Drawing Insights: Sequential Representation Learning in Comics

Sam Titarsolej[1]
s.titarsolej@gmail.com
Neil Cohn[2]
n.cohn@tilburguniversity.edu
Nanne van Noord[1]
n.j.e.vannoord@uva.nl

[1] MultiX Lab
University of Amsterdam
[2] Visual Language Lab
Tilburg University

Abstract

Comics present images in a sequence, where the spatially presented sequence is key to the narrative storytelling. To understand a comic, a comprehender must learn to encode this sequential nature. For this we present a novel self-supervised sequential representation learning method designed for comics. Our approach capitalises on the sequential structure of comics to incorporate contextual information. We conduct experiments on the TINTIN Corpus of 1,000+ comics from 144 countries, and show that our method outperforms baseline methods on both classification and retrieval tasks. These results affirm the effectiveness of sequential representation learning for comics, and may aid in uncovering new cultural insights within comics.

1 Introduction

Comics provide insight into cultural narratives, visual storytelling, and human expressiveness [1, 4, 26, 32]. They serve not only as entertainment but also as a mirror reflecting cultural and societal constructs. To analyse these constructs within comics at scale we require robust representations that are sensitive to the visual patterns within comics. Yet, dissecting comics into stylistic and semantic building bricks is a difficult endeavour [3, 8, 9]. Comics are inherently complex, characterised by a unique blend of visual artistry and textual narrative - also described as visual language [9]. This diversity in artistic styles poses a significant challenge for the quantification and annotation of comics. To overcome these issues and to learn comic representations we propose to leverage the sequential nature of comics panels as an inductive bias, and learn comic representations without supervision.

Through learned representations, we aim to work towards a quantitative understanding of comics. Such understanding could offer insights into difference in visual storytelling across different cultures. This analysis extends beyond the identification of basic stylistic elements to encompass a wide range of patterns depicted in comics. By systematically analysing these representations, we can uncover underlying patterns in how comics are constructed and their narrative structure [9].

© 2024. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

In this work we introduce a novel method for explicitly modelling the sequential nature of comics. To achieve this we introduce a training method for sequential image modelling taking inspiration from masked language modelling. We show that our method outperforms prior self-supervised methods such as DINO [6] and OpenCLIP [7] for comic representation learning on various evaluation tasks. Finally, we evaluate the individual components of our method through a comprehensive ablation study. All code for this research is publicly available at https://github.com/samtitar/ASTERX.

2 Related Work

Quantitative approaches to cultural analysis of comics, have enabled an increased understanding of visual storytelling across different cultures in recent years [26]. [8] and [11] introduced categorisations of scene framing techniques within panels. Their research revealed distinct framing practices among comics across cultures. This is extended in [12] to the layout of panel sequences, by introducing a systematic framework for evaluating page layouts, and facilitating a comparison of comics from six different cultural origins. Their findings provide insight into differences in layouts, further emphasising the cultural specificity in comics. A framework for analysing continuity of panel sequences is further introduced in [3], which enabled a detailed comparison of narrative structures across comics from different cultural origins, revealing variations in continuity between these origins.

Integration across several domains of structures have also revealed consistencies that cut across cultural dimensions, suggesting distinctive "visual languages" that correspond to systems of graphic representations and their storytelling [10]. However, though visual languages are implicated by patterns of storytelling, their corresponding visual "styles" have not been as thoroughly analysed. In addition, a challenge for such work is the reliance on manual annotation of comics.

In parallel, over a decade of computational approaches for comics have emerged using supervised machine learning [2, 31]. These have used a range of approaches from basic pattern recognition in comic art to complex narrative structure analysis. A substantial portion of existing work in computational comic analysis can be attributed to advancements in computer vision (CV) techniques, particularly in the automated extraction of key comic elements such as panels, speech bubbles, and characters. Methods for panel extraction have been addressed by [21] [28] [20] among others. Their methods leverage various CV techniques to identify and segment comic book panels. These approaches all leverage annotated comic panels to train object detection models for panel extraction. Automated text block extraction methods have been introduced in [30] and [16].

