MSc Artificial Intelligence
Master Thesis

Exploring the Impact of Textual Integration in Comic Book Representations

by
Kacper Bartosik
12624268

June 27, 2025

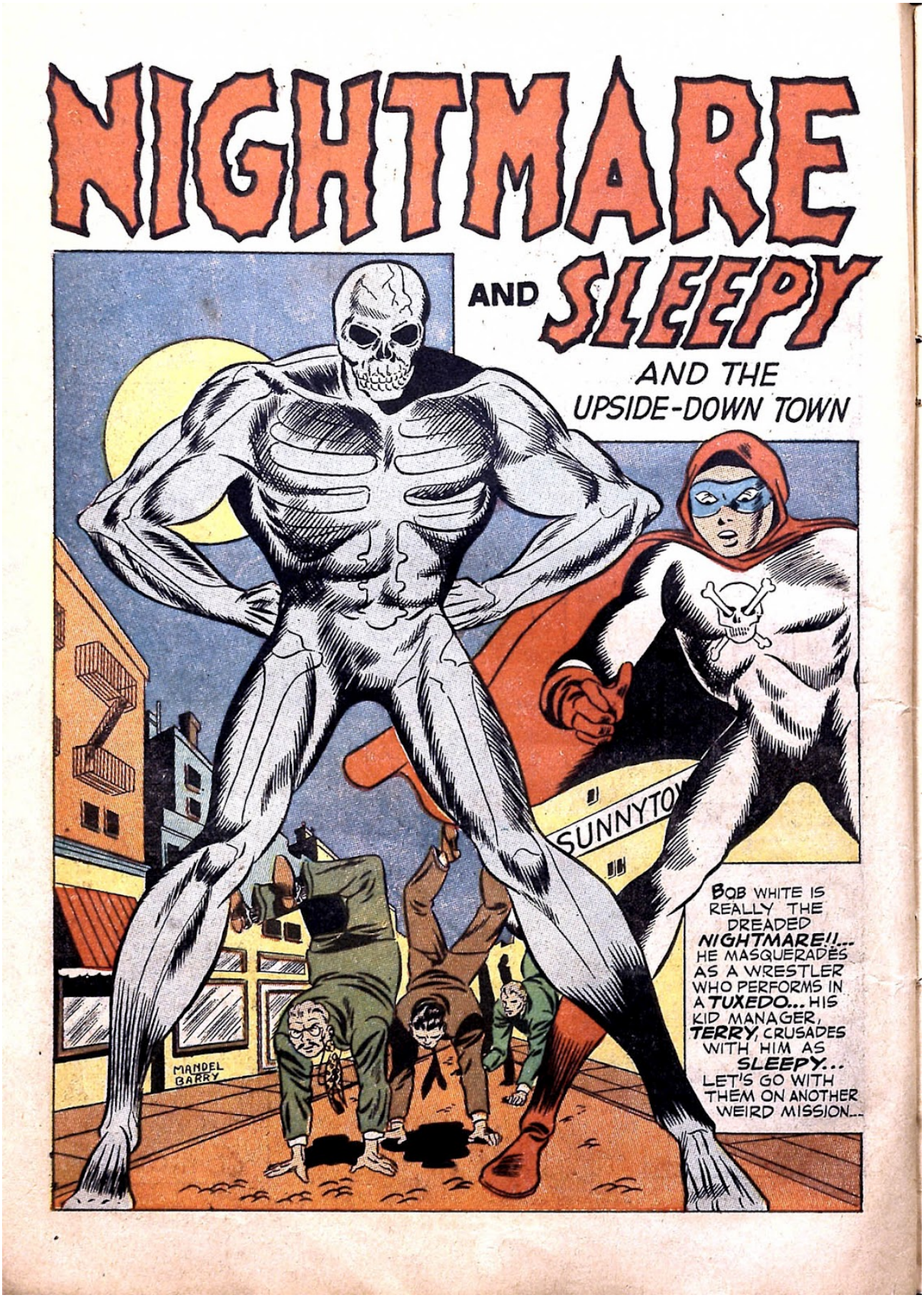
36 EC
January 2025 - June 2025

Supervisor:
Dr. Nanne van Noord

Examiner:
Dr. Yen-Chia Hsu

Second reader:
Dr. Nanne van Noord

Universiteit van Amsterdam



Artwork by **Alan Mandel** & **Dan Barry**

Contents

Abstract

Comics are a uniquely multimodal medium, blending visual and textual elements to convey narrative. While computational approaches to comic analysis have traditionally focused on visual content alone, recent advances suggest that integrating textual information can enhance model performance in understanding comic structure and storytelling. In this study, we investigate whether incorporating textual features into comic panel representations improves multimodal understanding compared to unimodal (image-only) models. We introduce NIGHTMARE (**N**eural **I**ntegrator of **G**raphics and **H**uman **T**ext for **M**ulti-mod**A**l **R**epresentation **E**ncoding), an extension of the ASTERX model, designed to process and fuse textual and visual modalities through dedicated representation channels. Using a Western comic dataset and next-panel prediction as a benchmark, we demonstrate that NIGHTMARE significantly outperforms both visual-only and simple multimodal baselines, achieving performance gains of over 20% (Recall - R@K) in some configurations. Furthermore, we explore partial use of textual embeddings and differentiation between dialogue and narration, finding both to offer additional performance improvements in specific scenarios. These findings underscore the importance of multimodal integration in comic book research and offer a robust foundation for future exploration in complex comic understanding tasks.

Chapter 1

Introduction

Comics are a multimodal medium of expression that combines language and visual components [6]. Comic books, in the form recognized today, originated in the 1930s [40]. The lack of preservation of early comic books has resulted in a scarcity of comprehensive datasets [21], which in turn has resulted in limited research in this area. Computational research on comic books has traditionally focused solely on their visual components (i.e., the images) [21], in contrast to the humanities, which has long examined the multimodal nature of comics [17]. This is starting to change, as some recent studies have explored how to combine text with images to capture a comic panel in a single feature representation, with the goal of predicting emotions or the next panels [15, 30, 37]. Not all tasks in comic book research require the same level of comic book understanding, that is, the ability of a model to learn the underlying characteristics of comic books. For example, emotion prediction is primarily a computer vision task that can be addressed using standard vision tools without deeper comic-specific comprehension [39]. Vivoli et al. [39] have categorized the tasks in comic book research into five categories based on the required level of understanding. These categories classify tasks based on their input and output modalities, as well as the spatio-temporal reasoning required to process comic data. The first category encompasses relatively simple tasks such as action detection. As we move through the layers, the tasks grow increasingly complex, culminating in the fifth layer, which involves advanced tasks like 3D model generation. In this study, we focus on the third level, Retrieval and Modification, which has so far been dominated by studies on Japanese comics or has focused on the visual aspects of comics [39]. In this study, we contribute by utilizing a Western comic dataset for retrieval and by incorporating the text from panels into the comic book representations. Our study can potentially lead to insights that could be used in different domains that use textual and visual features. We aim to build upon the findings of Titarsolej et al. [10], who developed the ASTERX (*A Self-supervised Transformer Encoder for comic panel Representation eXtraction*) model for this task. We will enhance it further by incorporating textual information into the ASTERX model. As shown in Fig 1.1, text can constitute a significant portion of a comic book panel, reinforcing the importance of incorporating it in the model. Textual information offers deeper insight into the storyline of a comic book, which can significantly enhance the retrieval of relevant images, to evaluate this, we use next-panel prediction as a benchmark for our methods. The main research question will be thus: **Can integrating textual features into comic representations improve comics understanding in multimodal models, compared to unimodal models (which only use image features)?**

Predicting the next panel can be a difficult task, as the images in a comic sequence differ much more compared to video sequences [9]. If our methods outperform the baseline, it would indicate a deeper understanding of comics beyond the methodologies currently in use. The next three sub-research questions will aid in investigating whether our methods improve upon

(a) (b) (c)

Figure 1.1: Three examples of panels with text. A panel with some text (**a**), one which is almost half text (**b**) and one that only has text and no other visual elements (**c**).

the baseline.

