UAB
Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

This is the **published version** of the bachelor thesis:

Singh, Amritpal. *Comics VQA*. Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Grau en Intel·ligència Artificial)

This version is available at https://ddd.uab.cat/record/336302

under the terms of the CC BY-SA license.

Comics VQA

Amritpal Singh

June 28, 2026

Abstract

Comics present a unique challenge for artificial intelligence due to their multimodal nature, combining visual storytelling, textual dialogue, and long-range narrative context across sequential pages. Although comic understanding encompasses a wide range of tasks, including visual question answering, object detection, segmentation, character recognition, and optical character recognition (OCR), existing approaches typically address these tasks independently, and there is currently no comprehensive benchmark for evaluating the general comic understanding capabilities of modern vision-language models (VLMs).

In this thesis, we introduce *Comics VQA*, a benchmark designed to evaluate multimodal comic understanding through a unified natural language interface. The benchmark contains both visual question answering and panel localization tasks, requiring models to perform OCR, visual reasoning, long-context understanding, and structural localization across multiple comic pages.

In addition, we investigate whether adapting a VLM to the comic domain through instruction fine-tuning on an automatically generated panel localization task improves its performance on broader comic understanding tasks. Training supervision is generated automatically from existing panel annotations, eliminating the need for manually annotated training data. Experimental results show that, while the proposed fine-tuning strategy substantially improves panel localization performance, these improvements do not transfer to the broader reasoning capabilities evaluated by *Comics VQA*. These findings demonstrate the challenges of comic-domain adaptation and highlight the need for richer training objectives that better capture the multimodal and narrative characteristics of comics.

Keywords: Comics Understanding, Vision-Language Models, Model Evaluation, Foundational Model

1 Introduction

Comics are a unique form of storytelling that combine visual content, textual dialogue, and narrative structure to convey complex stories. Understanding comics requires jointly interpreting visual elements, textual information, contextual relationships, and long-range narrative dependencies, making the domain particularly challenging for multimodal artificial intelligence systems. Furthermore, comic understanding encompasses a wide range of tasks, including object detection, segmentation, optical character recognition (OCR), dialogue understanding, question answering, and narrative reasoning. Existing approaches typically address these tasks using separate task-specific models, resulting in fragmented solutions rather than a unified framework for comprehensive comic understanding.

In recent years, artificial intelligence research has increasingly shifted toward foundation models, broadly trained models capable of adapting to multiple downstream tasks within a unified architecture [1]. Recent multimodal systems, including ChatGPT, Gemini, and open-source vision-language models (VLMs) such as Qwen, have demonstrated strong visual understanding and multimodal reasoning capabilities, making them promising starting points for broad comic understanding. Rather than developing separate models for each comic analysis task, a natural long-term research direction is to adapt a single VLM capable of supporting multiple comic understanding tasks. The concept of such a unified comic understanding model is illustrated in Figure 1.

Motivated by this vision, this thesis investigates whether an existing vision-language model can be effectively adapted to the comic domain. Instead of developing a new architecture, we use the open-source Qwen3-VL model as our base model and explore domain adaptation through instruction fine-tuning on a comic-specific panel localization task. Given one or more comic pages and a natural language description, the model is trained to identify the corresponding panel by predicting its page number and bounding box. The training annotations are generated automatically using an existing panel detector rather than being manually created by human annotators, enabling large-scale adaptation

• Contact E-mail: amritpal.singh@autonoma.cat
• Supervised by: Dimosthenis Karatzas (Department of Computer Science and Artificial Intelligence)
• Academic Year 2025/26

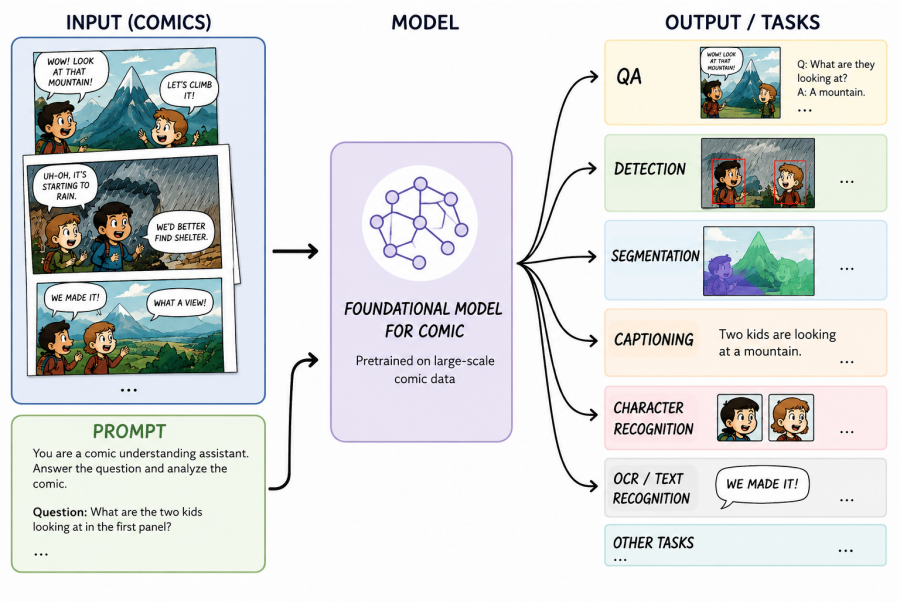


Fig. 1: Conceptual overview of a unified foundation model for comic understanding. Multiple comic pages are provided as input to a single vision-language model capable of supporting diverse downstream tasks, including question answering, object detection, segmentation, captioning, character recognition, and OCR/text understanding.

with minimal annotation effort.

Our research hypothesis is that adapting a vision-language model to the comic domain through the panel localization task enables it to learn the structural organization of comic pages, including panel boundaries and their spatial relationships. We hypothesize that this structural understanding transfers to visual question answering tasks, particularly those that require reasoning about page layout, locating relevant information, and integrating context across multiple panels or pages.

Evaluating this hypothesis requires a benchmark that measures whether the structural knowledge acquired during panel localization transfers to general comic understanding. Existing comic datasets, including eBDtheque [2], DCM772 [3], Manga109 [4], PopManga [5], and CoMix [6], have significantly contributed to comic analysis by providing datasets for individual tasks. However, they were not designed as unified evaluation benchmarks for modern vision-language models and therefore provide limited insight into whether domain adaptation transfers across different comic understanding capabilities.

To address this limitation, we introduce *Comics VQA*, a benchmark designed to evaluate diverse comic understanding tasks through a unified natural language interface. The benchmark consists of question answering tasks, where the expected output is a textual or numerical answer, together with panel localization tasks, where models must return the page number and bounding box of the panel described by a natural language query. These tasks require reasoning over visual content, textual information, and long narrative contexts spanning multiple comic pages, providing a comprehensive evaluation of multimodal comic understanding.

Finally, we evaluate the zero-shot performance of Qwen3-VL on Comics VQA to establish a baseline and subsequently assess the effect of comic-specific fine-tuning on the proposed benchmark. Through this study, we investigate whether adapting a general-purpose vision-language model using automatically generated supervision improves its ability to perform a broad range of comic understanding tasks.

The main contributions of this thesis are summarized as follows:

- We introduce *Comics VQA*, a benchmark for evaluating modern vision-language models on diverse comic understanding tasks, including both visual question answering and panel localization.
- We investigate comic-specific domain adaptation by instruction fine-tuning the Qwen3-VL model on a panel localization task using automatically generated annotations, with the goal of improving the model's understanding of comic page structure.
- We perform a comprehensive experimental evaluation of state-of-the-art vision-language models on Comics VQA, establishing zero-shot baselines and analyzing the effect of comic-specific fine-tuning on downstream comic understanding tasks.

Resource Availability. To facilitate reproducible research, all resources developed as part of this thesis are publicly available. The implementation of the proposed methods, evaluation framework, and annotation tool is available on GitHub: https://github.com/amritpal200/ComicsLib/tree/comicslib. The *Comics VQA* benchmark is publicly available on Hugging Face: https://huggingface.co/datasets/Amrit2412/ComicsQA.

2 Related Work

Our work builds upon two main research directions: benchmarking for comic understanding and adapting vision-language models to the comic domain through comic-specific fine-tuning. In this section, we review existing

comic understanding datasets and benchmarks, followed by recent advances in vision-language models and domain adaptation techniques.

2.1 Domain Adaptation for Comic Understanding

Vision-language models are typically pretrained on large-scale image-text datasets and therefore require adaptation to perform effectively in specialized domains such as comics. Recent work has explored different strategies for adapting models to the comic domain, including self-supervised learning, semi-supervised learning, and domain-adaptive pretraining.

Vivoli et al. [7] investigated domain-adaptive self-supervised pretraining for comic understanding using SimCLR-based representation learning to adapt visual encoders for a text-cloze task. Their results demonstrate that leveraging unlabeled comic data can improve downstream multimodal performance.