The exploration of computationally-aided analysis of comics has focused on supervised machine learning, as detailed by [31] and [2]. While these developments contributed to the field, the reliance on supervised learning again highlights a dependency on extensively annotated datasets. This requirement often limits the scalability of analysis and the adaptability to the vast stylistic diversity inherent in comics. To overcome these issues we propose to apply leverage self-supervised learning on the sequential structure of comics.

Self-supervised representation learning offers a rich array of methods that extract meaningful representations from data without the need for labeled datasets. Several strategies that have advanced the field are DINO (Self-Distillation with No Labels) [6], DINOv2 [27], and CLIP [29]. The two former methods employ self-distillation within a vision transformer (ViT) [34] in a student-teacher setup to construct rich image representations. Different views

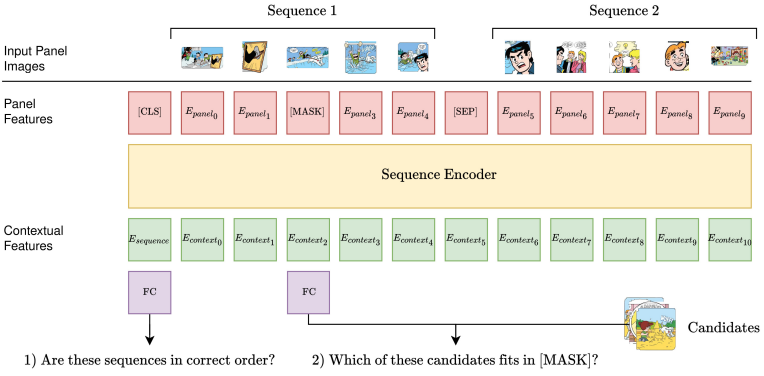


Figure 1: Overview of ASTERX. Individual panels are encoded with a backbone, after which the sequence of panels is encoded by our sequence encoder. This sequence encoder is optimised for both panel retrieval and panel order classification. In our experiments, the sequence encoder is a transformer encoder [34], where each input token is a panel feature vector encoded by an image encoder backbone.

of the same image are passed through the student and teacher models to optimise feature-level similarity. On the other hand, CLIP utilises image-caption pairs to jointly learn representations of both image and caption by minimises the distance between representations extracted from them. A challenge for comics is that, while panels may contain text, these do not take on the role of captions and would not be suitable for CLIP-style training.

A corner stone for self-supervised learning and a key enabler of existing methods, is data scale [27], where the performance of these methods greatly benefits from large-scale datasets. While the COMICS dataset [23] consists of 1.2M comic panels, this is still small-scale when compared to the 142M dataset for DINO [27] or the 400M for CLIP [29]. The limited availability of data for comics may make it challenging to apply these existing methods on them, so instead we propose a comic specific self-supervised learning method which leverages the sequential nature of the panels as inductive bias.

3 ASTERX

Our proposed method, **A S**elf-supervised **T**ransformer **E**ncoder for comic panel **R**epresentation e**X**traction (**ASTERX**, named after the comic character *Asterix*) is designed to learn from the sequential nature of comics. Our technique is inspired by masked language modelling [14], as we leverage narrative continuity embedded within adjacent panels to learn comic panel representations. As discussed in [6, 12, 22], the continuity of such aspects in sequences of panels in comics are of great significance to cultural analysis in comics.

ASTERX models sequences of comic panels with a transformer [34] encoder (Figure 1). To enable analysis at both panel-level and sequence-level a special [CLS] token is prepended to each sequence, with one input tokens per individual panel. The [CLS] token captures the sequence-level information, while panel-level representations are encoded by the matching output tokens for each input panel. To optimise these representations we train ASTERX with two tasks inspired by masked language modelling.