How can text-based and image-based embeddings be effectively combined for comic representation learning? To explore this, we draw inspiration from prior research that combines textual and visual features in comics. Approaches for fusing features from different modalities range from model-based methods [4] to simpler strategies like embedding addition [37]. We will study both types of methods, such as feed-forward neural networks, as well as direct combination strategies, to understand how best to fuse these modalities.

Can textual information be used partially to improve retrieval performance in specific scenarios? Rather than relying on fully incorporating textual features for all retrieval instances, we will explore whether incorporating text features partially in targeted situations, such as using them when a certain threshold is met, can lead to more efficient or accurate results. This partial use of text features may be particularly beneficial when using textual features fully would lead to a decrease in performance.

Does distinguishing between balloon text and description text improve retrieval performance? In our prior questions, all text within a panel has been treated as a single unit. We will investigate whether treating balloon text (dialogue) and description text (narration) as separate entities yields better results. This will involve adapting our fusion methods to account for this textual differentiation and evaluating its effect on retrieval accuracy.

Our main contributions are the expansions of ASTERX which incorporate text into a new model called NIGHTMARE (see Section 3). Certain NIGHTMARE variants result in a performance boost of over 20%. Our findings thus indeed suggest that textual information can enhance retrieval performance. The performance can be even further improved by partially using textual embeddings or incorporating the difference between dialogue- or narrative like text into the model.

Chapter 2

Related Work

This chapter surveys the various layers involved in the tasks utilized in comic book research. It then examines different methods for embedding text and images individually. Finally, it reviews existing approaches for combining text and images.

2.1 Layers of Comics Understanding

In [39], the authors develop a framework to classify the tasks that are involved in comic book research into five layers. The higher layers tasks' are considered more challenging and thus also require the used models to have greater comic book understanding [2]. This taxonomy categorizes tasks based on their input and output formats, as well as the spatial and temporal reasoning required for analysing comic data. It promotes a clear, layered approach to understanding comics with a goal of highlighting key challenges that remain unsolved in the field.

- *The first layer comprises tasks with unimodal input (single image or sequence of images) and unimodal output (Images or Text): Tagging and Augmentation.*
- *The second layer comprises three distinct task groups: grounding, analysis, and segmentation. These tasks dig deeper into the intricate components of comics, focusing on detailed elements like panels, text, and characters and their ordering and associations*
- *The third layer focuses on advanced image and text understanding through tasks such as Retrieval (uni-modal and multi-modal) and Modification*
- *The fourth layer proposes a set of tasks related to multimodal understanding, which comprehends often reasoning through a text-rich image given a text prompt and one or more images as input.*
- *The fifth layer explores the creative frontier in comics analysis, where the synthesis and generation of comics from various media sources play a pivotal role.*

Our study focuses on the third layer, specifically addressing the task of retrieving images from a sequence, which closely corresponds to the objective of this layer.

Tasks in this framework of layers complexity often face similar challenges. The first constraint being that the research in the domain of comic books suffers from a lack of datasets [36]. The majority of comic books are copyrighted, which prevents their use in research or, when possible, limits their publication for reproducibility purposes. As a result, only a few datasets are publicly available. An additional issue is the lack of transcribed text for the existing datasets. Many researchers often rely on existing optical character recognition (OCR)

methods, such as Textract, for text extraction [5, 37]. However, this often leads to suboptimal transcriptions.

Another constraint is the lack of variety in available datasets [35]. These datasets typically consist of comics from a single genre or comics with similar styles, making generalization beyond these traits challenging. Furthermore, combining different datasets is often difficult due to the varying configurations of each dataset. These constraints highlight the urgent need for more diverse, well-annotated, and openly accessible comic datasets to support progress across all levels of task complexity. Our research focuses on a single dataset, the COMICS dataset [21], which comprises a large collection of older open-source comic books. As a result, our methods may face challenges in generalizing to more modern or stylistically diverse comic datasets.

2.2 Textual and Visual Embedding Techniques

The embedding of textual and visual information into feature representations is detailed below.

There are three major methods for embedding text into an embedding [20]. First, we have simple techniques like Bag-of-Words [11] or TF-IDF [31]. These methods are statistical and more intuitive compared to neural networks. Although they are straightforward, they fail to capture the depth of meaning required for more complex tasks. Second, various pretrained methods can be used to create text-based embeddings. For example, Word2Vec [22] generates embeddings by training a neural network on large text corpora. BERT [13], on the other hand, uses a transformer-based architecture to learn bidirectional representations by masking words in a sentence and predicting the masked words. Methods like these are often optimal for capturing context-related information in the embeddings. However, they can be computationally expensive and may struggle with out-of-domain words, which can be particularly challenging for comic books. Finally, there is the option of training a model from scratch, for example, based on a transformer-like architecture [34]. The advantage of this approach is that it can be specifically tuned to the desired domain. However, training and fine-tuning a model from scratch can be challenging due to the high number of potential parameters and a lack of large datasets.

For embedding images into feature representations, several methods can be used. The most well-known are Convolutional Neural Networks (CNNs), such as VGG16 [29] and ResNet [18], which are commonly employed for standard image classification tasks and general feature extraction. CNNs are a type of feedforward neural network that learns features through filter optimization. Similar to text embeddings, pretrained models can be used in a process called transfer learning [24], where pretrained models are fine-tuned on domain-specific data to improve performance. In addition, just as transformer models have been developed for text, Visual Transformers (ViT) [14] have been created for images. ViTs have gained popularity due to their ability to model long-range dependencies within images. The input image is divided into patches, and the transformer learns the contextual relationships between them to create embeddings. As an alternative to these deep learning-based methods, Latent Dirichlet Allocation (LDA), though primarily used for text, can be used for generating image-based embeddings. Although some work has been done using LDA for image classification [27], further research is needed, particularly in the context of comic book research. LDA works by identifying a mixture of underlying "topics" or features that best describe the distribution of features within the images.

Building upon this discussion of general embedding techniques, we now turn our attention to the process by which comic book panels are converted into embeddings. For instance, the Comicsformer model [30] employs a transformer architecture to process visual features extracted from comic panels, along with textual information from speech and narrative boxes. This approach enables the model to generate embeddings that encapsulate the semantic content of the

panels, facilitating tasks like text-cloze (retrieving only the right text compared to visual-cloze which focusses on retrieving the image) and character attribution. Furthermore, advancements in vision-language models (VLMs) have made it possible for dense captioning of comic panels. The ComiCap [35] pipeline utilizes VLMs to produce detailed, grounded captions for comic panels, enhancing the understanding of their content. By generating embeddings that link visual elements to descriptive text, this method supports various downstream tasks, including scene interpretation and character identification. Finally, the ComicsPAP [38] benchmark introduces a framework for evaluating models on tasks that require understanding the sequential arrangement of comic panels. By embedding panels in a manner that preserves their temporal and spatial relationships, models can better comprehend the narrative flow inherent in comic strips.

In our research, we adopt a model-based approach for embedding feature representations. For image features, we utilize the pretrained self-supervised Vision Transformer known as DINO [7]. For textual features, we employ BERT, a transformer-based model pretrained on a large English-language corpus using a self-supervised learning objective [13]. Both models and their integration into our framework will be discussed in greater detail in Chapter 3, Methodology.

2.3 Merging Text and Visual Data for Comic Understanding

We examine recent research on the application and integration of text and image embeddings within the field of comic book research.