Topal et al. [8] explored domain-adaptive self-supervised pretraining for face and body detection in stylized visual domains, including comics and cartoons. Their work showed that utilizing large amounts of unlabeled domain-specific data substantially improves detection performance while reducing the reliance on manually annotated training data.

Soykan et al. [9] proposed a semi-supervised framework for comic character re-identification that combines identity-aware supervision with contrastive learning between facial and body representations. Their approach improves the robustness of character understanding across different panels and visual contexts.

Although these works demonstrate that domain adaptation is beneficial for comic understanding, they remain focused on specialized tasks such as character detection, character re-identification, or text-cloze reasoning. In contrast, recent large vision-language models provide the opportunity to address multiple comic understanding tasks within a single architecture. Our work investigates whether adapting an instruction-tuned vision-language model to the comic domain through panel localization fine-tuning, using automatically generated annotations, improves its performance on broader comic understanding tasks. To evaluate this hypothesis, we introduce the *Comics VQA* benchmark, which measures whether the knowledge acquired during fine-tuning transfers to question answering and long-context comic reasoning.

2.2 Benchmarks for Comic Understanding

Several works have explored benchmarking and dataset creation for comic understanding. One of the most comprehensive resources is CoMix [6], a large-scale dataset and benchmark designed for comic analysis. It includes four computational tasks: object detection, speaker identification, character re-identification, and panel-text sorting, as well as two multimodal reasoning tasks: character naming and dialogue generation. These tasks require models to detect visual objects, understand relationships between entities, and process textual information within comic panels. However, CoMix is primarily designed around predefined task-specific evaluations rather than assessing the capabilities of modern vision-language models (VLMs) through prompt-based, unified multimodal reasoning.

Another relevant benchmark is MangaVQA [10], which introduces a question-answering dataset consisting of 526 manually constructed question–answer pairs, alongside MangaOCR for text detection and recognition. MangaVQA includes questions requiring both panel-level and page-level reasoning. While this represents an important step toward comic question answering, the benchmark is limited to manga-style comics, and performance is significantly influenced by the ability to accurately recognize Japanese text.

ComicsPAP [11] presents a benchmark focused on identifying missing panels in comic sequences. This setup evaluates certain aspects of narrative understanding and sequential reasoning, but it remains limited to panel prediction rather than open-ended multimodal reasoning, where models must extract information from both text and images to answer natural language questions.

Re:Verse [12] introduces a benchmark for manga narrative understanding, including tasks such as story generation and temporal reasoning. While relevant to long-context reasoning, it remains constrained to manga-style content and therefore does not reflect the broader diversity of comic formats.

Humor in Pixels [13] evaluates the ability of large multimodal models to understand humor in comics. Although this benchmark provides valuable insights into higher-level semantic understanding, it focuses on a specific aspect of comic comprehension rather than general multimodal reasoning across diverse tasks.

Zooming into Comics [14] introduces AI4VA-FG, a fine-grained benchmark for comic understanding that includes tasks such as panel understanding, action recognition, and depth comparison. However, the benchmark relies heavily on automatically generated annotations, limiting its reliability for evaluating complex reasoning capabilities.

Overall, existing benchmarks address important but narrow aspects of comic understanding, including object detection, narrative prediction, humor comprehension, and fine-grained visual analysis. However, there remains a lack of benchmarks specifically designed to evaluate modern vision-language models through unified prompt-based reasoning across multiple comic understanding tasks, particularly in settings requiring long-context multimodal reasoning over multiple pages. This limitation motivates the development of our proposed benchmark, *Comics VQA*.

3 Comic Domain Adaptation

3.1 Methodology

The objective of this chapter is to investigate whether a general-purpose vision-language model can be adapted to the comic domain through comic-specific instruction fine-tuning. Rather than training the model directly for visual question answering, we formulate a panel localization task in which the model learns to identify the location of comic panels from natural language instructions. Our hypothesis is that learning the structural organization of comic pages will improve the model's internal representations and transfer to downstream comic understanding tasks.

The proposed methodology consists of four main stages. First, comic pages and their associated panel annotations are collected and preprocessed to produce multi-page training samples. Second, instruction-response pairs are generated automatically from the panel annotations, eliminating the need for manually written training data. Third, the generated samples are used to fine-tune the Qwen3-VL model using a standard instruction tuning objective. Finally, the adapted model is evaluated on the proposed *Comics VQA* benchmark to determine whether the learned structural knowledge improves performance on downstream comic understanding tasks.

The following sections describe each stage of the proposed methodology in detail.

3.2 Adaptation Strategy

The objective of this work is to investigate whether a general-purpose vision-language model can be effectively adapted to the comic domain through comic-specific instruction fine-tuning. We use Qwen3-VL as the base model, focusing primarily on the 4B and 8B variants. Rather than training the model directly for visual question answering, we formulate a panel localization task in which the model learns to identify the position of comic panels from natural language instructions.

Our motivation is based on the hypothesis that learning the structural organization of comic pages, including panel boundaries and their spatial relationships, will improve the model's understanding of comic layouts. Since many visual question answering tasks require locating relevant panels before interpreting their visual and textual content, we expect that knowledge acquired during panel localization will transfer to downstream comic understanding tasks, particularly those involving long-context reasoning across multiple panels or pages.

Although the training task uses supervision, the target annotations are not manually created by human annotators. Instead, they are generated automatically from existing panel annotations, allowing large-scale adaptation with minimal annotation effort. Consequently, our approach can be viewed as a form of automatic supervision for comic domain adaptation.

Because our benchmark evaluates complete comic books during inference, the ideal training setup would expose the model to the same long-context inputs. However, processing entire comic books during training is computationally infeasible, as many books contain more than thirty pages. Instead, we train the model using smaller groups of consecutive comic pages while preserving the sequential structure of the narrative. Throughout our experiments, we investigate inputs consisting of two to five consecutive pages, allowing the model to learn cross-page spatial relationships while keeping the computational cost manageable.

3.3 Training Pipeline

The overall training pipeline consists of four main stages: data preparation, automatic training sample generation, model fine-tuning, and evaluation.

First, comic pages together with their associated panel annotations are collected and preprocessed. Next, training samples are generated automatically by selecting a target panel and constructing a natural language instruction requesting its localization. The corresponding target output consists of the page identifier and bounding box coordinates of the selected panel. This process eliminates the need for manually written instruction-response pairs and enables large-scale training using automatically generated supervision.

The resulting multimodal input, composed of multiple comic pages and the generated instruction, is then processed by Qwen3-VL and optimized using an instruction fine-tuning objective. Finally, the adapted model is evaluated on the proposed *Comics VQA* benchmark to determine whether the structural knowledge learned during panel localization transfers to downstream comic understanding tasks such as visual question answering and long-context reasoning.

3.4 Data Preparation

Training is performed using a large subset of the DCM dataset introduced as part of the CoMix benchmark [6], comprising 11,248 comic books and a total of 517,666 comic pages. The dataset is stored in the WebDataset format, where each sample consists of a comic page image together with a JSON metadata file containing the original image dimensions and annotations for comic elements, including panel bounding boxes.

During preprocessing, each sample is first validated and decoded before the required annotation categories are extracted. To reduce computational requirements while preserving the overall page layout, all comic pages are resized to a fixed resolution of 650×450 pixels. Since the panel annotations are defined in the coordinate system of the original image, all bounding boxes are transformed to the resized image coordinates according to

$$x' = x \cdot \frac{W'}{W}, \qquad y' = y \cdot \frac{H'}{H}, \tag{1}$$

where (W, H) and (W', H') denote the original and resized image dimensions, respectively. This transformation ensures that the supervision remains spatially aligned with the resized images presented to the model.

The resized images are subsequently processed by the Qwen3-VL visual encoder. Qwen3-VL employs a visual downsampling factor of 28, resulting in

$$N_{\text{img}} = \frac{H'}{28} \cdot \frac{W'}{28} \tag{2}$$

visual tokens for an image of dimensions (H', W').

Using the original image resolutions, the selected subset corresponds to approximately 1.3 billion visual tokens. After resizing all pages to 650×450 pixels, the total number of visual tokens is reduced to approximately 190 million, making large-scale training feasible on the available hardware. When multiple pages are provided as input, the total sequence length is given by the sum of the visual tokens from all input pages together with the textual tokens generated from the instruction prompt.

Although the dataset is organized at the comic-book level, processing complete comic books during training is computationally infeasible because of memory limitations

and the large number of pages per book. Instead, each comic is partitioned into non-overlapping sequences of consecutive pages. In our main experiments, each training sample consists of three consecutive pages, although additional experiments using two to five consecutive pages are also performed. This strategy preserves the natural narrative continuity of comics while maintaining an input length that is tractable for training large vision-language models.