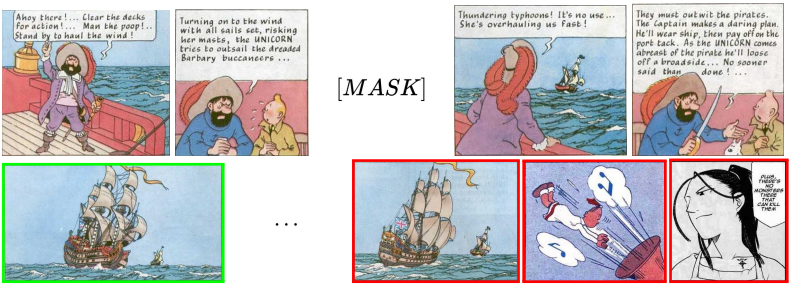


Figure 2: Sequence of panels from *Tintin: The Secret of the Unicorn* (first row) along with two candidate samples from the same comic and two from *Suske en Wiske* and *Fullmetal Alchemist* respectively (second row). The correct candidate in green, and incorrect in red.

3.1 Pretext Tasks

During training, the encoder is optimised by means of two tasks: panel retrieval and order classification. To facilitate these tasks the encoder receives two panel sequences - always sampled chronological in order - of the same length as input, separated by a [SEP] token. For the panel retrieval task a single input token (panel) is masked using the special [MASK] token in the first panel sequence, and the encoder is optimised to produce a representation that matches the masked panel. The second input sequence is used in a order classification task. The order of the two input sequences is shuffled at random to facilitate learning of longer panel distance relations. A linear classification layer takes the output embedding at the position of the [CLS] token is optimised to classify whether the sequences are in correct order.

Our approach follows the ideas behind masked language modelling introduced for BERT [14]. However, a key difference between natural language and visual language is the open set nature of visual language. Where natural language can be modelled through a discrete set of for words or tokens (i.e., the vocabulary), visual language is more continuous in nature and thus cannot be directly modelled in the same closed set fashion as masked language modelling. Even though for example framing of characters in comics [8, 21] often follow patterns across comics that can be discretised, the underlying data format (i.e. images) make pixel-wise reconstruction - in comparison to token-wise reconstruction in MLM - an infeasible task as discussed in [19, 35].

We address the challenges that come from the open set nature of visual language by casting it as a retrieval task. For the panel retrieval task we construct a collection of "candidate" panels. These panels are selected from the training set based on a predefined sampling method. With these candidates, the sequence encoder is not only optimised minimise the distance between the output embedding at the [MASK] position and the masked input token; it is also honed to maximise the distance from all incorrect candidate panel embeddings. This configuration allows us to optimise the sequence encoder by learning to select the correct panel. The probability assigned by the sequence encoder for an individual candidate is formulated as:

$$\hat{y}_i = \frac{\exp f(x, c_i)}{\sum_{c \in C} \exp f(x, c)}. \tag{1}$$

Here, $f(x,c)$ denotes a distance metric used to gauge the similarity between two embeddings. x represents the output embedding at the [CLS] token position, and C encompasses the set of candidate tokens including the authentic masked token. When expanded in the batch dimension, each x is paired with a unique C constructed according to the candidate sampling strategy. In essence, the optimisation objective for the panel retrieval task follows the principals of Noise-Contrastive Estimation loss [25]. Employing this formulation, the sequence encoder is optimised using the cross entropy loss, where the ground truth label aligns with the index of the masked input token within the candidate set.

3.2 Candidate Sampling

The candidate sampling strategy intricately shapes the embedding space of the panel sequence encoder. Consequently, we propose four sampling strategies: **1) Random sampling**: Candidates are randomly selected from the training dataset without any specific criteria. Each panel in the entire dataset has an equal probability of being selected as candidate **2) Pixel-intensity based sampling:** Candidate panels are chosen within a range of similar pixel-value intensities as the masked panel, based on histogram bins. **3) Panel shape based sampling:** Candidates are selected within a range of comparable panel shapes, defined by the height-to-width ratio, akin to the masked panel. **4) Same comic sampling:** Candidates are exclusively sourced from within the same comic as the input sequence. Crucially, all properties used in these sampling strategies do not require any annotation.