Emotion analysis is a frequently studied task in comic book research [42, 23, 28, 32]. Because comics are inherently multi-modal, exploring their multi-modal nature to predict the emotions and sentiments associated with comic scenes has been a key focus for the comic research community [15]. A popular approach to multi-modal problems in comics is to treat the different modalities separately and then combine their respective outputs. This approach is taken by Dutta et al. [15], who propose a multi-task model called EmoComicNet, which combines both image and textual modalities to predict discrete emotion labels and the corresponding sentiment associated with comic scenes. The proposed EmoComicNet consists of four modules: the Image Module, the Textual Module, the Decision Module, and the Fusion Module. The Image Module extracts hierarchical visual features from the input images, while the Textual Module retrieves textual features from the input text. The Decision Module then performs similarity checking to determine the bi-modal interactions in the Fusion Module by combining the comic's textual and image modalities.

However, despite the introduction of models like EmoComicNet, The comic book research domain suffers from the lack of dedicated foundation models specifically tuned for comic book-related tasks. Soykan et al. [30] attempt to address this issue by establishing a method that can serve as a foundation model and a self-supervised objective to pre-train it in the domain of comic book research. The base model they developed is called Comicsformer. Comicsformer is a transformer encoder structure designed to process visual information of panels, speech bubble texts, narrative text, and, ultimately, character visuals and identity information across sequences of comic panels. The self-supervision task that trains Comicsformer to become a foundation model is called Masked Comic Modeling. Overall, the Comicsformer model and the self-supervision task form a general framework referred to as ComicBERT.

In [30], Soykan et al. also propose two novel tasks that are useful for comic book understanding. One of them is scene-cloze; unlike other cloze-style tasks that focus on a specific aspect (visual-cloze or text-cloze), scene-cloze attempts to predict all the elements in the next scene, rather than just some of them. The other is contextual character-to-speech attribution; this

task helps determine whether a given speech bubble context belongs to an existing character when the sequential panel context is provided.

Even when combing the visual and textual features in comic book research, the focus often remains on the visual aspect. Vivoli et al. [37] address this by focusing on a text-based task, the text-cloze task. They make four main contributions. First, they introduce a novel Multimodal-LLM based architecture specifically designed for the comics text-cloze task, outperforming existing models by 10% (accuracy) in both easy and hard variants of the task. The easy task of text-cloze is about predicting the text on the next panel given the options from all comics in the dataset, while the hard variant gives options from the same comic book (scene). Secondly, they propose new OCR text for the COMICS dataset [21] (Which we also utilize in our study), for which they used Textract [5]. Thirdly, they introduce a new version of the text-cloze task. In the standard version the model is given a set of choices (texts to choose from), in their variant the model needs to generate the text without any choices being given. This is meant as a more challenging task than the standard text-cloze task. Lastly, the authors compare various image representations and demonstrate that fine-tuning ResNet-50 [18] to the domain of the comic in a self-supervised manner (SimCLR [8]) yields comparable results to advanced Multimodal LLM image encoders, whilst having one-fifth of the parameters.

2.4 Composed Image Retrieval

We draw inspiration from related domains that address comparable multimodal reasoning tasks, aiming to leverage cross-disciplinary insights to advance the retrieval of comic book panels.

Composed Image Retrieval (CIR) is an image retrieval task that allows users to search for target images by providing a reference image along with a textual description of desired modifications. This multimodal query enables more nuanced and flexible searches compared to traditional methods that rely solely on visual or textual inputs. CIR has gained significant attention due to its practical applications in areas such as fashion, design, and content creation, where users often seek images with specific visual attributes or alterations [43, 41, 33]. The task involves understanding and integrating the semantic information from both the image and the text to retrieve images that accurately reflect the described modifications. This multimodal objective offers valuable inspiration for related challenges, such as comic book panel retrieval, where combining visual content with narrative context can potentially enhance search relevance.

Recent advancements in CIR have led to the development of various models and approaches. For instance, the SEARLE model [3] utilizes textual inversion to map visual features into a pseudo-word token in the CLIP token embedding space, facilitating zero-shot CIR without the need for labelled datasets. Similarly, CompoDiff [16] employs latent diffusion models to handle diverse conditions, including negative text and image mask queries, achieving state-of-the-art performance on multiple benchmarks. Additionally, the work by Baldrati et al. [4] introduces a task-oriented fine-tuning of CLIP encoders, followed by a Combiner network that integrates image and text features using contrastive learning. This method has demonstrated better performance on datasets like FashionIQ and CIRR, surpassing more complex state-of-the-art approaches.

Chapter 3

Methodology

In this chapter, we introduce the **N**eural **I**ntegrator of **G**raphics and **H**uman **T**ext for **M**ulti-mod**A**l **R**epresentation **E**ncoding (NIGHTMARE, named after the comic character *Nightmare*[1]). We start by outlining our methodology, beginning with a detailed description of the NIGHTMARE model and its variants. We then briefly discuss the creation of feature embeddings, followed by an overview of the datasets used in our experiments: COMICS and OpenMantra.

3.1 NIGHTMARE

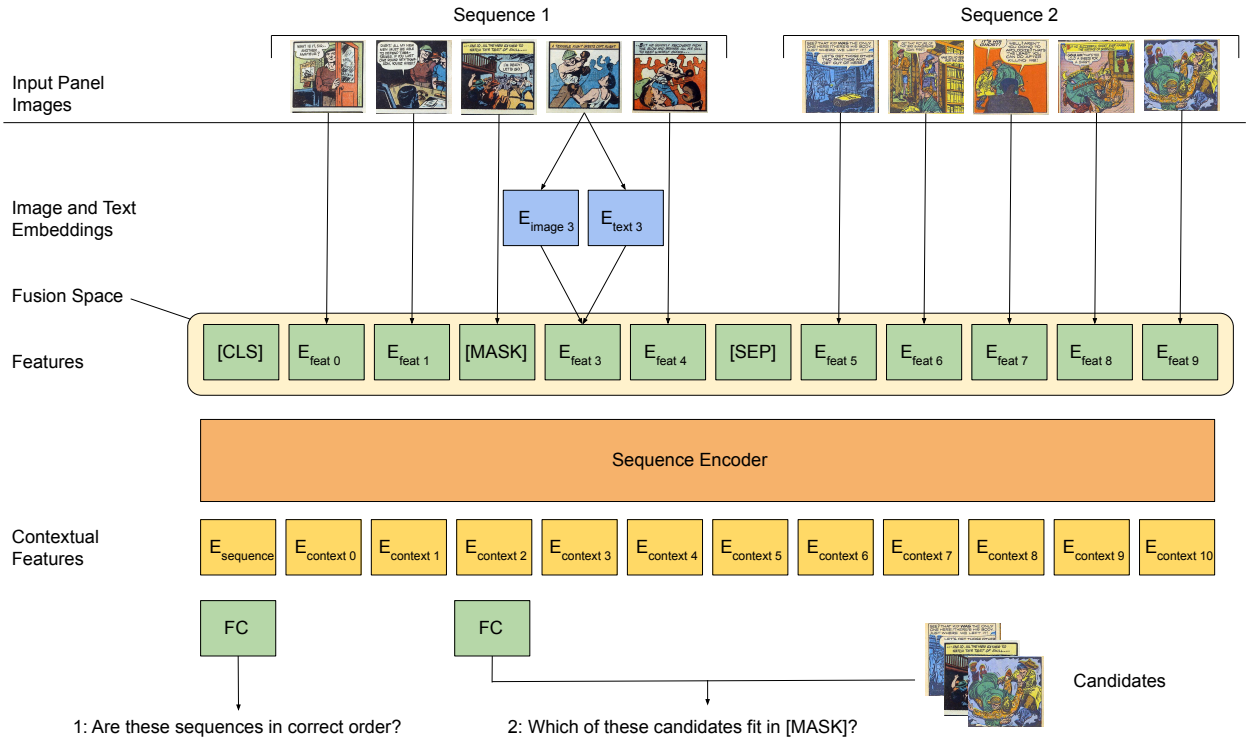


Figure 3.1: Overview of the NIGHTMARE model.

We use ASTERX, see Figure 3.1 (the ASTERX model would use image embeddings as the features), as the base point for our NIGHTMARE model. NIGHTMARE extends the ASTERX

[1]The hero Nightmare created by Alan Mandel & Dan Barry

model by introducing separate channels based on Gated Convolutional Networks (GCNs) [12] for the text and image embeddings.

