater computational resources.

CoSMo was also evaluated in a single-page setting, despite being primarily trained on full page sequences, to assess its robustness without explicit sequential context. In this configuration, the Vision-Only variant exhibited a significant drop in its ability to detect *First Page* transitions, with accuracy for this class falling from 96% (multi-page setting) to 77%. In stark contrast, the Multimodal CoSMo variant demonstrated remarkable resilience; its *First Page* accuracy only decreased from 97% (multi-page) to 92% in the single-page setting. These results offer two critical insights: First, the multimodal approach's benefit, particularly in challenging single-page scenarios, stems from its ability to extract crucial context from textual features. Second, both models significantly benefit from the contextualized Transformer architecture's capacity to capture implicit relations and order, even when presented with isolated pages.

7.2 Multi-Page Results

We now evaluate whether modeling sequential context across pages improves the segmentation of structurally important classes in comics. Table 2 summarizes the performance of various CoSMo architecture variants, progressively incorporating different modalities and design choices.

Detection Feature Fusion. Incorporating detection features into CoSMo led to competitive performance across metrics. However, results remained slightly below the vision-only baseline, suggesting that while these features provide useful layout biases, their benefit may be limited by extraction inconsistencies. In particular, the MAGI detector—originally trained on Manga—may not generalize reliably across diverse comic styles, introducing noise when combined with rich visual embeddings.

Multi-Visual Backbone Modeling. Combining multiple visual backbones obtained worse results than the vision-only model, which uses a single SigLIP as visual backbone. Showing that despite the different pretraining approaches of each backbone, the combination of their features, either through fusion or multi-token does not give any advantage.

Textual-Only CoSMo. Despite the absence of visual input, the textual-only model achieved a respectable F1-Macro score. However, it performed poorly on structural metrics such as PQ and MnDD, largely due to frequent misclassification of *First Page* transitions. These errors stem from the OCR model's difficulty in extracting stylized or decorative title elements critical for story boundary detection. While text provides useful semantic cues, it lacks the spatial and stylistic grounding required for precise segmentation, reinforcing its role as a complementary modality rather than a standalone input.

Visual-Only CoSMo. The vision-only CoSMo model, built on the SigLIP backbone, delivers strong performance

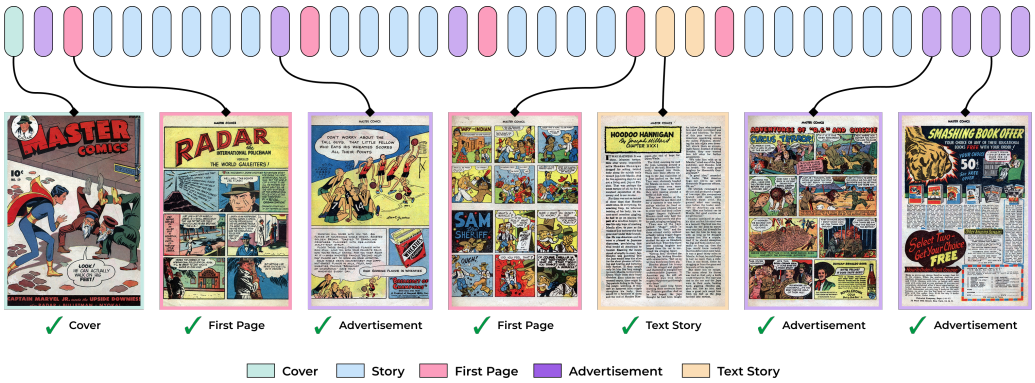

Fig. 6: Example of a comic page sequence showcasing robust and accurate page stream segmentation by both CoSMo Vision and CoSMo Multimodal.

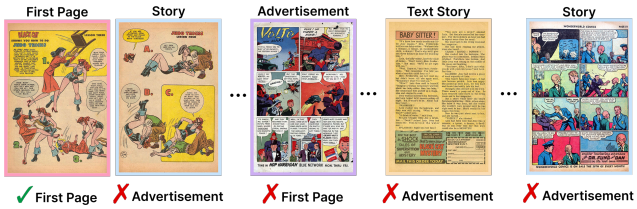

Fig. 7: Examples of pages correctly classified by CoSMo Multimodal and misclassified by CoSMo Vision, highlighting the critical role of textual features in challenging scenarios.

