Multimodal Cloze in Comics with Vision-Language Models
PanelPioneers

Anushree Udhayakumar (au11), Sonia Navale(snavale2), Harshvardhan Sekar(sekar6), Prisha Singhania(pds4)

1. Introduction

Comics are a unique narrative medium that convey stories through the interaction of sequential images and text. Unlike purely textual narratives, much of the storytelling in comics occurs implicitly, relying on the reader's ability to infer events, actions, and causal relationships that are not explicitly shown. This interpretive process is central to narrative comprehension and presents a compelling challenge for artificial intelligence systems.

Recent advances in multimodal machine learning, particularly vision-language models (VLMs), have shown promise in jointly reasoning over visual and textual inputs. However, understanding narratives that rely on implicit reasoning remains an open research problem. Comics provide an ideal testbed for this challenge due to their reliance on inference across panels rather than explicit depiction. This project explores whether modern multimodal models can approximate this human-like narrative reasoning by predicting missing content in comic sequences

2. Problem Statement

A defining feature of comics, as illustrated in Fig 1, is the **gutter**, the space between **panels** (which contains image and text) where readers mentally fill in missing actions or events. This cognitive process is known as **closure**, defined as the human ability to construct a continuous narrative from discrete visual fragments (McCloud, 1993). Closure allows readers to infer motion, causality, dialogue progression, and scene changes even when these are not explicitly depicted.

Fig 1: Comic Panel Structure in the COMICS Dataset. A sequence of comic panels illustrating the image, text, panel, and gutter, which together define the narrative structure used in our analysis.

From a machine learning perspective, modeling closure requires joint reasoning over visual context, textual dialogue, and temporal sequence. Prior work has demonstrated that traditional models struggle with this task, even when both image and text modalities are available (Iyyer et al., 2017). The core problem addressed in this project is therefore:

Can modern multimodal VLMs perform narrative closure in comics by accurately predicting missing dialogue and visual content between panels?

Solving this problem has broader implications for narrative understanding, multimodal reasoning, and creative AI systems capable of generating coherent story continuations.

3. Existing Approaches

Early computational approaches to comics understanding relied on separate encodings of text and images, typically combining convolutional neural networks (CNNs) for visual features with recurrent models for text (Iyyer et al., 2017). While these methods established foundational benchmarks, they suffered from limited cross-modal interaction and weak narrative reasoning capabilities.

More recently, large VLMs have demonstrated strong performance on tasks requiring joint image-text reasoning, including visual question answering and instruction following (Liu et al., 2023). Surveys of the field suggest that these models represent a significant step forward for comics understanding, though their ability to perform narrative inference and closure remains under-explored (Mittal & Singh, 2024).

In parallel, generative image models such as Stable Diffusion have enabled high-quality text-to-image synthesis, making it possible to visualize predicted narrative continuations (Rombach et al., 2022). However, few studies have combined vision-language reasoning with image generation to explicitly model closure in comics. This project addresses this gap by integrating state-of-the-art multimodal reasoning with generative image synthesis

4. Data Source and description

The primary dataset used in this project is the **COMICS dataset**, introduced by Iyyer et al. (2017). The dataset, as shown in Fig 2, is derived from public-domain comic books hosted by the Digital Comic Museum and contains approximately **1.2 million comic panels**, making it one of the largest multimodal narrative datasets available. The dataset consists of scanned pages from multiple comic books, which were already segmented into individual panels by the original authors. Each panel is provided as a separate image and is accompanied by a corresponding CSV file containing the extracted dialogue or narration text. In addition to the panel images and dialogue, the dataset also includes metadata such as panel coordinates and speech balloon locations, which were not used in this project.