The configuration *NIGHTMARE* from Figure 3.3 shows how those channels generate the features from the fusion space in Figure 3.1. Although originally developed for language modelling, GCNs [12] are a promising approach for extending and enhancing the ASTERX framework. The gating mechanism used in GCNs can be adapted to effectively control and combine multimodal features in a more flexible and efficient manner.

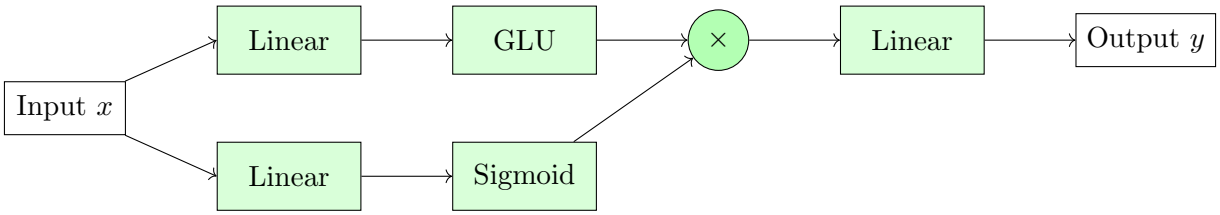


Figure 3.2: Compact GatedMLP architecture: Input x is processed in parallel through a GLU-activated MLP path and a gating path. The element-wise gated result is projected back to the original dimension. (The last Linear layer is our addition to the existing architecture by Dauphin et al. [12].)

Figure 3.2 shows the architecture of this mechanism. GCNs, as introduced by Dauphin et al. [12], employ a convolutional architecture for processing sequences, enabling efficient parallelization over sequential tokens. The core innovation of GCNs lies in their Gated Linear Unit (GLU), which modulates the output of convolutional layers to enhance gradient flow and capture intricate dependencies within the data. This approach contrasts with traditional recurrent models by offering a more efficient means of processing data with a finite context. The core innovation of GCNs lies in their Gated Linear Unit (GLU), which modulates the output of convolutional layers to enhance gradient flow and capture intricate dependencies within the data. This approach contrasts with traditional recurrent models by offering a more efficient means of processing data with a finite context. We refer the reader to our GitHub repository[2] for the full implementation of the NIGHTMARE model and the configurations of the parameters.

ASTERX [10], on which NIGHTMARE expands on, uses a novel self-supervised method for learning sequential representations from comics. This approach leverages the unique sequential nature of comics, where images are presented in a specific order to convey a narrative. By capturing the contextual relationships between panels, the method enhances the understanding of the storyline and visual elements. To optimise these representations ASTERX is trained with tasks inspired by masked language modelling (panel retrieval) and optimised using the cross-entropy loss. The authors evaluate their approach using the TINTIN Corpus (which we can not use due to the lack of text transcriptions), a dataset comprising over 1,000 comics from 144 countries, and demonstrate its effectiveness in both classification and retrieval tasks, outperforming baseline methods. These findings highlight the potential of sequential representation learning in uncovering cultural insights and advancing the analysis of comic narratives.

3.1.1 NIGHTMARE: Simple Variants

We briefly examine four simple NIGHTMARE model variants.

The first two variants employ relatively simple strategies for combining text and image embeddings. The first method uses element-wise addition (configuration *Element-Wise +*, Figure 3.3), where the visual and textual feature vectors are added together directly, assuming

[2] `https://github.com/Kacper970/NIGHTMARE`

Figure 3.3: Different configurations that can be used in the fusion space (see Figure 3.1). Highlighted in green is the configuration used in the final NIGHTMARE model. In configurations with only 1 text embedding block, the text encapsulates both dialogue and narration, not just dialogue.

they share the same dimensionality. This approach aims to create a fused representation that captures overlapping information between modalities in a straightforward manner. The second method relies on concatenation (configuration *Concatenation*, Figure 3.3), where the image and text embeddings are joined end-to-end to form a longer vector. This preserves the full information from both modalities and allows subsequent layers to learn how to integrate them effectively. Although these methods are computationally efficient and easy to implement, they serve primarily as baseline fusion techniques for comparison with more sophisticated approaches.

We proceed by using model based approaches for combining the text and image embeddings. Inspired by composed image retrieval we utilize the Combiner network [4] to combine the text and image representations in the NIGHTMARE framework (configuration *Combiner, FNN-Based*, Figure 3.3). During the training of the Combiner network, the authors train the general network in two stages. We focus on the second stage where the Combiner itself is trained as this is the architecture we will be utilizing. The design of the Combiner network leverages the progress made during the first stage of training (We replace that with the pretrained embeddings) and the enhanced additivity properties of the adapted embedding spaces. The network learns the residual of a convex combination of the image and text query features. Initially, both the image and text features are projected through a linear transformation followed by a ReLU activation. The resulting features are then concatenated and passed through two separate branches. The first branch computes the coefficients for the convex combination, using a linear layer followed by a ReLU function, another linear layer, and a sigmoid function. The output of the sigmoid provides the necessary coefficients for combining the image and text features. The second branch outputs the mixture contribution of the image and text features,

using a similar structure, but without the final sigmoid. In the end, the convex combination of the query features and the learned image-text mixture is obtained by summing the results of both branches.

Finally, we utilize a simple FNN-based method (configuration *Combiner, FNN-Based*, Figure 3.3). This approach utilizes a simple feedforward neural network, where concatenated text and image embeddings are passed through a linear layer, followed by a ReLU activation, and a final linear layer. Although the model is able to learn, its performance is lower compared to the default unimodal ASTERX. We hypothesize that the architecture has not been sufficiently fine-tuned for this specific task, which may explain the suboptimal results.

3.1.2 NIGHTMARE: 2 Channels Variants

Given that the main NIGHTMARE model utilizes two channels, we inspect variations on this design to further expand on the two channels principle.

First, we adopt various restrictions on the text embeddings channel. One NIGHTMARE variant includes the Single Digit Condition (configuration *Conditions, Masking, Film*, Figure 3.3). The Single Digit Condition applies a global gating value to interpolate between the original text-conditioned representation and a zeroed-out version, using weighted averaging.

Another NIGHTMARE variant introduces the Element-wise Condition (configuration *Conditions, Masking, Film*, Figure 3.3). This approach computes a separate gating value for each element in the representation, allowing for finer-grained control over feature integration.

A third variant includes Masking (configuration *Conditions, Masking, Film*, Figure 3.3), which selectively removes the text features under certain conditions, effectively simulating partial textual input.

Additionally, a NIGHTMARE variant includes the FiLM layer (configuration *Conditions, Masking, Film*, Figure 3.3). The FiLM (Feature-wise Linear Modulation) layer [25] applies a learned, feature-wise affine transformation to the visual feature representations, conditioned on a textual input. Given visual features f and a conditioning input c (e.g., text embeddings), the FiLM layer computes scaling and shifting parameters γ and β as linear projections of c, and modulates the features asFiLM$(f, c) = \gamma(c) \odot f + \beta(c)$, where $\odot$ denotes element-wise multiplication. This mechanism allows the model to adapt visual processing dynamically based on linguistic context.

Finally, another NIGHTMARE variant includes two channels in which the text embeddings are summed (configuration *Plus of Text Features*, Figure 3.3). This configuration simplifies the architecture by summing the two textual embeddings prior to processing, reducing the model to two channels. Compared to the the three channels variants in Section 3.1.3 which also distinguish between difference in text embeddings.

3.1.3 NIGHTMARE: 3 Channels Variants

This section presents additional NIGHTMARE variants that incorporate three channels. The existing NIGHTMARE approach, which utilizes two channels, performs well. Suggesting that expanding to more channels may be a worthwhile extension.

One such variant is the NIGHTMARE model with three channels (configuration *3 Channels*, Figure 3.3). This model processes dialogue, narration, and visual features in three independent channels.