For each strategy two variants exist: a pure sampling strategy and a mixed strategy. For the pure variants, panels are only sampled according to criteria described, while in the mixed variant half of the panels are sampled according to the criteria and the other half are sampled completely randomly. In each of these strategies the true target panel is always included in the set of candidate panels.

4 Experiments

Our experiments involve comparison of our method with various methods on two evaluation tasks: training a linear classifier using learned representations for a specific classification task and panel panel retrieval. We compare our method to the following existing works: 1) supervised ResNet-50 [18] trained to perform only style classification, 2) OpenCLIP [7] trained on the DataComp-1B [17] dataset, 3) DINO [6] and DINOv2 [27] trained on ImageNet [13] and 4) DINO trained on the COMICS [23] first, and then on the TINTIN corpus [5]. The OpenCLIP, DINO and DINOv2 methods are also evaluated in a "Pooled" variant, where the features of neighbouring panels surrounding the target panel are mean-pooled to obtain a contextualised representation.

4.1 Data

For our experiments we use two datasets, the COMICS dataset [23] and the TINTIN corpus [5]. The COMICS dataset consists of 1.2 million panel images and features American comics from the golden age (i.e., 1940s and 1950s). We use the COMICS dataset as a pre-training starting point due to the large quantity of available data within the domain of comics, despite the its larger uniformity in comic style.

The TINTIN Corpus features more diversity, with over 1,000 comics from over 144 countries/territories, and spanning more than 77,000 panels. The TINTIN Corpus is also accompanied by various types of comic-level annotations such as style, format and genre. For the style attribute, the majority class covers 24.5% of a total of 13 classes with classes such as "Manga" "Superhero" and "Cartoon". Format contains classes such as "Comic book", "Webcomic", and "Graphic Novel" with a total of 8 classes and the majority class covering 31.2%. Lastly, the genre attribute contains a total of 91 classes such as "Supernatural", "Action", and "Political commentary", with the majority class spanning 17.4%. All quantitative evaluation of our approach is performed on the TINTIN Corpus, as it encapsulates more varied data and provides annotations as ground truth information. The TINTIN Corpus is split into an 80% training split, a 10% validation split and a 10% test split for evaluation.

4.2 Experimental Setup

We use DINO-trained [6] ViT [15] features for the individual panel representations that serve as input tokens for the sequence encoder. To explore the influence of the backbone we experiment with variations of training the ViT feature extractor on different datasets, and with different ViT settings. For the main experiments the backbone - pre-trained on ImageNet - is first fine-tuned on the COMICS dataset, then fine-tuned on the TINTIN corpus.

All models are optimised using the AdamW optimiser [24], with batch size 256 and learning rate 0.0001 for 50 epochs. Unless specified otherwise, we sample 64 candidate panels during training. The architecture of the panel sequence encoder consists of four attention heads and four transformer encoder layers. Each layer has a dropout probability of 0.4. The input embedding size is 384, and the output embedding size is 768. We found a sequence length of 5 (two panels surrounding the masked panel on both sides) to be optimal during hyperparameter optimisation.