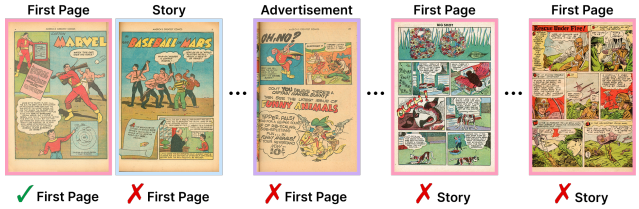

Fig. 8: Pages misclassified by both CoSMo variants, illustrating common error patterns that challenge both vision-only and multimodal models.

across all metrics, with an F1-Macro of 97.30 and a Panoptic Quality (PQ) of 94.50. Its low stream-level error (MnDD of 0.632) confirms that visual features alone are highly effective at modeling both semantic and structural aspects of comic narratives.

Multimodal CoSMo. Incorporating textual features yields the best overall performance, reaching a 98.10 F1-Macro, 95.08 PQ, and the lowest MnDD of 0.437. These gains suggest that text provides complementary cues that help resolve subtle ambiguities, especially in structurally complex or visually unconventional pages. However, improvements over the vision-only model remain relatively modest considering the added complexity of OCR and fusion.

7.3 Ablation Study

To identify the best visual backbone for CoSMo, we conducted an ablation study with various pretrained vision encoders, shown in Table 3.

SigLIP and SigLIP2 performed strongly, with SigLIP achieving the best results in all metrics. Despite its larger size, SigLIP2 slightly underperformed, suggesting that architecture alone doesn't guarantee better segmentation without task-specific alignment.

CLIP and DINOv2, while effective in classification, struggled with segmentation—especially on the critical *First Page* class—highlighting their limitations in capturing the stylistic and structural nuances of comics.

Across models, *First Page* remained the hardest class due to its similarity to *Story* pages. SigLIP handled this best, validating our hypothesis that Transformer-based sequence modeling enhances performance on structurally ambiguous cases. These findings position SigLIP as the best visual backbone for the CoSMo model, and hence it is the one used in all experiments.

7.4 Qualitative Results

To complement the quantitative analysis, this section provides visual examples of CoSMo's performance, highlighting its ability to accurately segment comic page streams, including common challenges, and demonstrate the nuanced contributions of multimodal information.

Successful Segmentation Examples. Figure 6 illustrates a full comic book stream that both CoSMo Vision-Only and CoSMo Multimodal successfully segmented. This example was carefully chosen for its inherent ambiguity, showcasing both standard and challenging page types, including a regular *Cover*, a typical *First-page* , a difficult *Advertisement* resembling a *Story*, a complex *First-page* with multiple small titles, a regular *Text-story*, another challenging *Advertisement* easily mistaken for a *First-page* or *Story*, and a standard *Advertisement*. This success demonstrates their robust capacity to learn complex visual and sequential patterns.

Advantages of the Multimodal Approach. While the CoSMo Vision-Only model performs well overall, it struggles with certain ambiguous cases due to the diverse visual styles of comic pages. Figure 7 illustrates scenarios where adding textual features resolves these ambiguities. The multimodal model correctly distinguishes *First Page* from *Advertisement*, detects atypical *Advertisements* with misleading layouts, and accurately classifies mixed-content pages like *Text Stories*. These examples highlight how textual cues enhance fine-grained semantic understanding, especially in cases where visual information alone is insufficient.

Analysis of Persistent Challenges. Despite strong ove-

rall performance, both CoSMo variants struggle with certain edge cases, as shown in Figure 8. Common failure modes include misclassifying non-initial *Story* pages as *First Page* due to local title cues without considering preceding context. Similarly, pages combining visual storytelling and ad-like content often confuse the model. Legitimate *First Page* examples with atypical layouts or weak title cues are also frequently mislabeled as *Story*. These errors underscore the difficulty of modeling long-range dependencies and handling visual and stylistic variability—key areas for future improvement.