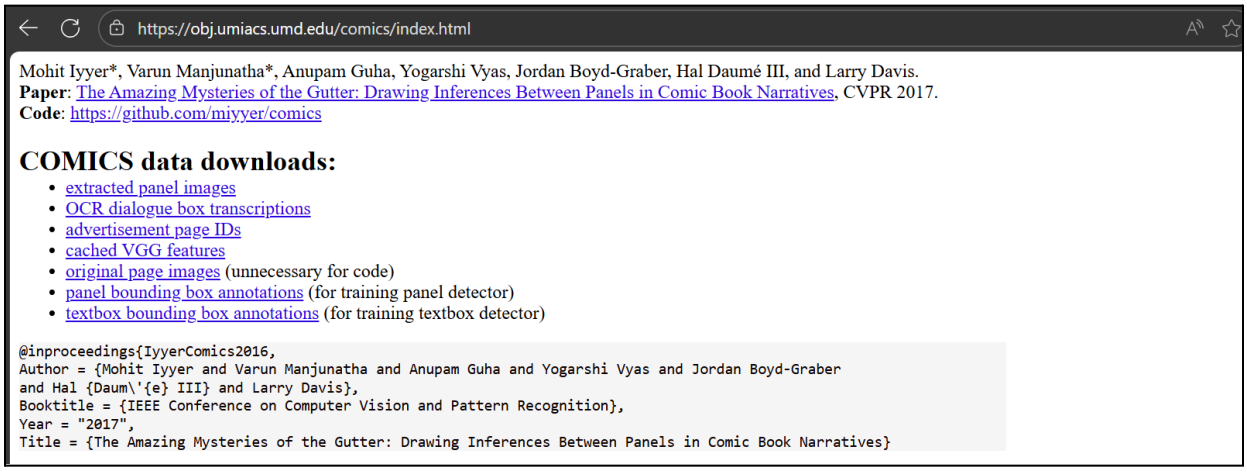


Fig 2: **Snapshot of the COMICS Dataset Resources.** Screenshot of the official COMICS dataset webpage showing available downloads, including extracted panel images, OCR dialogue transcriptions, and annotations used in this project.

5. Proposed Methodology and Expectations

5.1 Summary of Methodology

The task of modeling closure in comics is framed as a **multimodal cloze problem**, where a sequence of comic panels is provided and the model is required to infer the missing narrative content of the subsequent panel, including both dialogue and visual elements. This formulation requires reasoning over temporal continuity, character consistency, and implied actions-key aspects of narrative closure as defined in comics theory (McCloud, 1993).

The complete end-to-end pipeline starts from extracting panels from the entire comic book page, extract dialogue, have the model learn the story and predict the next panel, and use that to generate an image (described in detail in the project proposal). In brief, the approach consists of (1) finding panels from full page images, extract text, (2) multimodal narrative inference using VLMs, and (3) generative visualization of the inferred content using a text-to-image diffusion model.

5.2 Expectations and Hypotheses

Our primary expectation is that modern VLMs will outperform earlier CNN–LSTM-based approaches in capturing narrative continuity and implicit story progression. By jointly reasoning over visual and textual context, these models are expected to generate dialogue that is more coherent with preceding panels and better aligned with the underlying storyline. We further hypothesize that the predicted visual content will be semantically consistent with prior panels, even in cases where multiple plausible continuations exist.

Finally, we expect that instruction-tuned and interleaved multimodal representations will be particularly effective at capturing the implicit cues required for closure, supporting higher-level temporal and semantic reasoning. These expectations align with recent findings emphasizing the importance of tightly integrated image-text representations for complex narrative understanding tasks (Mittal & Singh, 2024).

6. Rationale for Methodological Changes Based on Experiments

6.1 Pivots in Data Pre-processing

We initially tried to process the raw comic page images ourselves using YOLOv8 to detect panel boundaries, CRAFT to find text regions, and EasyOCR to extract dialogue. However, this approach did not work well in practice. YOLOv8 often failed to correctly capture the true corners of comic panels and instead detected incorrect or incomplete regions. As a result, many panel images were poorly cropped or misaligned.

In addition, CRAFT and EasyOCR did not capture all the dialogue text. Some speech balloons were missed, and others were only partially extracted, leading to missing or incorrect dialogue. Overall, the quality of the extracted panels and text was clearly worse than the data prepared by the original authors of the COMICS dataset.

Since the main goal of this project is to evaluate whether modern VLMs perform better than earlier methods, it was important to use high-quality and reliable input data. Using noisy or incomplete data would make it unclear whether poor results were caused by model limitations or data errors. Therefore, we decided to use the **author-provided panel images and extracted dialogue**, which ensured well-aligned inputs and allowed us to focus on evaluating the model's abilities.