Another set of variants explores different methods for combining the two text channels before integration with visual features:

- Summed Text Channels (configuration *3 Channels (+, Gate, Bilinear)*, Figure 3.3): The dialogue and narration embeddings are summed element-wise before being fused with the

visual features.

- GatedMLP Fusion (configuration *3 Channels (+, Gate, Bilinear*), Figure 3.3): The two text channels are combined using a GatedMLP network to enable adaptive feature integration.
- Bilinear Fusion (configuration *3 Channels (+, Gate, Bilinear*), Figure 3.3): The dialogue and narration embeddings are fused using a Bilinear layer, allowing multiplicative interaction between the channels.

Two additional variants modify the pathway structure for processing the text channels:

- Double Straight Wrap (configuration *Double Straight Wrap*, Figure 3.3): Each text type is processed independently through sequential GatedMLP layers before being fused with the visual modality.
- Double Cross Wrap (configuration *Double Cross Wrap*, Figure 3.3): This architecture diversifies the processing paths across the text channels by applying crossed GatedMLP layers, enabling cross-talk between dialogue and narration streams.

3.2 Baselines

For our baselines we use the Dino and Bert pretrained embeddings, see Section 3.3 and the CLIP embeddings. Radford et al. [26] introduced CLIP, a contrastive learning framework that jointly trains image and text encoders to align visual and linguistic representations in a shared embedding space. Trained on 400 million image–text pairs sourced from the internet, CLIP learns to associate images with their corresponding natural language descriptions without relying on manually curated labels. The model uses a contrastive objective that brings matching image-text pairs closer together while pushing mismatched pairs apart. Once trained, CLIP can perform a wide range of downstream vision tasks, such as image classification, retrieval, and zero-shot transfer, by simply comparing text prompts with image embeddings, eliminating the need for task-specific fine-tuning. This ability to generalize across tasks has made CLIP a foundational model in the field of vision and language research. We will utilize CLIP to generate embeddings for texts and images. The reasoning for the use of clip is that text and image embeddings originating from the same feature space should increase the performance compared to text and image embeddings that occupy different embeddings spaces.

Our last baseline is ComicBERT, which is a novel transformer-based architecture designed to process and understand the complex interplay of visual and textual elements in comics [30]. The model introduces a self-supervised pre-training objective called Masked Comic Modeling (MCM), inspired by BERT's masked language modeling, to train a foundational model capable of contextual understanding in comics. To fine-tune and validate the model, ComicBERT adopts existing cloze-style tasks and proposes new tasks, such as scene-cloze, which better capture the narrative and contextual intricacies unique to comics. Ultimately, ComicBERT aims to serve as a universal comic processor. The architecture leverages the Comicsformer model, which processes multimodal inputs, including images, text, and character information, to generate contextual embeddings. These embeddings are used in various downstream tasks, such as Text-Cloze, Visual-Cloze, Scene-Cloze, and Character Coherence, to evaluate the model's understanding of comic narratives. Given the absence of a publicly available code repository, we re-implemented ComicBERT from scratch. In our implementation, we excluded the character and body image channels, as such annotations are not available in our dataset. Additionally, we applied the same loss function used in ASTERX and NIGHTMARE to ensure a fair comparison between the models.

3.3 Feature Embedding

To represent the multimodal content of comic panels, we use pretrained models to embed both image and text information. For the visual modality, we use DINO [7], a self-supervised Vision Transformer that learns meaningful image representations without requiring labelled data. DINO is well-suited for capturing high-level semantic features in complex visual domains like comics. For the textual modality, we utilize the BERT model in its uncased variant (the text in our data is also uncased) [13], which ignores case distinctions and is pretrained on a large English corpus using a masked language modelling objective. This variant is particularly effective for general-purpose language understanding, we hope this leads to better generalization to the comic domain. The extracted features from both modalities are used as input to our models.

3.4 Data and Evaluation

In this paper we use the COMICS dataset, which is a widely used comics dataset [21]. This dataset is drawn from public domain Golden Age comic books and consists of over 1.2 million panels, extracted from approximately 4,000 comic book pages.

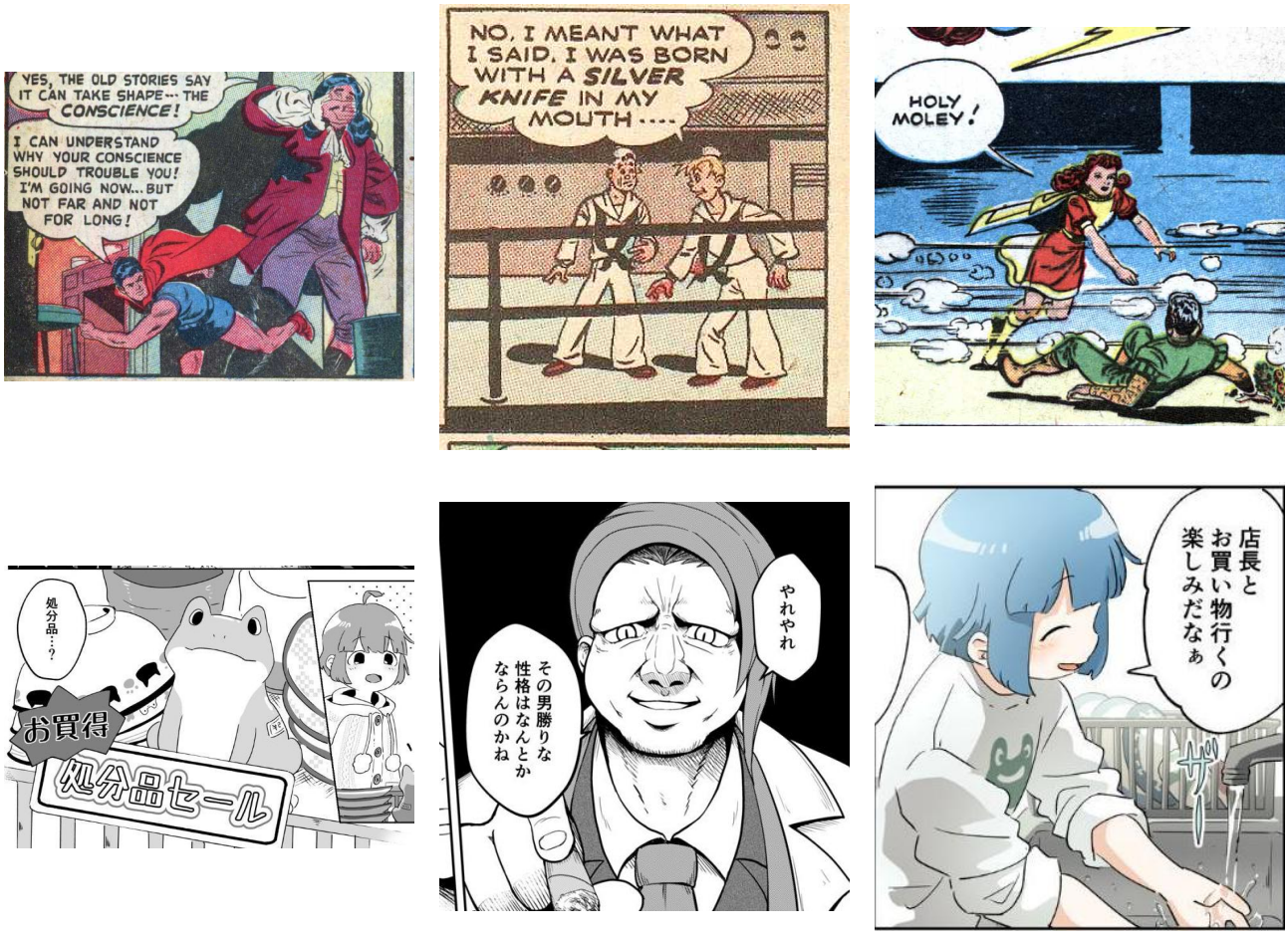


Figure 3.4: The top row shows examples from the COMICS dataset, while the bottom row shows examples from the OpenMantra dataset.