3.5 Automatic Training Sample Generation

To enable large-scale comic-specific fine-tuning without requiring manually written instruction-response pairs, all training samples are generated automatically from existing panel annotations. These annotations are obtained using the Faster R-CNN panel detector proposed by Ren et al. [15], which provides the location of every panel on a comic page.

For each training sample, one of the input pages is selected uniformly at random, followed by the random selection of a panel from that page. Before generating the training instruction, all panels are sorted according to their approximate reading order, from top to bottom and from left to right. This ordering assigns a unique index to every panel, allowing the model to reference panels consistently using natural language descriptions.

A natural language instruction is then generated automatically by specifying the page number and the target panel index. An example instruction is shown below:

```
Give me the page number and bounding
box of panel 3 on page 2.
```

The corresponding target output is also generated automatically from the panel annotations and consists of a structured JSON response containing the page identifier together with the bounding box coordinates of the selected panel:

```
{"page": p, "bbox": [x1, y1, x2,
y2]}
```

This automatic generation process enables the creation of an effectively unlimited number of instruction-response pairs without manual annotation. Unlike conventional instruction tuning, where human annotators write task-specific prompts and responses, both the instruction and the target output are derived directly from the available panel annotations. Consequently, the model can be adapted to the comic domain using automatically generated supervision while minimizing the need for additional human labeling.

3.6 Model

The base model used throughout this work is Qwen3-VL, a large vision-language model capable of jointly processing visual and textual inputs within a unified transformer architecture. An overview of the model is shown in Figure 2. Given one or more comic pages together with a natural language instruction, the model generates a textual response. In our panel localization task, this response consists of a structured JSON object containing the predicted page number and the corresponding bounding box coordinates.

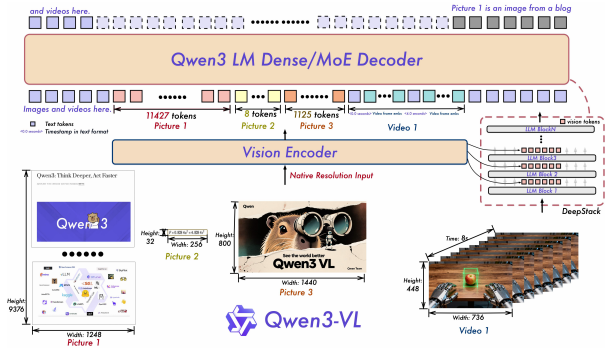


Fig. 2: Overview of the Qwen3-VL architecture.

Our experiments primarily use the 4B and 8B variants of Qwen3-VL. All models are loaded in `bfloat16` precision, and FlashAttention-2 is enabled whenever supported by the hardware. This optimization is particularly important for our setting because multi-page comic inputs produce long multimodal sequences. By reducing both memory consumption and the computational cost of the attention mechanism, FlashAttention-2 enables more efficient training and inference while supporting longer input contexts.

Directly fine-tuning all parameters of Qwen3-VL would require substantial computational resources. Therefore, we adopt Low-Rank Adaptation (LoRA), which keeps the pretrained model weights frozen while introducing a small number of trainable parameters. Given a pretrained weight matrix W, LoRA learns a low-rank update according to

$$W' = W + \Delta W, \qquad \Delta W = BA, \tag{3}$$

where W denotes the frozen pretrained weight matrix, and A and B are trainable low-rank matrices. In our implementation, LoRA adapters are inserted into the query, key, value, and output projection layers of the transformer attention blocks. We use a rank of $r = 16$, a scaling factor of $\alpha = 16$, and a dropout rate of 0.05.

This approach dramatically reduces the number of trainable parameters while preserving the knowledge contained in the pretrained model, making comic-specific adaptation computationally efficient without requiring full model fine-tuning.

3.7 Training Objective

Although the training samples are generated automatically, the model is optimized using the standard instruction fine-tuning objective employed by Qwen3-VL. Each training sample is formatted as an instruction-response pair. The instruction contains the input comic pages together with the automatically generated localization query, while the target response consists of a structured JSON object specifying the page number and bounding box of the target panel.

Let x_i denote the multimodal input, consisting of the comic page images and the corresponding textual instruction, and let $y_i = (y_{i,1}, \ldots, y_{i,T})$ denote the target assistant response. The model parameters θ are optimized using the standard autoregressive language modeling objective

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N}\sum_{t=1}^{T} m_{i,t} \log p_\theta(y_{i,t} \mid y_{i,<t}, x_i),$$

where $m_{i,t}$ is a binary mask indicating whether token $y_{i,t}$ belongs to the target response. Consequently, the loss is computed only over the target output tokens, while the input instruction, image tokens, and other conditioning information are excluded from the optimization objective.

3.8 Distributed Training

Training vision-language models on multi-page comic inputs is computationally demanding due to the large number of visual tokens processed by the model. To scale training across multiple GPUs, we considered two distributed training strategies: Distributed Data Parallelism (DDP) and Fully Sharded Data Parallelism (FSDP).

In DDP, each GPU maintains a complete copy of the model parameters and processes a different mini-batch. The local gradient computed on GPU k is

$$g_k = \nabla_\theta \mathcal{L}_k(\theta), \tag{4}$$

and the gradients from all GPUs are synchronized using an all-reduce operation:

$$g = \frac{1}{G} \sum_{k=1}^{G} g_k. \tag{5}$$

The synchronized gradient is then used to update the model parameters:

$$\theta_{t+1} = \theta_t - \eta g. \tag{6}$$

The effective batch size is therefore

$$B_{\text{eff}} = B_{\text{gpu}} \times G \times A, \tag{7}$$

where B_{gpu} is the per-GPU batch size, G is the number of GPUs, and A is the number of gradient accumulation steps.

In contrast, FSDP partitions the model parameters, gradients, and optimizer states across multiple GPUs. Instead of storing the complete parameter set $|\theta|$ on every device, each GPU stores approximately

$$\frac{|\theta|}{G}, \tag{8}$$

where $|\theta|$ denotes the total number of model parameters. During training, the required parameter shards are gathered for computation and redistributed afterward, reducing memory usage at the cost of additional communication overhead.

To determine the most suitable strategy, we performed a pilot comparison using the Qwen3-VL 2B model. The objective was to evaluate the computational trade-offs between DDP and FSDP rather than model performance. Since the final experiments use LoRA-based fine-tuning, where only a small subset of the parameters is updated, this pilot study was used to select the distributed training framework.

As shown in Table 1, DDP achieved the shortest runtime for the same batch size, requiring 608 seconds compared with 630–640 seconds for the FSDP configurations. Because LoRA significantly reduces memory requirements, the additional memory savings provided by FSDP were less beneficial than its communication overhead. Consequently, DDP was selected as the distributed training strategy for all subsequent experiments due to its lower runtime overhead and simpler implementation.

TABLE 1: Runtime comparison of DDP and FSDP.

Method	Batch	FlashAttn	Time (s)
FSDP	15	No	640
FSDP + Compile	15	No	630
FSDP + Compile	15	Yes	635
DDP + Compile	15	Yes	608
DDP + Compile	20	Yes	825

3.9 Optimization Details

Model optimization is performed using the fused AdamW optimizer with a learning rate of 10^{-4}. Gradient checkpointing is enabled to reduce activation memory consumption, which is particularly important when training with multi-page comic inputs. To improve training stability, the gradient norm is clipped to a maximum value of 1.0, and a warmup ratio of 0.03 is applied at the beginning of training. A constant learning-rate schedule is used throughout training.

Training is performed in `bfloat16` precision, with TensorFloat-32 (TF32) enabled whenever supported by the hardware, providing a good balance between computational efficiency, memory usage, and numerical stability. The maximum sequence length is set to 256k tokens to accommodate multiple comic pages together with the corresponding textual instruction. Model checkpoints are saved periodically throughout training. The final model consists of the pretrained Qwen3-VL weights, the learned LoRA adapters, and the corresponding processor required for inference.

4 Comics VQA Benchmark

4.1 Methodology

The objective of the proposed benchmark is to evaluate the comic understanding capabilities of modern vision-language models. Rather than focusing on a single task, the proposed benchmark is designed to assess multiple aspects of comic understanding through a unified natural language interface. In particular, it evaluates a model's ability to reason over visual and textual information, interpret long narrative contexts spanning multiple comic pages, and generate both free-form textual answers and structured outputs.

The benchmark construction process consists of four main stages. First, comic books are collected from publicly available datasets and additional scanned comic books. Second, human annotators manually create and validate benchmark questions covering different comic understanding capabilities. Third, the questions are organized into two complementary task types: visual question answering and panel localization. Finally, a collection of state-of-the-art vision-language models is evaluated on the benchmark to establish baseline performance and identify current limitations in comic understanding.

The following sections describe each stage of the benchmark construction process in detail.