6.2 Model Selection Adjustment

We initially began by trying to implement lightweight and mid-scale vision-language models on our local machines. However, due to the low computational capacity of our systems, these attempts were largely unsuccessful. Although we were able to download some of the models, running them reliably and debugging issues related to memory, dependencies, and inference took a significant amount of time and did not lead to meaningful results. From a theoretical standpoint, these smaller models were also expected to have limited capacity for capturing long-range narrative context across multiple panels.

We next considered **OpenFlamingo**, as originally proposed. However, OpenFlamingo is primarily designed for **single-image vision-language tasks** and does not provide native architectural or training support for **multi-panel, narrative-level reasoning**, which is essential for comic cloze prediction. For this reason, we decided not to proceed with OpenFlamingo and instead focused our efforts on LLaVA-based models that better matched the requirements of our task.

7. Methodology

7.1 Dataset Overview

To train models to understand comic narratives, we reorganized the COMICS dataset into structured panel sequences. Instead of treating panels independently, we grouped them into short narrative units that preserve story flow and context.

Component	Count
Total panels in COMICS dataset	~1,200,000
Training sequences constructed	249,576
Test sequences constructed	54,530
Context panels per sequence	5
Target panels per sequence	1 (Panel 6)
Image resolution (max)	672 × 672 pixels

7.2 Sequence Construction

Comics are inherently sequential-each panel builds on what came before. To model this narrative structure, we construct **panel sequences** so the model can reason over story progression rather than isolated images.

Each example is built as a **6-panel sequence**:

- **Panels 1–5** serve as *context*, including both images and dialogue
- **Panel 6** is the *target*, which the model must predict
- The task is to generate a scene description for the target panel using only the prior context (see Section 7.3)

Sequences were constructed by sliding a 6-panel window across each comic book's panel sequence, ensuring that context panels and target panels belong to the same narrative thread. Sequences spanning page boundaries were included to capture cross-page narrative continuity.

7.3 Ground Truth Label Generation with Gemini 2.5 Flash

A critical challenge in this task is defining appropriate ground truth for scene prediction. Since the original COMICS dataset provides only OCR-extracted dialogue (which is often noisy or absent), we needed high-quality scene descriptions as training targets. We employed Google's Gemini 2.5 Flash model via the Vertex AI Batch API to generate these descriptions at scale.

Two Labeling Paradigms

We explored two distinct approaches to ground truth generation, each with different implications for training and evaluation:

Paradigm 1: Scene Descriptions (DESCP) - Gemini is shown all 6 panels including Panel 6 and asked to describe what it sees in Panel 6. This provides accurate visual descriptions but creates a train-test mismatch since the model during inference cannot see Panel 6.

Paradigm 2: Scene Predictions (PRED) - Gemini is shown only Panels 1-5 and asked to predict what will happen in Panel 6. This aligns training and evaluation conditions but may produce less accurate descriptions since Gemini is also predicting.

Aspect	DESCP Paradigm	PRED Paradigm
Gemini sees	All 6 panels	Only panels 1-5
Label type	Visual description	Narrative prediction
Train-eval alignment	Misaligned	Aligned
Expected accuracy	Higher ceiling	Better generalization

Batch Processing Infrastructure

Ground truth generation was performed using Google Cloud's Vertex AI Batch API with the following configuration:

- **Model:** gemini-2.5-flash-001
- **Processing:** 8 shards of ~35,000 sequences each
- **Total sequences:** 249,576 training + 54,530 test
- **Cost:** ~$30 USD for complete training set
- **Processing time:** ~4-6 hours per shard

Images were uploaded to Google Cloud Storage and referenced via GCS URIs in the batch request. The Batch API's 50% cost reduction compared to online inference made large-scale label generation economically feasible.