Each panel includes images, transcribed text (speech bubbles and narration), and panel ordering information, allowing for the study of both visual and textual storytelling elements. The dataset supports several tasks that require common sense reasoning and inference across

panels, inspired by the narrative phenomenon known as the "gutter", the space between panels where readers infer what happens. The authors define three cloze-style tasks (Text Cloze, Visual Cloze, and Character Coherence) to evaluate a model's ability to understand context and continuity in comics.

To determine whether our methods also generalize into a different domain and style we also use the OpenMantra dataset [19]. The OpenMantra dataset is designed as a benchmark for evaluating machine translation of manga. It consists of content from five Japanese manga series spanning various genres, such as fantasy, romance, action, mystery, and slice of life. The original Japanese dialogue has been professionally translated into both English and Chinese, providing high-quality parallel text for multilingual translation tasks. Given the small size of this dataset (around 200 panels) we only use it for testing purposes. Figure 3.4 shows some examples of the COMICS and OpenMantra datasets.

To evaluate NIGHTMARE and the other methods used in this study we use the Recall (R@K) metric. This and similar measures have been widely used in information retrieval [1]. We will focus on using Recall@1 (R@1), Recall@5 (R@5) and Recall@10 (R@10). Using Recall@1 (R@1) in retrieval tasks is particularly valuable because it directly measures the system's ability to return the most relevant result at the very top of the list. In many real-world applications, such as image retrieval, recommendation systems, or question answering users often rely on the first result to be correct or highly relevant. R@1 provides a clear and strict performance indicator: it shows the percentage of times the correct item is ranked first. This makes it a strong benchmark for assessing the precision and reliability of a retrieval model in high-stakes or user-facing scenarios where accuracy is critical. By focusing on the very first prediction, R@1 sets a high bar for retrieval quality and helps guide improvements in model performance. R@5 and R@10 can provide an insight in whether the model is on the right 'track'. Since this study focuses on the novel use of text in comic book representation achieving a high R@1 can be challenging. R@5 and R@10 can provide some more information about the performance of the used methods.

Chapter 4

Results and Discussion

Method	R@1	R@5	R@10
ASTERX (Image Only)	54.23	88.22	94.40
ASTERX (Text Only)	0.65	2.41	4.32
NIGHTMARE (Element-Wise +)	55.24	88.09	94.33
NIGHTMARE (Concatenation)	57.94	89.05	94.09
NIGHTMARE (Combiner)	0.06	0.56	1.17
NIGHTMARE (FNN-Based)	33.28	70.52	83.43
NIGHTMARE	**74.10**	**96.43**	**98.51**
ComicBERT (Multimodal)	62.85	92.26	96.04
Pretrained Clip (Multimodal)	11.02	25.21	32.28
Pretrained Clip (Image Only)	10.72	24.32	31.45
Pretrained Bert/Dino (Multimodal)	10.31	22.68	28.95
Pretrained Bert/Dino (Image Only)	10.32	22.40	28.82

Table 4.1: Different methods for combining text and image embeddings. Lower part of the table is the baselines we compare against.

Table 4.1 presents the results of various methods for combining text and image embeddings. Among these, CLIP and the combination of DINO and BERT baselines perform poorly. We suspect that their low performance is due to a lack of comic book related training material during their pre-training. In contrast, the ComicBERT baseline achieves better results than most evaluated methods, though it still falls short of our best-performing approach, NIGHTMARE. This highlights the effectiveness of our proposed enhancements, particularly the addition of distinct channels to the ASTERX network architecture. NIGHTMARE consistently outperforms competing methods by a significant margin.

As shown in Table 4.1, the ASTERX text-only variant fails to train effectively, resulting in notably poor performance. Meanwhile, two simpler fusion strategies, element-wise addition and embedding concatenation, achieve modest but noticeable improvements (compared to the image-only ASTERX). These results underscore the advantage of incorporating both visual and textual modalities, even with basic fusion techniques. Furthermore, the Combiner network, when paired with ASTERX (to form a NIGHTMARE variant), fails to train successfully. We hypothesize that this is due to architectural incompatibility, as ASTERX diverges significantly from the architecture the Combiner was originally trained to integrate.

Lastly, we evaluate a simple feed-forward neural network (FNN) approach, in which concatenated text and image embeddings are passed through a linear layer, followed by a ReLU activation and a final linear output layer. While this model is capable of learning, its performance

remains below that of the default unimodal ASTERX, suggesting that it lacks task-specific fine-tuning necessary for optimal results.

The introduction of NIGHTMARE suggests that incorporating dedicated channels for different modalities provides an effective solution to the sub-research question: **How can text-based and image-based embeddings be effectively combined for comic representation learning?**

Method	R@1	R@5	R@10
NIGHTMARE, (Single Digit Condition)	74.49	96.51	98.58
NIGHTMARE, (Element Wise Condition)	**75.33**	96.82	**98.80**
NIGHTMARE, (Masking)	75.05	**96.91**	98.74
NIGHTMARE, (FILM)	55.85	88.18	94.16

Table 4.2: Partial use of textual features.

To investigate how partially using text features affect multimodal fusion, we evaluate several variations of the NIGHTMARE model, as presented in Table 4.2. All variants, except for FILM, result in a slight improvement over the best-performing Channels baseline. Although FILM is effective in other contexts, it underperforms in this setting, likely due to its architectural assumptions not aligning well with the NIGHTMARE structure. Overall, the results indicate that simple gating and masking strategies are more effective for combing multimodal features in this task. This helps us answer us the sub-research question: **Can textual information be used partially to improve retrieval performance in specific scenarios?**. While using text features partially result in improvements, those improvements are not significant enough to make an definitive claim.

Method	R@1	R@5	R@10
NIGHTMARE, (3 Channels)	75.85	96.31	98.21
NIGHTMARE, (3 Channels, Text Channels: +)	74.99	**96.81**	98.60
NIGHTMARE, (3 Channels, Text Channels: Gate)	72.96	96.54	98.65
NIGHTMARE, (3 Channels, Text Channels: Bilinear)	74.13	96.43	98.53
NIGHTMARE, (3 Channels, Double Cross Wrap)	**76.17**	96.80	**98.78**
NIGHTMARE, (3 Channels, Double Straight Wrap)	75.94	96.28	98.25
NIGHTMARE, (2 Channels, Plus of Text Features)	74.50	96.63	98.59

Table 4.3: Splitting the text into narration and dialogue.

To investigate the impact of distinguishing between different types of text (dialogue and narration), we explore several variants of the NIGHTMARE model using different methods for combining textual (dialogue and narration separately) and visual features. This can be seen in Table 4.3, the variants achieve strong overall performance. We can see that simple using 3 channels in the NIGHTMARE model already yields a higher performance compared to the base NIGHTMARE model. Simple addition of adding the results of the text channels before processing it with the image channels decreases the performance. Using a GatedMLP network or Bilinear layer to combine the text channels provides more nuanced fusion but slightly underperforms compared to simple addition, suggesting that increased complexity does not necessarily translate to better retrieval performance in this context.

We also experiment with Double Wrap strategies to deepen textual processing. The Double Straight Wrap architecture processes each text type through sequential GatedMLP layers before

fusion, yielding one of the top performances. The Double Cross Wrap variant, which further diversifies processing paths across text channels by using crossed GatedMLP layers, achieves the highest R@1 score, suggesting that deeper and more diverse text-specific pathways may better capture the complementary nature of dialogue and narration.

Finally, NIGHTMARE (2 Channels, Plus of Text Features) simplifies the architecture by summing the two textual embeddings prior to processing, reducing the model to two channels. While this variant performs reasonably well, its performance lags behind the three-channel configurations, reinforcing the value of preserving and distinctly processing different textual modalities.