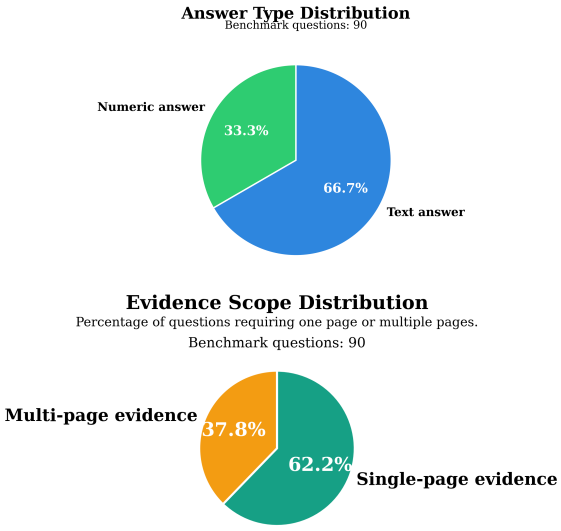


Fig. 3: Statistics of the visual question answering (VQA) portion of the *Comics VQA* benchmark. (a) Distribution of answer types. (b) Distribution of evidence scope, showing the proportion of questions requiring single-page and multi-page reasoning.

4.2 Benchmark Task Design

The *Comics VQA* benchmark is designed to evaluate the reasoning capabilities of modern vision-language models in the comic domain. It consists of a collection of manually curated question-answer pairs that require models to jointly interpret visual and textual information across comic panels and pages. The comic data is collected from multiple sources, including the CoMix dataset [6], which contains both Western-style comics and manga.

The benchmark comprises two complementary task types. The first consists of 90 visual question answering (VQA) examples, in which models are required to generate textual or numerical answers. These questions evaluate a broad range of comic understanding capabilities, including OCR and text understanding, counting, character identification, visual understanding, and narrative reasoning across multiple panels or pages. Some questions require extracting information from speech balloons or captions, whereas others require reasoning about visual relationships, character interactions, or the progression of events throughout the story.

The distribution of the VQA questions is summarized in Figure 3. Figure 3(a) shows the distribution of answer types. Most questions (66.7%) require textual answers, while the remaining 33.3% require numerical answers. This distribution reflects the emphasis of the benchmark on semantic reasoning while still including quantitative reasoning tasks such as counting.

Figure 3(b) illustrates the evidence scope required to answer each question. A total of 62.2% of the questions can be answered using information from a single comic page, whereas the remaining 37.8% require reasoning across multiple pages. These multi-page questions are designed to evaluate long-context understanding, requiring models to integrate visual and textual information distributed throughout the narrative.

The second task type consists of 8 panel localization questions, where models must identify the panel corresponding to a natural language query by returning both the page number and the panel bounding box. These tasks extend the benchmark beyond conventional question answering by evaluating a model's ability to ground textual descriptions in the structural layout of comic pages. Together, the two task types enable the evaluation of both free-form textual responses and structured outputs through a unified natural language interface.

Examples of the benchmark tasks are shown in Figures 4 and 5. Figure 4 illustrates a panel localization task, whereas Figure 5 presents a multi-page reasoning example requiring the integration of visual and textual information across multiple comic pages.

Each benchmark question is associated with either a single panel, multiple panels, or multiple comic pages, depending on the reasoning required. All question-answer pairs are manually created and validated by human annotators to ensure correctness, clarity, and consistency. By combining visual question answering with panel localization, *Comics VQA* provides a comprehensive benchmark for evaluating multimodal reasoning, long-context understanding, and structural comic understanding in modern vision-language models.

4.3 Annotation Process

All benchmark questions are manually created and validated by human annotators using a dedicated annotation tool developed for this project. The annotation interface allows annotators to browse comic pages, formulate questions, record the corresponding ground-truth answers, and assign the pages required to answer each question.

Annotators are instructed to design questions that require genuine multimodal reasoning rather than simple text extraction or object recognition. Particular emphasis is placed on questions involving long-context reasoning across multiple pages, character identification, visual understanding, OCR, and narrative comprehension. In addition, a small set of panel localization questions is created to evaluate the ability of vision-language models to ground natural language descriptions in the structural layout of comic pages.

Annotators followed a common set of annotation guidelines to ensure consistency in question formulation, answer formatting, and task difficulty across the benchmark. The guidelines also specify how to handle ambiguous cases and define the expected output format for different question types. The complete annotation guidelines are provided in Appendix A.3.

Each annotation is manually reviewed to ensure correctness, clarity, and consistency before being added to the benchmark. Examples of the annotation interface are provided in Appendix A.2.

4.4 Evaluation Protocol

To evaluate model performance on *Comics VQA*, different evaluation metrics are used depending on the expected answer type. All evaluated models receive the same input format, consisting of the comic pages, the benchmark question,



Fig. 4: Panel localization example. The model must return the page number and the bounding box of the panel corresponding to the query: *"Give me the page number and bounding box of the panel where Samson says, 'NOW I WISH I KNEW WHERE DAVID IS!' "*

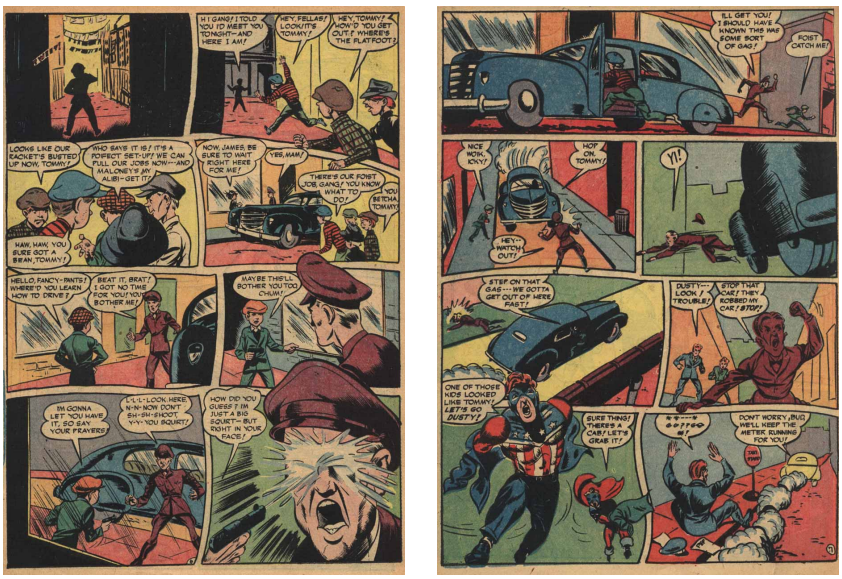


Fig. 5: Multi-page reasoning example. The model must integrate visual and textual information across multiple comic pages to answer the question: *"What is the name of the character who says, 'Stop that car! They robbed my car! STOP!'?"*

and a predefined system prompt. The complete prompt template is provided in Appendix A.1.

Numerical answers, including those containing units, are evaluated using exact matching, where both the numerical value and the unit must match the ground truth exactly. Date-based answers are also evaluated using exact matching.

For textual answers, we employ the Approximate Normalized Levenshtein Similarity (ANLS) metric. Before computing ANLS, both the predicted answer and the ground-truth answer are normalized by converting all characters to lowercase, removing punctuation, and eliminating articles such as "a", "an", and "the".

Given two strings s_1 and s_2, ANLS is defined as

$$\mathrm{ANLS}(s_1, s_2) = 1 - \frac{d(s_1, s_2)}{\max(|s_1|, |s_2|)}, \tag{9}$$

where $d(s_1, s_2)$ denotes the Levenshtein edit distance between the two strings and $|s|$ denotes the length of string s.

When multiple ground-truth answers are available, the final similarity score is computed as

$$\mathrm{ANLS}(p, G) = \max_{g \in G} \mathrm{ANLS}(p, g), \tag{10}$$

where p denotes the predicted answer and G the set of valid ground-truth answers.

A prediction is considered correct if

$$\mathrm{Correct}(p, G) = \begin{cases} 1, & \mathrm{ANLS}(p, G) \geq \tau, \\ 0, & \text{otherwise}, \end{cases} \tag{11}$$

where the threshold is set to $\tau = 0.9$. This relatively strict threshold ensures high precision when evaluating textual responses.

4.5 Baseline Models

To establish benchmark baselines, we evaluate both proprietary and open-source vision-language models. The proprietary models include ChatGPT and Gemini, while the open-source models are drawn from the Qwen3-VL family. All models are evaluated using the same benchmark questions, identical input formatting, and the same system prompt to ensure a fair comparison.

For each benchmark question, the model receives the required comic pages together with the corresponding natural language query and generates a prediction. The predicted answer is then compared with the ground-truth annotation using the evaluation protocol described in Section 4.4. We report the overall benchmark accuracy together with performance across the different task categories, allowing us to analyze the strengths and limitations of current vision-language models on comic understanding.