8. Model Architecture and Selection

8.1 Vision-Language Model Selection

Modern vision-language models process both visual and textual inputs within a fixed context window, which limits how much information can be provided to the model at once. Since our task requires reasoning over multiple comic panels-each containing an image and associated dialogue-it was necessary to analyze context window requirements to ensure that the full narrative context could be provided without truncation. This analysis directly informed our choice of model architecture.For the 5-panel input configuration, we estimated token requirements as follows:

- 5 images × ~576 tokens/image = ~2,880 visual tokens
- 5 dialogue transcriptions × ~100 tokens = ~500 text tokens
- Prompt template + generation buffer = ~600 tokens
- **Total requirement: ~4,000 tokens**

LLaVA-OneVision's 32K context window provides ample headroom for this configuration, whereas earlier LLaVA-1.5 variants (4K context) would require truncation.

After evaluating multiple VLM architectures (detailed in Section 6.1), we selected **LLaVA-OneVision-Qwen2-7B** as our primary model. This decision was driven by several key requirements:

Requirement	LLaVA-OneVision	LLaVA-1.5
Context window	32,768 tokens	4,096 tokens
Multi-image support	Native	Limited
Base LLM	Qwen2-7B	Vicuna-7B
Instruction tuning	Strong	Moderate

8.2 Model Architecture Details

Component	Specification
Model ID	llava-hf/llava-onevision-qwen2-7b-ov-hf
Vision encoder	SigLIP (400M parameters)
Language model	Qwen2-7B (7B parameters)
Total parameters	~8 billion
Vision-language projector	MLP (2-layer)
Image tokens per image	~576

9. Training Methodology

This section describes the training setup used to fine-tune the vision-language model. We outline the LoRA-based parameter-efficient fine-tuning strategy, key training hyperparameters, and the label masking approach used for causal language modeling. We also summarize the computational infrastructure and resources used to support large-scale training.

9.1 LoRA Fine-Tuning Configuration

We employed **Low-Rank Adaptation (LoRA)** for parameter-efficient fine-tuning to reduce memory and compute requirements while maintaining model performance. LoRA works by **freezing the original model weights** and injecting small, trainable **low-rank matrices** into selected transformer layers (such as attention projections). During training, only these added parameters are updated, which significantly lowers GPU memory usage and speeds up training while preserving the model's pre-trained knowledge.

Parameter	Value
LoRA rank (r)	16
LoRA alpha	32
LoRA dropout	0.05
Target modules	q_proj, k_proj, v_proj, o_proj
Trainable parameters	43,245,568 (0.54%)
Frozen parameters	~7.96 billion (99.46%)
Quantization	4-bit (NF4)

9.2 Training Hyperparameters

Hyperparameter	Value
Training examples	10,000
Batch size per GPU	4
Gradient accumulation steps	4
Effective batch size	16
Number of epochs	1
Total optimization steps	625
Learning rate	2e-5
Learning rate scheduler	Cosine with warmup
Warmup ratio	0.03
Optimizer	AdamW (8-bit)

Weight decay	0.01
Max sequence length	4096 tokens

9.3 Label Masking for Causal Language Modeling

Label masking is used in causal language modeling to ensure the model is trained only on the tokens it is expected to generate. In our task, the input includes both prompt text and the scene description, but the model should learn only to generate the scene description.

We use **attention-based label masking** to exclude prompt tokens from loss computation. This prevents the model from learning to reproduce the prompt and instead focuses training on the desired output.

Our implementation follows:

- **Prompt tokens**: Labels set to -100 (ignored in loss computation)
- **Response tokens**: Labels set to actual token IDs (included in loss)

Verification confirmed that about **99% of tokens were masked as prompt tokens**, with only the **scene description tokens (~1%)** contributing to the training loss. This ensures the model learns to generate appropriate scene descriptions rather than memorizing prompts.

9.4 Computational Infrastructure

All training was conducted on NCSA's Delta high-performance computing cluster:

- **GPU:** NVIDIA H200 (141 GB HBM3 memory)
- **Partition:** gpuH200 (latest generation nodes)
- **Resource:** gpu1e
- **Job scheduler:** SLURM
- **Training time per model:** ~4 hours (625 steps)
- **Iteration speed:** ~15-17 seconds per step

10. Evaluation Framework

10.1 Evaluation Metrics

To evaluate how well the model predicts the next panel’s scene description, we measure both text overlap and semantic similarity between the predicted output and the reference. These metrics below assess wording, fluency, and meaning alignment.