In our prior questions, all text within a panel has been treated as a single unit. However, treating balloon text (dialogue) and description text (narration) as separate entities thus yields better results (see Table 4.1 and 4.3). This clearly answers the sub-research question: **Does distinguishing between balloon text and description text improve retrieval performance?**

Our main research question was: **Can integrating textual features into comic representations improve comics understanding in multimodal models, compared to unimodal models (which only use image features)?** Based on our results, we conclude that NIGHTMARE and its variants indeed enhance comics understanding in the multimodal setting. This is particularly significant given the inherent difficulty of predicting the next panel in a comic sequence, where visual continuity is often much weaker than in video sequences [9]. Our findings suggest that introducing modality specific components into model architectures is crucial for achieving deeper narrative understanding and improved performance in comic representation tasks.

4.1 Ablation Analysis

Method	R@1	R@5	R@10
NIGHTMARE (Multimodal → Image Only, Element Wise +)	**55.74**	**88.34**	**94.41**
ASTERX (Image Only → Multimodal, Element Wise +)	53.44	87.85	94.36
NIGHTMARE* (Multimodal → Image Only)	17.86	47.81	63.37
NIGHTMARE (Image Only → Multimodal)	n/a	n/a	n/a

Table 4.4: Exploring using a multimodal model in an unimodal setting.

It is important to briefly consider the use of multimodal pretrained models in unimodal evaluation settings. As discussed earlier, the lack of comprehensive multimodal comic datasets means that, at evaluation time, textual information is often unavailable. This raises the question of whether a model trained with both text and image features can still perform effectively when only one modality is available during inference. Table 4.4 illustrates that, in the case of element-wise addition, the use of a multimodal pretrained model does result in a slight performance improvement compared to the base ASTERX performance shown in Table 4.1. For completeness, we also include our best-performing method, NIGHTMARE. However, due to its inherently multimodal architecture, adapting it to a unimodal evaluation setting results in a significant drop in performance (we use the image embeddings twice to account for the absent text embeddings). Despite this limitation, the results highlight a promising direction for future work in developing more adaptable and robust multimodal models. Training NIGHTMARE on a unimodal dataset and evaluating in a multimodal manner is not very logical as we can not

use the advantage of the channels during training properly, that is why this method is left out (see Table 4.4, left in the table for a complete overview).

Method	**R@1**	**R@5**	**R@10**
ASTERX (Image Only), English	1.04	4.64	10.06
ASTERX (Image Only), Chinese	0.66	5.59	9.28
NIGHTMARE (Element-wise +), English	0.66	6.75	11.54
NIGHTMARE (Element-wise +), Chinese	0.66	6.09	11.82
NIGHTMARE, English	**5.38**	26.48	**38.97**
NIGHTMARE, Chinese	5.19	**27.84**	38.28

Table 4.5: Evaluation on OpenMantra dataset with the English and Chinese languages.

To evaluate whether our methods generalize beyond the style and language of the COMICS dataset used during training, we also test them on the Japanese comic dataset OpenMantra. For this evaluation, we use the panel images from OpenMantra along with the professionally translated English and Chinese texts. We opt for these translations instead of the original Japanese text because they can be embedded using language specific BERT models, one for English and one for Chinese, enabling a more consistent and comparable representation across modalities. As shown in Table 4.5, our best-performing method, NIGHTMARE, achieves up to a fivefold improvement over the unimodal baseline (for the **R@1** score). Although the absolute performance remains low, this result is nonetheless promising, as the OpenMantra dataset represents entirely unseen data for our model. These findings suggest that the proposed approach has the potential to generalize across diverse comic styles and languages, despite being trained on a single, Western-centric dataset.

4.2 Sample Output Analysis

While quantitative metrics provide a clear picture of overall performance, they cannot fully capture the nuanced differences observed in the outputs of our NIGHTMARE model. In particular, many of the images correctly retrieved by NIGHTMARE depict more dynamic or contextually rich scenes, such as expressive character poses, shifting environments, or visually implied actions, that are difficult to quantify using standard evaluation scores. These dynamic elements often convey subtle narrative cues that benefit from multimodal understanding. NIGHTMARE appears to be more adept at interpreting these cues, likely due to its ability to integrate both visual and textual context.

Figure 4.1 presents examples that NIGHTMARE retrieves correctly but are missed by the ASTERX model. Conversely, Figure 4.2 shows examples that ASTERX classifies correctly but NIGHTMARE does not. These qualitative results suggest that NIGHTMARE excels in cases with significant dynamic movement or contextual shifts between panels, whereas ASTERX performs better on sequences that are more visually similar. For a fair comparison, all examples are selected from the same comic books. See the captions of Figures 4.1 and 4.2 for additional context.

Figure 4.1: Correct next-panel predictions by the NIGHTMARE model. The predicted panel is highlighted in green and is shown in context, surrounded by its preceding and succeeding panels to illustrate narrative coherence. The top row corresponds to comic 998, the middle row to comic 720, and the bottom row to comic 896 from the COMICS dataset.

Figure 4.2: Correct next-panel predictions by the ASTERX model. The predicted panel is highlighted in green and is shown in context, surrounded by its preceding and succeeding panels to illustrate narrative coherence. The top row corresponds to comic 998, the middle row to comic 720, and the bottom row to comic 896 from the COMICS dataset.

Chapter 5

Conclusion

The main research question studied in this thesis was: **Can integrating textual features into comic representations improve comics understanding in multimodal models, compared to unimodal models (which only use image features)?** To investigate this, we developed NIGHTMARE, a novel model architecture designed to process and integrate both visual and textual information from comic panels. A key feature of NIGHTMARE is the incorporation of modality specific channels, allowing the model to treat text and image features separately before fusion.

Through comprehensive quantitative and qualitative evaluation, we have shown that NIGHTMARE outperforms traditional unimodal models as well as baseline multimodal fusion techniques. Notably, the model demonstrates a particular strength in interpreting dynamic or narrative driven comic sequences. While experiments with separating textual input into dialogue and narration channels yielded slightly better performance, we recommend using the base version of NIGHTMARE. This is primarily due to the limited availability and inconsistency of annotated data that distinguishes between these text types, which may hinder generalizability. Overall, our findings support the effectiveness of multimodal representation learning for comics.

Acknowledgements

I would like to express my gratitude to my supervisor ***Nanne van Noord*** for his invaluable guidance, support, and encouragement throughout this work. His insights and advice were instrumental in shaping this project.

I also wish to thank ***Rénan van Dijk*** for helping with proofreading the final version and ensuring the clarity and accuracy of the final text.

Finally, I am truly grateful to my family for their unwavering support, patience, and understanding during this journey.

Bibliography

[1] Monika Arora, Uma Kanjilal, and Dinesh Varshney. Evaluation of information retrieval: precision and recall. *International Journal of Indian Culture and Business Management*, 12(2):224–236, 2016. URL: https://www.researchgate.net/publication/292671645_Evaluation_of_information_retrieval_precision_and_recall.

[2] Olivier Augereau, Motoi Iwata, and Koichi Kise. A survey of comics research in computer science. *Journal of imaging*, 4(7):87, 2018. URL: https://arxiv.org/abs/1804.05490.

[3] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15338–15347, 2023. URL: https://arxiv.org/abs/2303.15247.

[4] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Composed image retrieval using contrastive learning and task-oriented clip-based features. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3):1–24, 2023. URL: https://arxiv.org/abs/2308.11485.

[5] Edouard Belval, Thomas Delteil, Martin Schade, and Srividhya Radhakrishna. Amazon Textractor, 2024. URL: https://github.com/aws-samples/amazon-textract-textractor.

[6] Casey Brienza and Paddy Johnston. *Cultures of comics work*. Springer, 2016. URL: https://link.springer.com/book/10.1057/978-1-137-55090-3.

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. URL: https://arxiv.org/abs/2104.14294.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. URL: https://arxiv.org/abs/2002.05709, arXiv:2002.05709.