8 Conclusions

We introduced CoSMo, a novel multimodal Transformer for Page Stream Segmentation (PSS) in comic books. This work formalized this underexplored task, and curated a high-quality annotated dataset. Our extensive evaluation, conducted using a comprehensive suite of metrics spanning page, document, and stream levels, demonstrated CoSMo's superior performance. By establishing strong baselines and achieving a new state-of-the-art for comic PSS across these various metrics, we underscored the efficacy of our end-to-end Transformer-based model. The vision-only CoSMo, using a SigLIP backbone, proved remarkably strong and efficient, highlighting the importance of visual semantics as the dominant and most influential features. However, to achieve the absolute maximum state-of-the-art performance, especially in the most challenging and ambiguous cases, the complementary signal from textual features in the full multimodal CoSMo proved essential, demonstrating our exploration of diverse multimodal representations. This critical synergy, however, comes with a trade-off of increased computational overhead and a dependency on high-quality OCR. Crucially, CoSMo consistently outperformed significantly larger, general-purpose vision-language models, affirming the practical value of specialized architectures for structured document understanding. This work marks a significant step towards automated comic book analysis, serving as a vital component in the Layer of Comics Understanding (LoCU) framework.

8.1 Limitations

Despite CoSMo's robust performance for standard PSS, certain inherent limitations exist. Our current multimodal approach relies on a large, general-purpose OCR system, which incurs significant computational overhead and may not be optimally tailored for highly stylized comic book typography. Furthermore, while CoSMo effectively segments page types, its current scope does not extend to "advanced PSS,"which involves extracting rich metadata like authors or story titles. This latter limitation stems primarily from the absence of sufficiently large and specialized datasets required for training models capable of such granular extraction.

8.2 Future Work

Building on CoSMo's foundation, future work will focus on two key areas. First, we aim to develop a more cost-effective and task-specific OCR pipeline, potentially integrating contextualized features to enhance multimodal efficiency. Second, leveraging CoSMo to generate large-scale pseudo-annotations for a new dataset, that combined with manual annotations, will enable an "advanced PSS"model capable of extracting valuable metadata, thus scaling beyond current data limitations and unlocking richer semantic understanding. Further avenues include investigating more advanced multimodal fusion mechanisms, exploring novel strategies for challenging classes like *First Page* (e.g., contrastive learning or dedicated boundary detection), expanding the dataset to cover a wider range of comic eras and styles for enhanced generalization, and integrating CoSMo into interactive annotation tools to streamline digitization workflows.

Acknowledgements

I am deeply grateful to Dr. Dimosthenis Karatzas for the invaluable opportunity and confidence he placed in me, as well as for his mentorship. My sincere thanks also go to Emanuele, whose guidance ensured a truly inspiring introduction to the field of research. I extend my appreciation to my family for their unwavering support, my friends and CVC colleagues for creating such an enriching atmosphere, and to Liah for her constant encouragement and joy.

Referències

[1] Shuai Bai et al. *Qwen2.5-VL Technical Report*. 2025. arXiv: 2502.13923 [cs.CV]. URL: https://arxiv.org/abs/2502.13923.

[2] Mehmet Arif Demirtaş et al. "Semantic Parsing of Interpage Relations". A: *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE. 2022, pàg. 1579 - 1585.

[3] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". A: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019, pàg. 4171 - 4186.

[4] Albert Gordo et al. "Document Classification and Page Stream Segmentation for Digital Mailroom Applications". A: *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*. Ag. de 2013, pàg. 621 - 625. DOI: 10.1109/ICDAR.2013.128.

[5] Hunter Heidenreich et al. *Large Language Models for Page Stream Segmentation*. 2024. arXiv: 2408.11981 [cs.CL]. URL: https://arxiv.org/abs/2408.11981.