Metric	Description
ROUGE-1	Unigram overlap between prediction and reference
ROUGE-2	Bigram overlap, captures phrase-level similarity
ROUGE-L	Longest common subsequence, captures fluency

BLEU	N-gram precision with brevity penalty
BERTScore	Semantic similarity using contextual embeddings

10.2 Experimental Conditions

We evaluate four experimental conditions to understand the impact of fine-tuning and labeling paradigms. (Refer to Section 7.3 for notation details about the labeling paradigm):

- **Zero-Shot DESCP:** Base LLaVA model evaluated against Gemini scene descriptions
- **Zero-Shot PRED:** Base LLaVA model evaluated against Gemini scene predictions
- **Fine-Tuned DESCP:** LoRA fine-tuned model trained and evaluated on descriptions
- **Fine-Tuned PRED:** LoRA fine-tuned model trained and evaluated on predictions

This 2×2 design allows us to isolate the effects of fine-tuning from the effects of labeling paradigm choice.

10.3 Inference Configuration

To ensure a fair comparison between zero-shot and fine-tuned models, we use the same inference settings for both, so that any performance differences are due to training and not decoding choices.

Parameter	Value
Max new tokens	800
Temperature	0.3
Top-p (nucleus sampling)	0.9
Repetition penalty	1.1
Do sample	True

11. Training Dynamics and Observations

11.1 Loss Curves

Training loss trajectories for both paradigms demonstrate successful learning:

Training Step	DESCP Loss	PRED Loss
Step 10	2.226	1.640
Step 70	1.951	1.271
Step 120	~1.95	~1.27
Reduction	12%	23%

The PRED paradigm shows lower starting loss and faster convergence, which we attribute to better alignment between training and the model's natural generation tendencies. The DESCP paradigm, while having higher initial loss, still achieves substantial improvement.

11.2 Label Masking Verification

We verified proper label masking by analyzing token distributions in training batches:

Component	DESCP Model	PRED Model
Total tokens	13,503	13,576
Masked (prompt)	13,430 (99.5%)	13,430 (98.9%)
Active (response)	73 (0.5%)	146 (1.1%)

The high masking ratios confirm that the model is learning to generate scene descriptions rather than reproducing prompts, which was a critical concern after identifying a label masking bug in earlier training attempts.

12. Stable Diffusion Integration

The final stage of our pipeline uses the predicted scene descriptions to generate visual representations of the predicted panel using Stable Diffusion. This provides a qualitative evaluation of whether the scene predictions are sufficiently descriptive to guide image generation.

12.1 Generation Pipeline

- LLaVA generates scene description from 5 context panels
- Scene description is formatted as Stable Diffusion prompt
- Stable Diffusion generates candidate image
- Generated image compared visually with actual Panel 6 (Future work is to evaluate image similarity between the actual panel image and the generated image)

While quantitative evaluation of generated images against original panels is challenging (the original panels have specific artistic styles, character designs, and layouts that Stable Diffusion cannot replicate), qualitative inspection reveals whether the semantic content of predictions is reasonable.

SAMPLE LLAVA OUTPUTS (Extracted from Finetuned PRED Paradigm Model. The text boxes in each image is the output of LlaVa we executed which indicates the predicted next panel):