[9] Neil Cohn, Irmak Hacımusaoğlu, 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. URL: https://www.tandfonline.com/doi/full/10.1080/21504857.2022.2152067.

[10] Neil Cohn, Nanne van Noord, and Sam Titarsolej. Drawing insights: Sequential representation learning in comics. In *35th British Machine Vision Conference 2024*, 2024. The 35th British Machine Vision Conference, BMVC 2024 ; Conference date: 25-11-2024 Through 28-11-2024. URL: https://bmva-archive.org.uk/bmvc/2024/papers/Paper_650/paper.pdf.

[11] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague, 2004. URL: https://www.cs.cmu.edu/~efros/courses/LBMV07/Papers/csurka-eccv-04.pdf.

[12] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR, 2017. URL: https://arxiv.org/abs/1612.08083.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, 2019. URL: https://arxiv.org/abs/1810.04805.

[14] 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, 2021. URL: https://arxiv.org/abs/2010.11929, arXiv:2010.11929.

[15] Arpita Dutta, Samit Biswas, and Amit Kumar Das. Emocomicnet: A multi-task model for comic emotion recognition. *Pattern Recognition*, 150:110261, 2024. URL: https://www.sciencedirect.com/science/article/pii/S0031320324000128, doi:10.1016/j.patcog.2024.110261.

[16] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. Compodiff: Versatile composed image retrieval with latent diffusion. *arXiv preprint arXiv:2303.11916*, 2023. URL: https://arxiv.org/abs/2303.11916.

[17] Robert C Harvey. *The art of the comic book: An aesthetic history*. Univ. Press of Mississippi, 1996. URL: https://slims.umn.ac.id/index.php?p=show_detail&id=13170.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. URL: https://ieeexplore.ieee.org/document/7780459.

[19] Ryota Hinami, Shonosuke Ishiwatari, Kazuhiko Yasuda, and Yusuke Matsui. Towards fully automated manga translation, 2021. URL: https://arxiv.org/abs/2012.14271, arXiv:2012.14271.

[20] Francesca Incitti, Federico Urli, and Lauro Snidaro. Beyond word embeddings: A survey. *Information Fusion*, 89:418–436, 2023. URL: https://www.sciencedirect.com/science/article/pii/S1566253522001233, doi:10.1016/j.inffus.2022.08.024.

[21] Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan Boyd-Graber, Hal Daume, and Larry S Davis. The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 7186–7195, 2017. URL: https://arxiv.org/abs/1611.05118.

[22] Tomas Mikolov. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 3781, 2013. URL: https://arxiv.org/abs/1301.3781.

[23] Nhu-Van Nguyen, Xuan-Son Vu, Christophe Rigaud, Lili Jiang, and Jean-Christophe Burie. Icdar 2021 competition on multimodal emotion recognition on comics scenes. In *Document Analysis and Recognition – ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part IV*, page 767–782, Berlin, Heidelberg, 2021. Springer-Verlag. doi:10.1007/978-3-030-86337-1_51.

[24] SJ—Yang Pan. Q.: A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. URL: https://ieeexplore.ieee.org/abstract/document/5288526.

[25] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer, 2017. URL: https://arxiv.org/abs/1709.07871, arXiv:1709.07871.

[26] 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, 2021. URL: https://arxiv.org/abs/2103.00020, arXiv:2103.00020.

[27] Nikhil Rasiwasia and Nuno Vasconcelos. Latent dirichlet allocation models for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2665–2679, 2013. URL: https://ieeexplore.ieee.org/document/6494575.

[28] Charles Lima Sanches, Olivier Augereau, and Koichi Kise. Manga content analysis using physiological signals. In *Proceedings of the 1st International Workshop on CoMics ANalysis, Processing and Understanding*, MANPU '16, New York, NY, USA, 2016. Association for Computing Machinery. doi:10.1145/3011549.3011555.

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. URL: https://arxiv.org/abs/1409.1556.

[30] Gürkan Soykan, Deniz Yuret, and Tevfik Metin Sezgin. Comicbert: A transformer model and pre-training strategy for contextual understanding in comics. In *Document Analysis and Recognition – ICDAR 2024 Workshops: Athens, Greece, August 30–31, 2024, Proceedings, Part I*, page 257–281, Berlin, Heidelberg, 2024. Springer-Verlag. doi:10.1007/978-3-031-70645-5_16.

[31] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972. URL: https://dl.acm.org/doi/10.5555/106765.106782.

[32] Hideki Tanaka, Ryosuke Yamanishi, and Junichi Fukumoto. Relation analysis between speech balloon shapes and their serif descriptions in comic. In *2015 IIAI 4th International Congress on Advanced Applied Informatics*, pages 229–233, 2015. doi:10.1109/IIAI-AAI.2015.235.

[33] Prateksha Udhayanan, Srikrishna Karanam, and Balaji Vasan Srinivasan. Learning with multi-modal gradient attention for explainable composed image retrieval. *arXiv preprint arXiv:2308.16649*, 2023. URL: https://arxiv.org/abs/2308.16649.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. URL: https://arxiv.org/abs/1706.03762.

[35] Emanuele Vivoli, Niccolò Biondi, Marco Bertini, and Dimosthenis Karatzas. Comicap: A vlms pipeline for dense captioning of comic panels, 2024. URL: https://arxiv.org/abs/2409.16159, arXiv:2409.16159.

[36] Emanuele Vivoli, Irene Campaioli, Mariateresa Nardoni, Niccolò Biondi, Marco Bertini, and Dimosthenis Karatzas. Comics datasets framework: Mix of comics datasets for detection benchmarking, 2024. URL: https://arxiv.org/abs/2407.03540, arXiv:2407.03540.

[37] Emanuele Vivoli, Joan Lafuente Baeza, Ernest Valveny Llobet, and Dimosthenis Karatzas. Multimodal transformer for comics text-cloze. In *International Conference on Document Analysis and Recognition*, pages 128–145. Springer, 2024. URL: https://arxiv.org/abs/2403.03719.

[38] Emanuele Vivoli, Artemis Llabrés, Mohamed Ali Souibgui, Marco Bertini, Ernest Valveny Llobet, and Dimosthenis Karatzas. Comicspap: understanding comic strips by picking the correct panel, 2025. URL: https://arxiv.org/abs/2503.08561, arXiv:2503.08561.

[39] Emanuele Vivoli, Mohamed Ali Souibgui, Andrey Barsky, Artemis LLabrés, Marco Bertini, and Dimosthenis Karatzas. One missing piece in vision and language: A survey on comics understanding, 2025. URL: https://arxiv.org/abs/2409.09502, arXiv:2409.09502.

[40] Bradford W Wright. *Comic book nation: The transformation of youth culture in America*. JHU Press, 2003. URL: https://academic.oup.com/jah/article-abstract/89/1/295/689509?login=false.

[41] Cairong Yan, Meng Ma, Yanting Zhang, and Yongquan Wan. Dual-path multimodal optimal transport for composed image retrieval. In *Proceedings of the Asian Conference on Computer Vision*, pages 1741–1755, 2024. URL: https://link.springer.com/chapter/10.1007/978-981-96-0960-4_15.

[42] Yi-Ting Yang and Wei-Ta Chu. Manga text detection with manga-specific data augmentation and its applications on emotion analysis. In Duc-Tien Dang-Nguyen, Cathal Gurrin, Martha Larson, Alan F. Smeaton, Stevan Rudinac, Minh-Son Dao, Christoph Trattner, and Phoebe Chen, editors, *MultiMedia Modeling*, pages 29–40, Cham, 2023. Springer Nature Switzerland. URL: https://link.springer.com/chapter/10.1007/978-3-031-27818-1_3.

[43] Gangjian Zhang, Shikui Wei, Huaxin Pang, Shuang Qiu, and Yao Zhao. Enhance composed image retrieval via multi-level collaborative localization and semantic activeness perception. *IEEE Transactions on Multimedia*, 26:916–928, 2023. URL: https://ieeexplore.ieee.org/document/10120671.