[6] Ruben van Heusden, Jaap Kamps i Maarten Marx. "OpenPSS: An Open Page Stream Segmentation Benchmark". A: *Linking Theory and Practice of Digital Libraries: 28th International Conference on Theory and Practice of Digital Libraries, TPDL 2024, Ljubljana, Slovenia, September 24–27, 2024, Proceedings, Part I*. Ljubljana, Slovenia: Springer-Verlag, 2024, pàg. 413 - 429. ISBN: 978-3-031-72436-7. DOI: 10.1007/978-3-031-72437-

4_24. URL: https://doi.org/10.1007/978-3-031-72437-4_24.

[7] Mohit Iyyer et al. "The Amazing Mysteries of the Gutter: Drawing Inferences Between Panels in Comic Book Narratives". A: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pàg. 6478 - 6487. URL: https://api.semanticscholar.org/CorpusID:215826810.

[8] Alexander Kirillov et al. "Panoptic segmentation". A: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pàg. 9404 - 9413.

[9] Thisanaporn Mungmeeprued et al. "Tab this Folder of Documents: Page Stream Segmentation of Business Documents". A: set. de 2022. DOI: 10.1145/3558100.3563852.

[10] Maxime Oquab et al. "DINOv2: Learning Robust Visual Features without Supervision". A: *Transactions on Machine Learning Research Journal* (2024), pàg. 1 - 31.

[11] Alec Radford et al. "Learning transferable visual models from natural language supervision". A: *International conference on machine learning*. PmLR. 2021, pàg. 8748 - 8763.

[12] Dillon Reis et al. *Real-Time Flying Object Detection with YOLOv8*. 2024. arXiv: 2305.09972 [cs.CV]. URL: https://arxiv.org/abs/2305.09972.

[13] Shaoqing Ren et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". A: *Advances in neural information processing systems* 28 (2015).

[14] Ragav Sachdeva, Gyungin Shin i Andrew Zisserman. "Tails Tell Tales: Chapter-wide Manga Transcriptions with Character Names". A: *Proceedings of the Asian Conference on Computer Vision (ACCV)*. Des. de 2024, pàg. 2053 - 2069.

[15] Ragav Sachdeva i Andrew Zisserman. *From Panels to Prose: Generating Literary Narratives from Comics*. 2025. arXiv: 2503.23344 [cs.CV]. URL: https://arxiv.org/abs/2503.23344.

[16] Ragav Sachdeva i Andrew Zisserman. "The manga whisperer: Automatically generating transcriptions for comics". A: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pàg. 12967 - 12976.

[17] Michael Tschannen et al. *SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features*. 2025. arXiv: 2502.14786 [cs.CV]. URL: https://arxiv.org/abs/2502.14786.

[18] Emanuele Vivoli, Marco Bertini i Dimosthenis Karatzas. "Comix: A comprehensive benchmark for multi-task comic understanding". A: *Advances in Neural Information Processing Systems* 37 (2024), pàg. 140828 - 140846.

[19] Emanuele Vivoli et al. "Comics Datasets Framework: Mix of Comics datasets for detection benchmarking". A: *International Conference on Document Analysis and Recognition*. Springer. 2024, pàg. 154 - 167.

[20] Emanuele Vivoli et al. "Multimodal transformer for comics text-cloze". A: *International Conference on Document Analysis and Recognition*. Springer. 2024, pàg. 128 - 145.

[21] Emanuele Vivoli et al. *One missing piece in Vision and Language: A Survey on Comics Understanding*. 2025. arXiv: 2409.09502 [cs.CV]. URL: https://arxiv.org/abs/2409.09502.

[22] Gregor Wiedemann i Gerhard Heyer. "Page Stream Segmentation with Convolutional Neural Nets Combining Textual and Visual Features". A: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

[23] Xiaohua Zhai et al. "Sigmoid loss for language image pre-training". A: *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, pàg. 11975 - 11986.

[24] Yanzhao Zhang et al. *Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models*. 2025. arXiv: 2506.05176 [cs.CL]. URL: https://arxiv.org/abs/2506.05176.