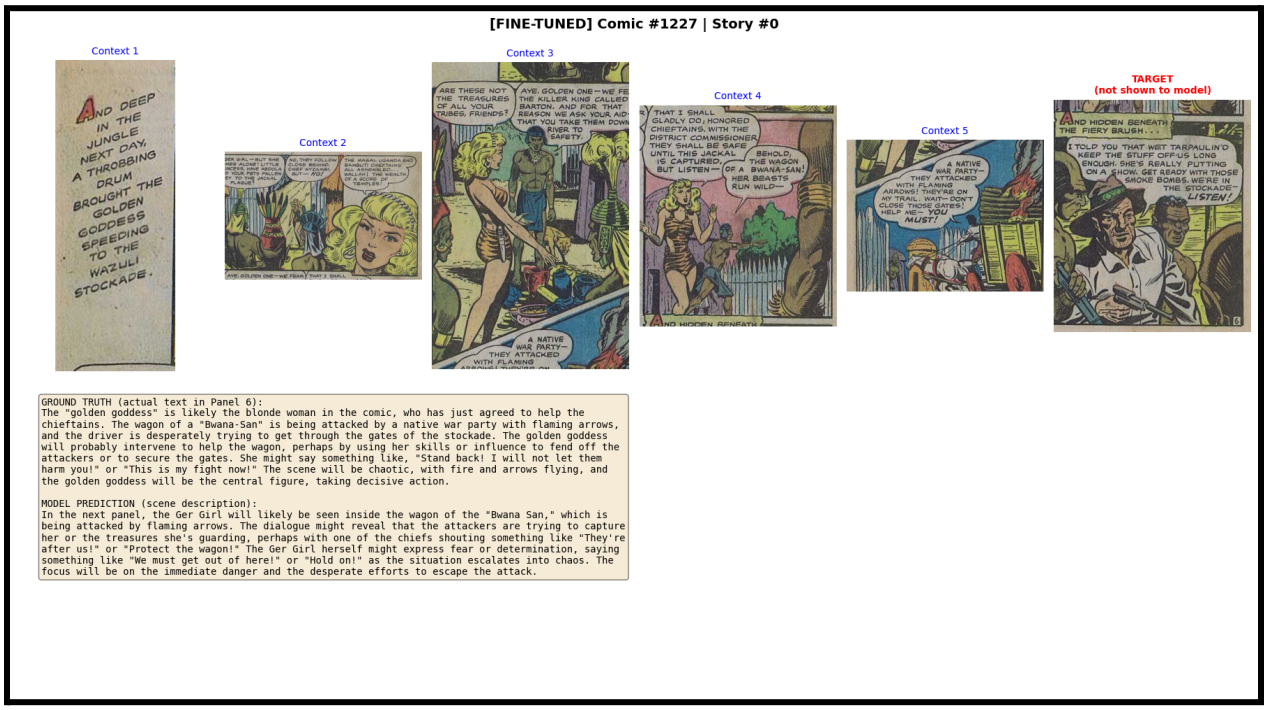


Fig 3 - The Golden Goddess & BWANA SAN against the raiders

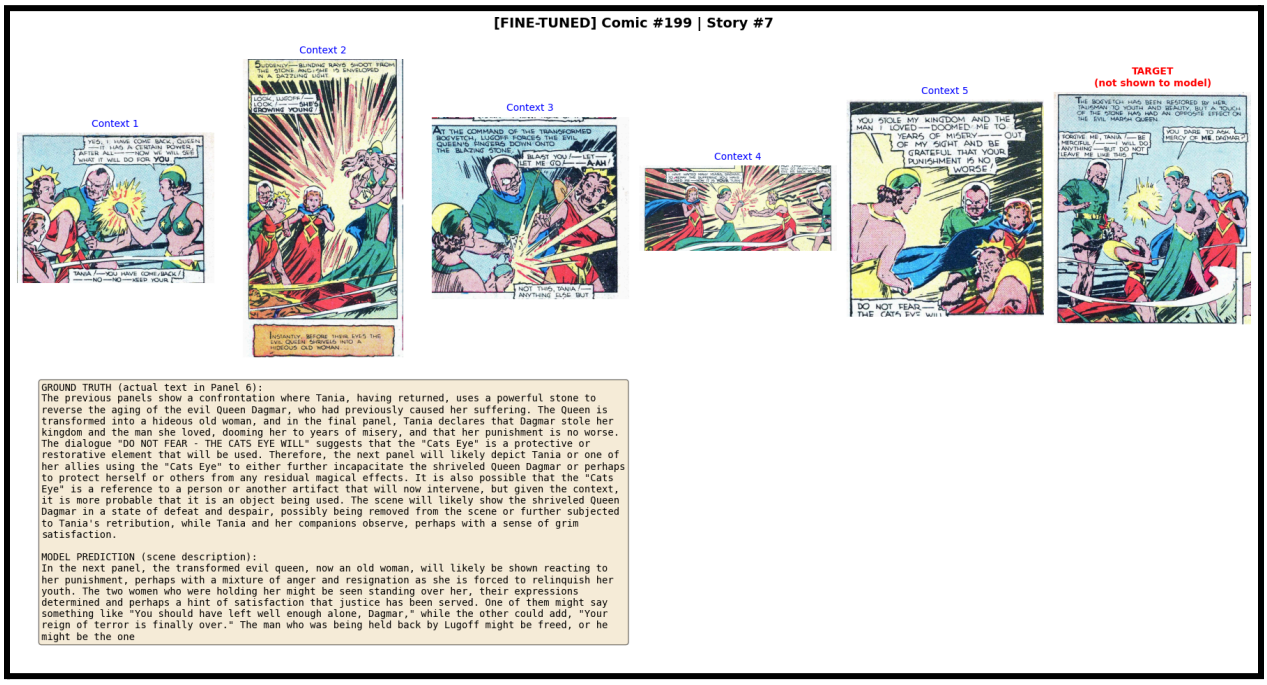
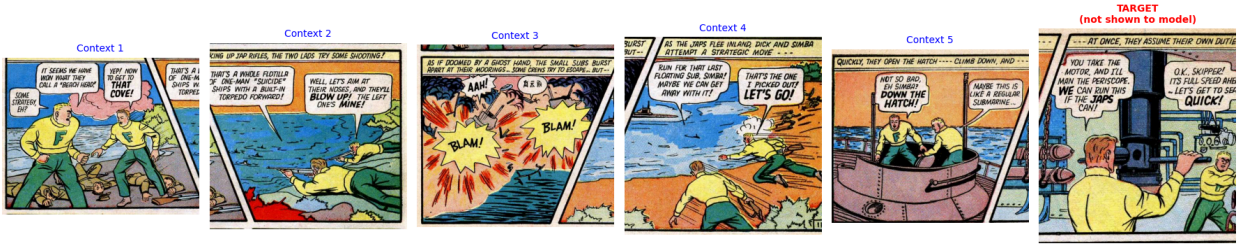