5 Work Plan

The project was organized into several stages, beginning with a review of the literature on comic understanding, vision-language models, domain adaptation, and benchmarking. This initial stage provided an overview of existing approaches and helped identify the research gaps addressed in this thesis.

The second stage focused on the development of the *Comics VQA* benchmark. This involved collecting comic books from existing datasets and additional sources, developing an annotation tool to support efficient dataset creation, and manually creating and validating benchmark questions. The resulting benchmark was designed to evaluate multiple comic understanding capabilities, including vi-

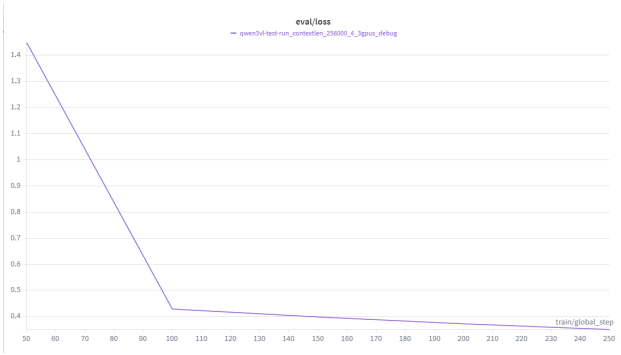


Fig. 6: Validation loss during panel localization fine-tuning.

sual question answering, long-context reasoning, OCR, and panel localization.

The third stage consisted of evaluating several state-of-the-art vision-language models on the proposed benchmark to establish zero-shot baselines and analyze their strengths and limitations in comic understanding.

Building on these results, the fourth stage investigated comic-specific domain adaptation through instruction fine-tuning. A large-scale training dataset was generated automatically using panel annotations, and the Qwen3-VL model was fine-tuned on a panel localization task to evaluate whether learning the structural organization of comic pages improves performance on downstream comic understanding tasks.

Finally, the benchmark results and the fine-tuning experiments were analyzed to evaluate the effectiveness of the proposed benchmark and to assess the impact of comic-specific adaptation on vision-language models.

Throughout the project, regular meetings were held with the supervisor to discuss progress, address technical challenges, and receive feedback. These meetings helped ensure that the project remained aligned with its objectives and progressed according to the planned schedule.

6 Results

6.1 Panel Localization Fine-Tuning

To adapt the model to the comic domain, we fine-tuned Qwen3-VL on the panel localization task described in Section **??**. Training was performed on the MareNostrum 5 supercomputer using the optimization settings described in Section 3.9.

For the 4B model, training was conducted on a single node equipped with four NVIDIA H100 GPUs. The training run lasted approximately 71 hours and processed roughly 33 million tokens. The 8B model was trained using four nodes (16 NVIDIA H100 GPUs in total), requiring approximately 48 hours and processing roughly 59 million tokens.

To evaluate whether the model successfully learned the proposed task, we first monitored the validation loss throughout training. As shown in Figure 6, the validation loss decreases steadily, indicating that the model successfully optimized the panel localization objective.

Although the validation loss provides evidence of successful optimization, it does not directly measure the model's localization performance. Therefore, we additionally evaluate three task-specific metrics:

TABLE 2: Performance on the panel localization validation task.

Model	FA (%)	JSON (%)	IoU Acc. (%)
4B Base	100.0	100.0	0.6
4B FT	100.0	100.0	36.5

- Presence of the `FINAL ANSWER` token.
- Valid JSON generation rate.
- Bounding-box localization accuracy.

Bounding-box localization accuracy is measured using an Intersection-over-Union (IoU) threshold of 0.5. Given a predicted bounding box $\hat{B}_i$ and the corresponding ground-truth bounding box B_i, the IoU is computed as

$$\mathrm{IoU}(\hat{B}_i, B_i) = \frac{|\hat{B}_i \cap B_i|}{|\hat{B}_i \cup B_i|}. \tag{12}$$

A prediction is considered correct if

$$\mathrm{IoU}(\hat{B}_i, B_i) \geq \tau, \tag{13}$$

where $\tau = 0.5$. The localization accuracy is then computed as

$$\mathrm{Acc}_{\mathrm{IoU}@\tau} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[\mathrm{IoU}(\hat{B}_i, B_i) \geq \tau\right]. \tag{14}$$

Table 2 summarizes the performance on the panel localization validation set.

Here, FA denotes the percentage of predictions containing the required `FINAL ANSWER` token, JSON denotes the percentage of syntactically valid JSON outputs, and IoU Acc. corresponds to the bounding-box localization accuracy.

The results demonstrate that the proposed fine-tuning strategy successfully teaches the model to perform the panel localization task. Although both the pretrained and fine-tuned models consistently generate valid JSON responses in the expected format, the localization accuracy improves substantially after fine-tuning, increasing from 0.6% to 36.5%. This indicates that the model effectively learns to associate natural language queries with the structural organization of comic pages.

Having established that the model successfully learns the panel localization task, the next section investigates whether this knowledge transfers to the broader comic understanding capabilities evaluated by the proposed *Comics VQA* benchmark.

6.2 Transfer to Comics VQA

Having established that the model successfully learned the panel localization task, we now investigate whether this knowledge transfers to broader comic understanding capabilities. To answer this question, we evaluate the fine-tuned

TABLE 3: Zero-shot accuracy on the proposed *Comics VQA* benchmark.

Model	Accuracy
Gemini-3-flash-preview	56.7%
Qwen3-VL-4B	5.6%
Qwen3-VL-8B	21.1%
Qwen3-VL-32B	17.8%

TABLE 4: Benchmark accuracy before and after panel localization fine-tuning.

Model	Base	Fine-tuned
Qwen3-VL-4B	5.6%	2.2%
Qwen3-VL-8B	21.1%	16.7%

models on the proposed *Comics VQA* benchmark and compare their performance with the corresponding zero-shot baselines.

As an initial reference, Table 3 presents the zero-shot performance of several modern vision-language models on *Comics VQA*. Since no comic-specific adaptation is applied, these results reflect the out-of-the-box capabilities of general-purpose vision-language models.

The results show that comic understanding remains a challenging problem for current vision-language models. Gemini-3-flash-preview achieves the highest accuracy, substantially outperforming the open-source Qwen3-VL models. Interestingly, within the Qwen family, the 8B model outperforms the larger 32B variant, suggesting that increased model size alone does not necessarily lead to better comic understanding.

Next, we evaluate whether panel localization fine-tuning improves benchmark performance. Table 4 compares the zero-shot and fine-tuned models.

Contrary to our hypothesis, fine-tuning on the panel localization task does not improve performance on *Comics VQA*. Both the 4B and 8B models achieve lower benchmark accuracy after fine-tuning than their corresponding zero-shot baselines. These results indicate that learning to localize comic panels does not automatically transfer to the broader comic understanding capabilities evaluated by the benchmark.

To better understand this behavior, we manually inspected the generated predictions. During this analysis, we observed that many responses contained the correct answer somewhere in the generated text but did not follow the required output format. The benchmark expects every prediction to be returned using the format

```
FINAL ANSWER: ...
```

However, the models frequently omitted the required prefix or embedded the correct answer within intermediate reasoning text. In such cases, the prediction was considered incorrect even though the generated output contained the correct information.

To quantify this effect, we introduce two additional metrics:

- **Answer in Text Rate**: the percentage of predictions in which the ground-truth answer appears anywhere in the generated output.

TABLE 5: Analysis of answer occurrence within generated responses.

Model	Acc.	In Text	First 5W
4B Base	0.056	0.356	0.289
4B FT	0.022	0.300	0.211
8B Base	0.211	0.525	0.448
8B FT	0.167	0.389	0.293

- **Answer Length $\leq$ 5 Words Rate**: the percentage of predictions containing the ground-truth answer where the generated response consists of at most five words.

The results are summarized in Table 5.

The additional metrics show that the benchmark accuracy underestimates the amount of relevant information contained in the generated responses, as the correct answer is often present somewhere in the output. Nevertheless, both the benchmark accuracy and the two auxiliary metrics decrease after fine-tuning. Therefore, the observed performance degradation cannot be explained solely by formatting issues. For a more detailed comparison of the evaluation outcomes, the corresponding confusion matrices are provided in Appendix A.4, while the per-question results for all evaluated models are presented in Appendix A.5.

Overall, these results indicate that although the model successfully learns the panel localization task, the acquired knowledge does not transfer effectively to the broader comic understanding capabilities measured by *Comics VQA*. This suggests that panel localization alone is not a sufficiently rich adaptation objective for improving general comic reasoning, motivating the exploration of more informative comic-specific training objectives in future work. The complete model outputs, including both processed predictions and raw generated responses, are available in Appendix A.6.