Method	ViT	Style	Format	Genre
Fully Supervised				
ResNet-50 [18]	-	85.6	53.2	47.7
ResNet-50 Pooled	-	**87.3**	**56.1**	**49.8**
Linear Fine-tuned				
OpenCLIP [7]	B/16	35.6	38.1	25.4
DINO [6]	S/16	34.8	37.7	23.5
DINOv2 [27]	S/14	33.4	36.6	21.7
OpenCLIP Pooled	B/16	**36.2**	**38.4**	**26.0**
DINO Pooled	S/16	35.4	38.2	23.9
DINOv2 Pooled	S/14	34.0	36.7	22.3
Fully Fine-tuned				
DINO	B/16	54.8	51.4	42.6
DINO Pooled	B/16	56.2	54.8	43.6
ASTERX (ours)	B/16	**65.6**	**63.2**	**51.2**

Table 1: Linear classification results on various attributes of the TINTIN Corpus [5]. The ResNet-50 [18] model is trained for style classification, the models in the Not Fine-tuned section are trained on DataComp-1B [17] and ImageNet [13]. The methods in the Fully Fine-tuned section are first fine-tuned on the COMICS dataset [23] and then Fine-tuned on the TINTIN Corpus.

Method	ViT	R@1	R@5	R@10
Fully Supervised				
ResNet-50 [18] Pooled	-	**0.1**	**0.1**	**0.2**
Not Fine-tuned				
OpenCLIP Pooled [7]	B/16	**0.5**	**4.5**	**15.6**
DINO [5] Pooled	S/16	0.1	0.5	8.1
DINOv2 [21] Pooled	S/14	0.1	0.4	7.6
Fully Fine-tuned				
DINO Pooled	B/16	1.5	19.7	37.1
ASTERX (ours)	B/16	**15.7**	**72.3**	**89.6**

Table 2: Panel retrieval performance on the TINTIN Corpus [5]. The ResNet-50 [18] model is trained for style classification, the models in the Not Fine-tuned section are trained on DataComp-1B [17] and ImageNet [13]. The methods in the Fully Fine-tuned section are first fine-tuned on the COMICS dataset [23] and then Fine-tuned on the TINTIN Corpus.

4.3 Linear Fine-tuning

We follow a common approach for quantitative evaluation of self-supervised training methods. This approach involves training a linear classifier on the feature space derived from the self-supervised model for a specific classification task. In our configuration, we leverage diverse comic-level annotations extracted from the TINTIN Corpus. Specifically, we employ style, format, and genre annotations to train a linear classifier. Subsequently, this classifier's performance is evaluated based on classification accuracy, which provides insight into how well the feature space captures specific information.

For this analysis we compare models in three different settings: (1) Fully Supervised, here the ResNet-50 models are pre-trained for Style and then fine-tuned on Format and Genre. This model is only pre-trained on Style classification, as the expected results when pre-training on the other two aspects would be similar to this setting. (2) Linear Fine-tuned, which utilises pre-trained backbone on image datasets, and a linear classifier that is fine-tuned for each classification task. OpenCLIP is pre-trained on DataComp-1B, and DINO and DINOv2 are pre-trained on ImageNet. In (3) the Fully Fine-tuned section all methods are first pre-trained on the COMICS dataset, and then finonesidee-tuned on the TINTIN Corpus.

Table 1 shows the classification performance of a linear classifier trained to classify three aspects of a comic in the TINTIN Corpus using the features of various methods. The ResNet-50 trained to classify style outperforms all other methods for style classification, yet generalizes poorly to format and genre. Our method outperforms all unsupervised methods, as well as the ResNet-50 on format and genre classification, indicating that supervised pre-training does not result in a model that generalises well.

4.4 Panel Retrieval

To further evaluate our method, we pose a retrieval task that aims to retrieve a missing panel within a sequence. To determine how well the model handles large diversity in style and content we use the entire test set, consisting of approximately 7.700 panel images, as the pool of candidate panels. We compute Recall@K across different values of k to gauge the retrieval performance, and thus the ability of the model to encode contextual information