6.3 Discussion

The results presented in this chapter provide several insights into comic-specific adaptation of vision-language models. First, the panel localization experiments demonstrate that Qwen3-VL can successfully learn the proposed fine-tuning task. The substantial improvement in bounding-box localization accuracy confirms that the model acquires knowledge about the structural organization of comic pages, including panel boundaries and their spatial arrangement.

However, this improvement does not translate into better performance on the proposed *Comics VQA* benchmark. Despite successfully learning the training objective, both the 4B and 8B models achieve lower benchmark accuracy after fine-tuning than their corresponding zero-shot baselines. Furthermore, the additional answer-in-text analysis shows that this degradation cannot be explained solely by differences in output formatting.

These findings suggest that learning comic page structure alone is insufficient to improve general comic understanding. Although panel localization encourages the model to reason about page layout and spatial relationships, most benchmark questions additionally require OCR, character identification, long-context reasoning, and integration of visual and textual information across multiple panels and

pages. Consequently, the learned representations appear to be highly specialized to the localization task and do not generalize effectively to the broader reasoning capabilities evaluated by *Comics VQA*.

One possible explanation is that the proposed fine-tuning objective does not provide a sufficiently rich learning signal for developing transferable comic representations. The training task focuses exclusively on predicting panel locations, whereas the benchmark evaluates a much broader set of multimodal reasoning abilities. More diverse adaptation objectives that jointly exploit page structure, textual content, and narrative context may therefore be required to improve downstream comic understanding.

Although the proposed adaptation strategy did not improve benchmark performance, the results remain valuable. They demonstrate that automatically generated supervision can successfully train a vision-language model to perform comic-specific localization tasks while also highlighting the difficulty of transferring this knowledge to more general comic understanding. These findings provide a useful starting point for future research on comic-domain adaptation and suggest several promising directions, including richer instruction-tuning tasks, multi-task learning, and self-supervised objectives that incorporate both structural and semantic information.

7 Conclusions

This thesis addressed two complementary challenges in comic understanding: the lack of a comprehensive benchmark for evaluating modern vision-language models and the adaptation of existing models to the comic domain through instruction fine-tuning.

The first contribution of this work is the introduction of *Comics VQA*, a benchmark designed to evaluate a broad range of comic understanding capabilities through a unified natural language interface. Unlike existing comic datasets that primarily target individual tasks, *Comics VQA* combines visual question answering and panel localization, requiring models to perform OCR, visual understanding, long-context reasoning across multiple pages, and structural understanding of comic layouts. The zero-shot evaluation demonstrates that comic understanding remains a challenging problem for current vision-language models. Although proprietary models such as Gemini achieve promising performance, a substantial gap remains for open-source models, highlighting the need for improved comic-specific adaptation methods.

The second contribution of this work is an instruction fine-tuning strategy based on automatic panel localization. Instead of relying on manually annotated training data, the proposed approach automatically generates training instructions and supervision from existing panel annotations, enabling large-scale adaptation with minimal human effort. Experimental results show that the fine-tuned model successfully learns the panel localization task, achieving a substantial improvement in bounding-box localization accuracy while maintaining valid structured outputs.

However, the main research hypothesis was not supported by the experimental results. Although the model learned the structural organization of comic pages through panel localization, this knowledge did not transfer to the broader comic understanding capabilities evaluated by *Comics VQA*. Fine-tuning on the panel localization task resulted in lower benchmark performance than the corresponding zero-shot models, indicating that learning comic page structure alone is insufficient to improve general comic reasoning. This finding suggests that successful optimization of a domain-specific training objective does not necessarily lead to transferable representations for more complex multimodal reasoning tasks.

These findings provide valuable insights for future research. They indicate that richer adaptation objectives incorporating narrative understanding, dialogue grounding, character tracking, and long-context reasoning may be necessary to improve general comic understanding. Likewise, multi-task instruction tuning or representation learning approaches that jointly exploit both the structural and semantic characteristics of comics represent promising research directions.

Overall, this thesis contributes a new benchmark for comic understanding, an automatic pipeline for comic-specific instruction fine-tuning, and an empirical analysis of knowledge transfer from panel localization to visual question answering. We hope that both the benchmark and the experimental findings, including the negative transfer results, will provide a useful foundation for future research on adapting vision-language models to the comic domain.

Acknowledgments

We thank all annotators for their contributions to the annotation process.

References

[1] R. Bommasani, D. A. Hudson, E. Adeli, and R. Altman, “On the opportunities and risks of foundation models,” 2022. [Online]. Available: https://arxiv.org/abs/2108.07258

[2] C. Guérin, C. Rigaud, A. Mercier, F. Ammar-Boudjelal, K. Bertet, A. Bouju, J.-C. Burie, G. Louis, J.-M. Ogier, and A. Revel, “ebdtheque: A representative database of comics,” in *International Conference on Document Analysis and Recognition (ICDAR)*, 2013, pp. 1145–1149.

[3] N. V. Nguyen, C. Rigaud, and J.-C. Burie, “Digital comics image indexing based on deep learning,” *Journal of Imaging*, vol. 4, no. 7, p. 89, 2018.

[4] A. Fujimoto, T. Ogawa, K. Yamamoto, Y. Matsui, T. Yamasaki, and K. Aizawa, “Manga109 dataset and creation of metadata,” in *Proceedings of the 1st International Workshop on Comics Analysis, Processing and Understanding (MANPU)*, 2016, pp. 1–5.

[5] R. Sachdeva and A. Zisserman, “The manga whisperer: Automatically generating transcriptions for comics,” *arXiv preprint arXiv:2401.10224*, 2024.

[6] E. Vivoli, M. Bertini, and D. Karatzas, “Comix: A comprehensive benchmark for multi-task comic understanding,” *arXiv preprint arXiv:2407.03550*, 2024.

[7] E. Vivoli, J. L. Baeza, E. V. Llobet, and D. Karatzas, "Multimodal transformer for comics text-cloze," 2024. [Online]. Available: https://arxiv.org/abs/2403.03719

[8] B. B. Topal, D. Yuret, and T. M. Sezgin, "Domain-adaptive self-supervised pre-training for face body detection in drawings," 2023. [Online]. Available: https://arxiv.org/abs/2211.10641

[9] G. Soykan, D. Yuret, and T. M. Sezgin, "Identity-aware semi-supervised learning for comic character re-identification," 2023. [Online]. Available: https://arxiv.org/abs/2308.09096

[10] J. Baek, K. Egashira, S. Onohara, A. Miyai, Y. Imajuku, H. Ikuta, and K. Aizawa, "Mangavqa and mangalmm: A benchmark and specialized model for multimodal manga understanding," in *Findings of the Association for Computational Linguistics: EACL*, 2026.

[11] E. Vivoli, A. Llabrés, M. A. Souibgui, M. Bertini, E. L. Valveny, and D. Karatzas, "Comicspap: Understanding comic strips by picking the correct panel," in *Document Analysis and Recognition – ICDAR 2025*, 2025, pp. 337–350.

[12] A. Baranwal, M. Kataria, N. Agrawal, S. Vyas, and Y. S. Rawat, "Re:verse – can your vlm read a manga?" in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCV-W)*, 2025.

[13] Y. Ryan, R. Y. Tan, K. T. W. Choo, and R. K.-W. Lee, "Humor in pixels: Benchmarking large multimodal models understanding of online comics," *arXiv preprint arXiv:2509.12248*, 2025.

[14] Y. Chen, Y. Ren, and S. Süsstrunk, "Zooming into comics: Region-aware rl improves fine-grained comic understanding in vision-language models," *arXiv preprint arXiv:2511.06490*, 2025.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2015, pp. 91–99.

A APPENDIX

A.1 Prompt

The following prompt template was used to evaluate the models in the Comics VQA benchmark:

System Prompt:

ACT as an expert in Comics Visual Question Answering (ComicsVQA) system. ANALYZE the provided images to extract precise information.

Mandatory Response Rules:

1. **Source adherence:** If the question is unanswerable from the document, respond only with "Unknown".
2. **List formatting:** List multiple answers in order of appearance, separated by a comma and a single space (e.g., "Answer A, Answer B"). Do not use "and".
3. **Numbers and units:**
   - Convert units to their standardized abbreviation (e.g., use "kg" instead of "kilograms", "m" instead of "meters").
   - Place a single space between the number and the unit (e.g., "50 kg", "10 USD").
4. **Percentages:** Attach the % symbol directly to the number with no space (e.g., "50%", not "50 %").
5. **Date formatting:** Convert all dates to the YYYY-MM-DD format (e.g., "Jan 1st 24" → "2024-01-01").
6. **Decimal formatting:** Use a period as the decimal separator (e.g., "3.14", not "3,14").
7. **Thousands separator:** Do not use commas as thousands separators (e.g., "1000", not "1,000").
8. **No filler:** Output only the result. Do not use phrases such as "The answer is...".