Method	Linear Classifier ↑			Retrieval ↑		
	Style	Format	Genre	R@1	R@5	R@10
Random	56.8	55.0	43.7	2.3	45.5	60.0
Pure Sampling						
Intensity	57.9	56.8	45.4	4.1	52.7	67.9
Ratio	57.8	57.0	45.7	3.6	51.2	67.5
Comic	**60.9**	**59.5**	**47.0**	**4.5**	**51.6**	**68.7**
Mixed Sampling						
Intensity + Random	59.0	58.2	45.9	11.3	67.2	82.6
Ratio + Random	59.3	58.6	46.4	11.5	67.1	82.9
Comic + Random	**63.8**	**62.5**	**47.6**	**12.3**	**68.8**	**84.2**

Table 3: Classification and retrieval performance of different candidate sampling strategies. All mixed sampling strategies outperform pure sampling strategies, and the mixed comic sampling strategy performs best overall.

within panel sequences.

The panel retrieval performance of the compared methods are shown in Table 2. Across the board we see that this is a highly challenging task. The ResNet-50 model that has been trained in a supervised manner for style classification performs poorly in this retrieval setting, failing to generalise to retrieval. Similarly, the pre-trained OpenCLIP and DINO backbones do not perform well, with OpenCLIP outperforming DINO and DINOv2, which may presumably be due to the the presence of comic(-like) images in the larger pre-training set. Fine-tuning DINO on comic data boost the performance somewhat, but presumably more data is required to obtain good performance. Comparatively, our method performs notably better, thereby showcasing increased contextual understanding of panels.

Comparing the retrieval performance in Table 2 to the classification performance in Table 1, there is an obvious gap in the performance of the ResNet-50 where the strong classification performance does not translate to retrieval performance. This gap showcases the reality of the generalisability of features learned with supervision, as good classification performance does not guarantee good retrieval performance. Comparing the retrieval performance with the classification performance for the unsupervised methods, we observe much more balance in performance between these varied tasks, indicating that more general representations are learned.

4.5 Ablations

We perform three ablations to increase understanding of our method. First, Table 3 demonstrates both the linear classifier performance and the retrieval performance of our method trained using various candidate sampling strategies. All mixed sampling strategies outperform the pure sampling counterpart, indicating that a diverse mix of sampling candidates is beneficial in all evaluation aspects. Furthermore, in both mixed and pure sampling categories sampling candidates from the same comic outperforms the other sampling strategies. The fully random sam-

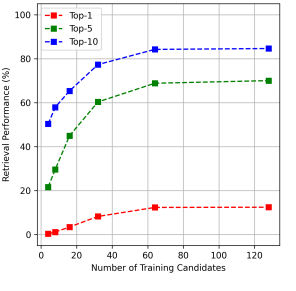


Figure 3: Retrieval performance of our method against the number of candidates sampled during training.

Backbone	ASTERX	R@1	R@5	R@10
COMICS [23]	COMICS	0.3	13.8	32.4
COMICS	TINTIN [6]	0.6	29.3	54.7
COMICS → TINTIN	TINTIN	**12.3**	**68.8**	**84.2**
TINTIN	TINTIN	10.8	65.3	79.7

Table 4: Comparison of the retrieval performance for the backbone and ASTERX trained on COMICS [23] and/or TINTIN [6], and subsequently evaluated on COMICS or TINTIN to evaluate cross-dataset performance. We observe that cross-dataset pre-training works best.

pling strategy is outperformed by all other sampling strategies.

Secondly, Table 4 provides insight into the generalisability over the data domain. Three configurations of fine-tuning both the feature extraction backbone and the sequence encoder on different datasets are shown. All retrieval evaluation results are on the TINTIN Corpus validation set. Unsurprisingly, fine-tuning both models on the target dataset results in the best performance. However, only training both models on the COMICS dataset already approaches similar performance of DINO trained on the target dataset (as in Table 2).

Finally, Figure 3 displays the evaluation retrieval performance of our method based on the number of candidates used during training. The graph clearly depicts a stark increase in performance between 0 and 50 candidates, and flattens beyond 50 candidates, indicating an optimal value around 64 candidates.

4.6 Qualitative Results

In an effort to form a qualitative understanding of the learned representations of our method in comparison to other methods, two retrieval results are presented in Figure 4. In both cases, the original comic page with the masked panel of interest are shown, as well as retrieval results from ASTERX (ours), DINO and the supervised ResNet-50. In both cases only ASTERX retrieved the correct target panel. In the left case, DINO retrieved a panel from the same comic, in which one of the main characters of this comic is portrayed. The supervised ResNet-50 retrieves a visually similar panel, but is unable to retrieve a panel from the correct

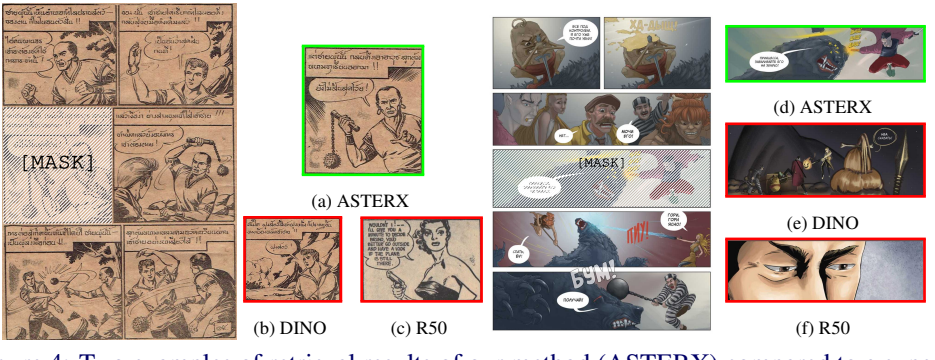
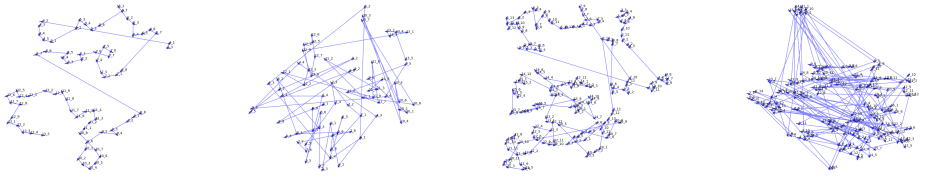


Figure 4: Two examples of retrieval results of our method (ASTERX) compared to a supervised ResNet-50 [13] and the unsupervised DINO [6] method. In both cases our method is able to retrieve the correct panel while both other methods are not.



(a) *The Bluecoats* (b) *Suske en Wiske*

Figure 5: t-SNE [33] projections of ASTERX features in comparison to t-SNE projections of DINO [6] features for two comics.

comic. On the right all results are close in visual style, but DINO returns a panel from a different edition from the same comic series, whereas ResNet-50 selects a panel from a different series altogether. Interestingly, all three results are approximately the right shape.

Figure 5 displays the difference between ASTERX and DINO in terms of feature space organisation. Two examples are shown in which all panels of two comics are encoded and then projected using into 2D space using T-SNE [33]. Each example displays how ASTERX clearly encapsulates the sequential nature of panels, while DINO organises panels without regard for panel order.

5 Conclusion

In this work we introduced ASTERX, a novel method for learning representations for panels in comics, while capturing the sequential nature of comics. We showed through several experiments how this learning strategy specifically designed for comics outperforms baseline methods that do not take into account their sequential nature. Crucially, we demonstrate the feasibility of self-supervised fine-tuning on domains which are not as data rich by leveraging domain-specific aspects. This allows us to present a strong feature basis that enables diverse cultural analysis of comics.

6 Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 850975).

References

[1] *Comics and Power: Representing and Questioning Culture, Subjects, and Communities*. Cambridge Scholars, 2015.

[2] Olivier Augereau, Motoi Iwata, and Koichi Kise. A survey of comics research in computer science. *Journal of imaging*, 4(7):87, 2018.