Reasoning Protocol:

1. Perform exhaustive step-by-step reasoning to locate and verify the data.
2. Verify whether the data contains a date, number, or unit.
3. Transform the extracted data step by step to match the mandatory response rules.

Output Format:

After the analysis, the model must provide the final result in the following format:

```
FINAL ANSWER: [Your exact formatted answer]
```

The content inside the final answer must strictly follow the mandatory response rules.

A.2 Annotation Tool Interface

Figure 7 illustrates the annotation interface developed for the creation of the *Comics VQA* benchmark. The tool enables human annotators to inspect comic pages, create question–answer pairs, and validate annotations efficiently.

A.3 Annotation Guidelines

To ensure consistency across annotators, the following guidelines were provided during the creation of the *Comics VQA* benchmark.

1. **Question design.** Questions should require genuine multimodal reasoning whenever possible, rather than simple OCR or object recognition. Annotators were encouraged to create questions involving character identification, visual understanding, OCR, long-context reasoning across multiple panels or pages, and narrative comprehension.

2. **Answer format.** Answers should be concise whenever possible. The expected answer is typically a short textual phrase (e.g., a character name, object, or location) or a numerical value. Sentence-length answers should only be used when necessary.

3. **Multiple valid answers.** If a question admits more than one correct answer, all valid answers should be annotated. For example, if a character is referred to by both a full name and a shortened name within the comic, both forms should be included as acceptable answers.

4. **Question difficulty.** Annotators were encouraged to create challenging questions that require integrating information from multiple panels or multiple comic pages whenever possible. The goal was to evaluate reasoning capabilities rather than isolated visual recognition.

5. **Model verification.** The annotation tool includes an option to evaluate newly created questions using the Gemini model. Annotators were encouraged to retain questions that the model answered incorrectly, as these questions are generally more challenging and provide greater discriminative power for benchmark evaluation. Although this verification step was not applied to every question, it was used for a substantial portion of the benchmark.

6. **Avoid ambiguous questions.** Questions with subjective or ambiguous answers should be avoided. For example, questions requiring color descriptions such as "light green" or "dark blue", or other subjective interpretations, should not be included unless the answer is unambiguous.

7. **Unique answer.** Every question should have a clearly defined ground-truth answer that can be determined solely from the provided comic pages. Questions admitting multiple interpretations without a unique correct answer should be avoided.

8. **Consistency.** Questions and answers should follow a consistent style and formatting throughout the benchmark. Annotators were instructed to verify spelling, character names, and numerical values before submitting each annotation.

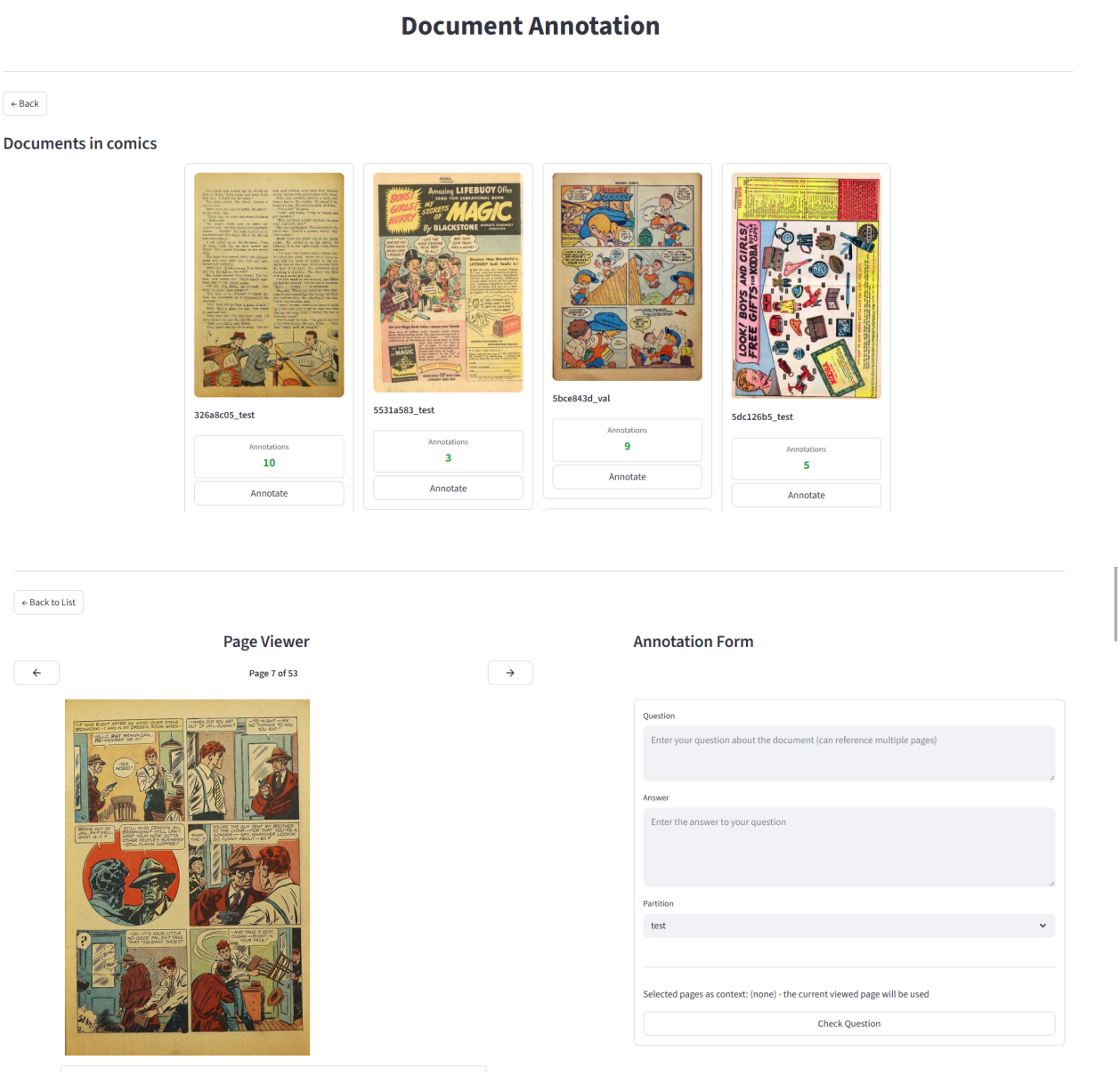


Fig. 7: Examples of the annotation interface developed for creating the *Comics VQA* benchmark. The tool allows human annotators to inspect comic pages and efficiently generate or validate question-answer pairs.

A.4 Confusion Matrices

This appendix presents the confusion matrices corresponding to the three evaluation settings discussed in Chapter **??**. The first matrix is computed using the standard benchmark evaluation. The second considers a prediction to be correct whenever the ground-truth answer appears anywhere in the generated response (Answer in Text). The third applies the same criterion while additionally requiring the generated response to contain at most five words (Answer in Text, ≤ 5 Words).

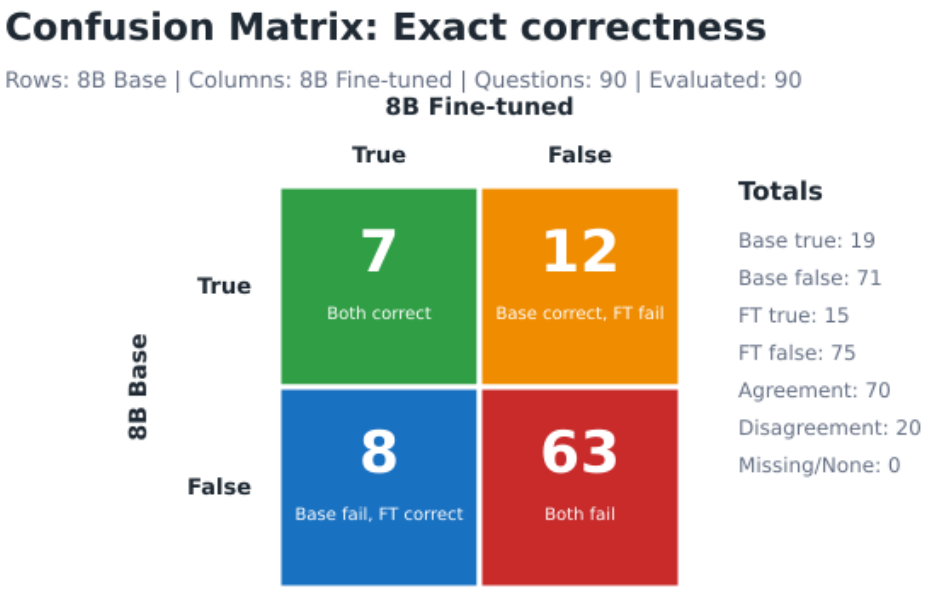


Fig. 8: Confusion matrix using the standard Comics VQA evaluation protocol.