[3] Irmak Hacımusaoğlu Bien Klomberg and Neil Cohn. Running through the who, where, and when: A cross-cultural analysis of situational changes in comics. *Discourse Processes*, 59(9):669–684, 2022.

[4] C. Brienza and P. Johnston. *Cultures of Comics Work*. Palgrave Studies in Comics and Graphic Novels. Palgrave Macmillan US, 2016.

[5] Bruno Cardoso and Neil Cohn. The multimodal annotation software tool (MAST). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6822–6828. European Language Resources Association, June 2022.

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021.

[7] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning, 2022.

[8] Neil Cohn. A different kind of cultural frame: An analysis of panels in american comics and japanese manga. *Image and Narrative*, 12(1):120–134, 2011. ISSN 1780-678x.

[9] Neil Cohn. Visual narrative structure. *Cognitive science*, 37 3:413–52, 2013.

[10] Neil Cohn. *The patterns of comics*. Bloomsbury Academic, December 2023.

[11] Neil Cohn, Amaro Taylor-Weiner, and Suzanne Grossman. Framing attention in japanese and american comics: Cross-cultural differences in attentional structure. *Frontiers in Psychology*, 3, 2012. ISSN 1664-1078.

[12] Neil Cohn, Jessika Axnér, Michaela Diercks, Rebecca Yeh, and Kaitlin Pederson. The cultural pages of comics: cross-cultural variation in page layouts. *Journal of Graphic Novels and Comics*, 10(1):67–86, 2019.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2018.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

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

[17] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets, 2023.

[18] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CoRR*, abs/2111.06377, 2021.

[20] Zheqi He, Yafeng Zhou, Yongtao Wang, Siwei Wang, Xiaoqing Lu, Zhi Tang, and Ling Cai. An end-to-end quadrilateral regression network for comic panel extraction. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, page 887–895. Association for Computing Machinery, 2018.

[21] Anh Khoi Ngo Ho, Jean-Christophe Burie, and Jean-Marc Ogier. Panel and speech balloon extraction from comic books. In *2012 10th IAPR International Workshop on Document Analysis Systems*, pages 424–428, 2012.

[22] Bien Klomberg Irmak Hacımusaoğlu and Neil Cohn. Navigating meaning in the spatial layouts of comics: A cross-cultural corpus analysis. *Visual Cognition*, 31(2):126–137, 2023.

[23] Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan L. Boyd-Graber, Hal Daumé III, and Larry S. Davis. The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. *CoRR*, abs/1611.05118, 2016.

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[25] Zhuang Ma and Michael Collins. Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3698–3707, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1405. URL https://aclanthology.org/D18-1405.

[26] Irmak Hacımusaoğlu Neil Cohn and Bien Klomberg. The framing of subjectivity: Point-of-view in a cross-cultural analysis of comics. *Journal of Graphic Novels and Comics*, 14(3):336–350, 2023.

[27] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang,

Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024.

[28] Xufang Pang, Ying Cao, Rynson W.H. Lau, and Antoni B. Chan. A robust panel extraction method for manga. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, page 1125–1128. Association for Computing Machinery, 2014.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021.

[30] Christophe Rigaud, Jean-Christophe Burie, and Jean-Marc Ogier. Text-independent speech balloon segmentation for comics and manga. In *Graphic Recognition. Current Trends and Challenges*, pages 133–147. Springer International Publishing, 2017.

[31] Rishabh Sharma and Vinay Kukreja. Image segmentation, classification and recognition methods for comics: A decade systematic literature review. *Engineering Applications of Artificial Intelligence*, 131:107715, 2024. ISSN 0952-1976.

[32] A. Silbermann. *Comics and Visual Culture: Research Studies from ten Countries*. De Gruyter, 2010.

[33] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[35] Haoqing Wang, Yehui Tang, Yunhe Wang, Jianyuan Guo, Zhi-Hong Deng, and Kai Han. Masked image modeling with local multi-scale reconstruction, 2023.