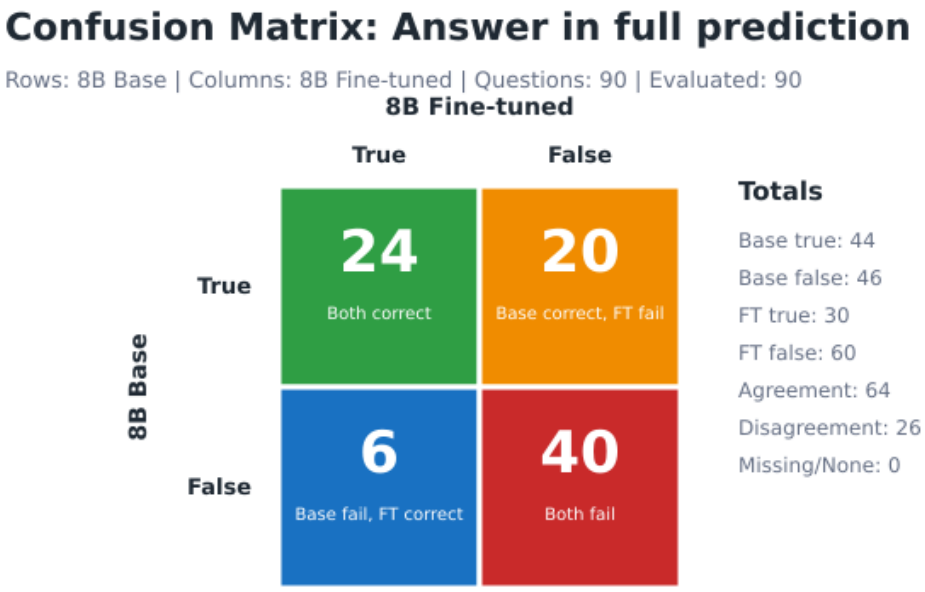


Fig. 9: Confusion matrix using the Answer in Text evaluation protocol.

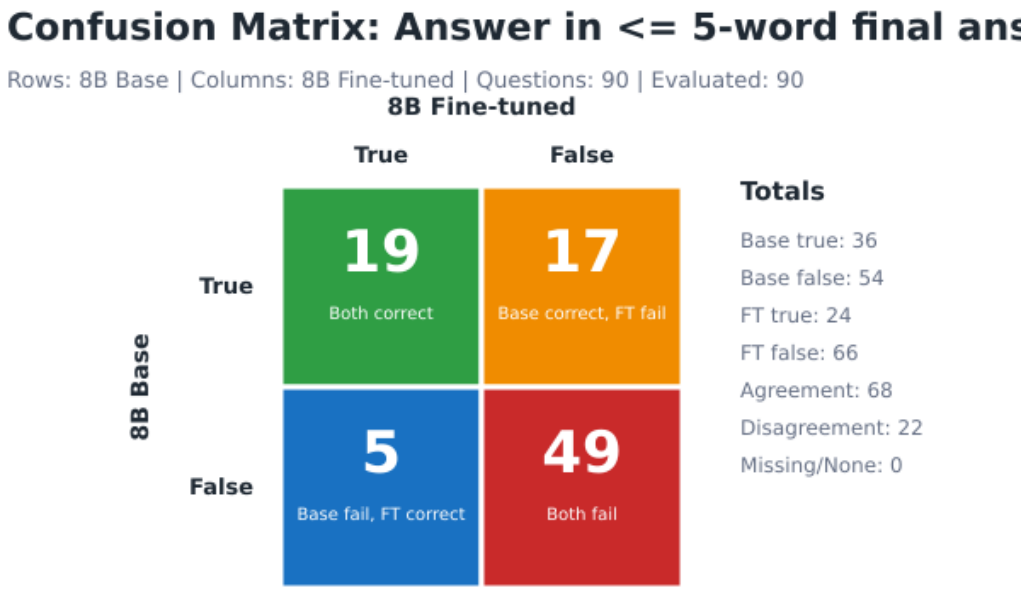


Fig. 10: Confusion matrix using the Answer in Text ($\leq$ 5 Words) evaluation protocol.

A.5 Per-Question Benchmark Results

This appendix provides a per-question comparison of model performance under the three evaluation settings. Each matrix summarizes the predictions for every benchmark question, making it possible to identify which questions remain challenging and which benefit from the relaxed evaluation criteria.

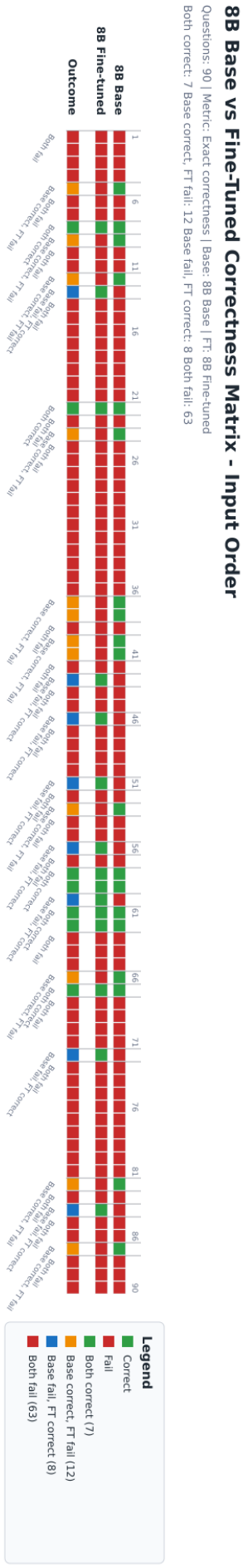


Fig. 11: Per-question benchmark results using the standard Comics VQA evaluation protocol. Each row corresponds to a benchmark question and each column corresponds to an evaluated model.

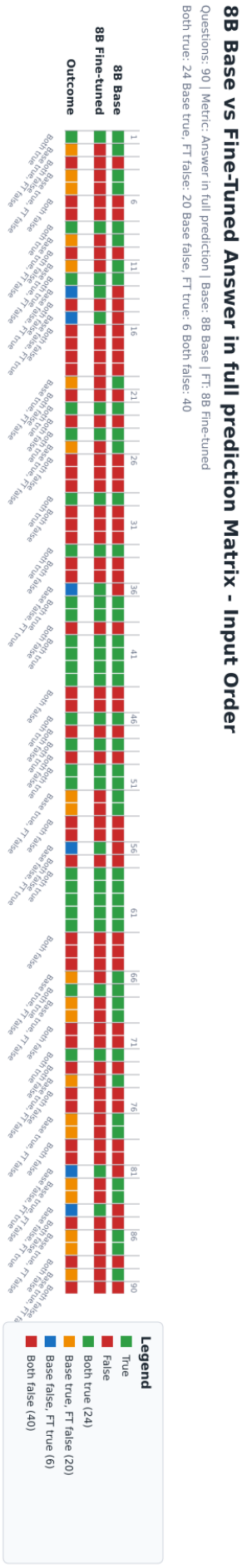


Fig. 12: Per-question benchmark results using the Answer in Text evaluation protocol, where a prediction is considered correct if the ground-truth answer appears anywhere in the generated response.

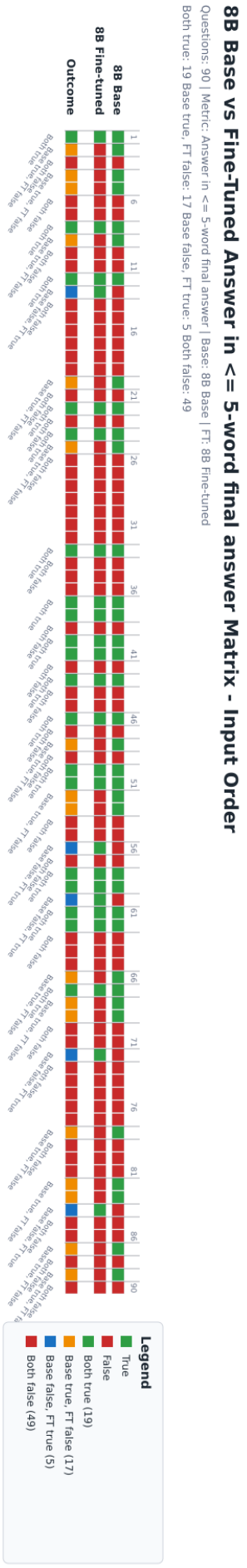


Fig. 13: Per-question benchmark results using the Answer in Text ($\leq$ 5 Words) evaluation protocol.

A.6 Complete Model Predictions

The complete evaluation results, including the ground-truth answer (`answer`), processed predictions from the base and fine-tuned models (`base_prediction`, `ft_prediction`), and the corresponding raw model outputs (`base_raw_prediction`, `ft_raw_prediction`), are available in the supplementary repository:

https://github.com/amritpal200/ComicsLib/blob/comicslib/answers