Fig 4 - The Evil Queen Story

[FINE-TUNED] Comic #54 | Story #1

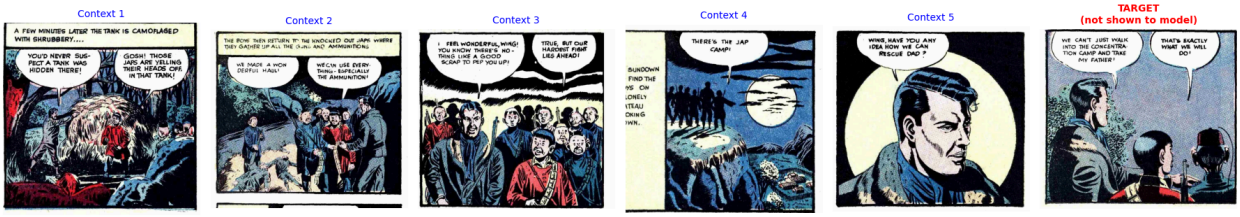


GROUND TRUTH (actual text in Panel 6):
Dick and Simba, having successfully commandeered a Japanese submarine, are now inside, exploring their new vessel. Simba might remark, "This is quite different from our usual mode of transport, Dick," while Dick, ever the optimist, could reply, "Indeed, Simba, but I have a feeling this will be quite useful in our continued efforts." They will likely begin to familiarize themselves with the submarine's controls, perhaps with Dick attempting to start the engine or Simba looking for navigational equipment. The scene will probably show them in the cramped interior of the submarine, with the sound of the engine starting or the periscope being raised, signaling their departure from the cove and their next move in the ongoing conflict.

MODEL PREDICTION (scene description):
The next panel will likely depict Dick and Simba successfully entering the submersible hatch of the last floating sub they spotted. They might be shown climbing down into the cramped interior, with one of them exclaiming something like "Not so bad, eh Simba?" as they begin their escape. The scene could show the hatch closing behind them, leaving the rest of the shipwrecked crew to their fate, or perhaps a brief glimpse of the surrounding water as they start their journey underwater. Dialogue could include their thoughts about their predicament, such as "Maybe this is like a regular submarine..

Fig 5 - Dick and Simba against the Jap Submarines

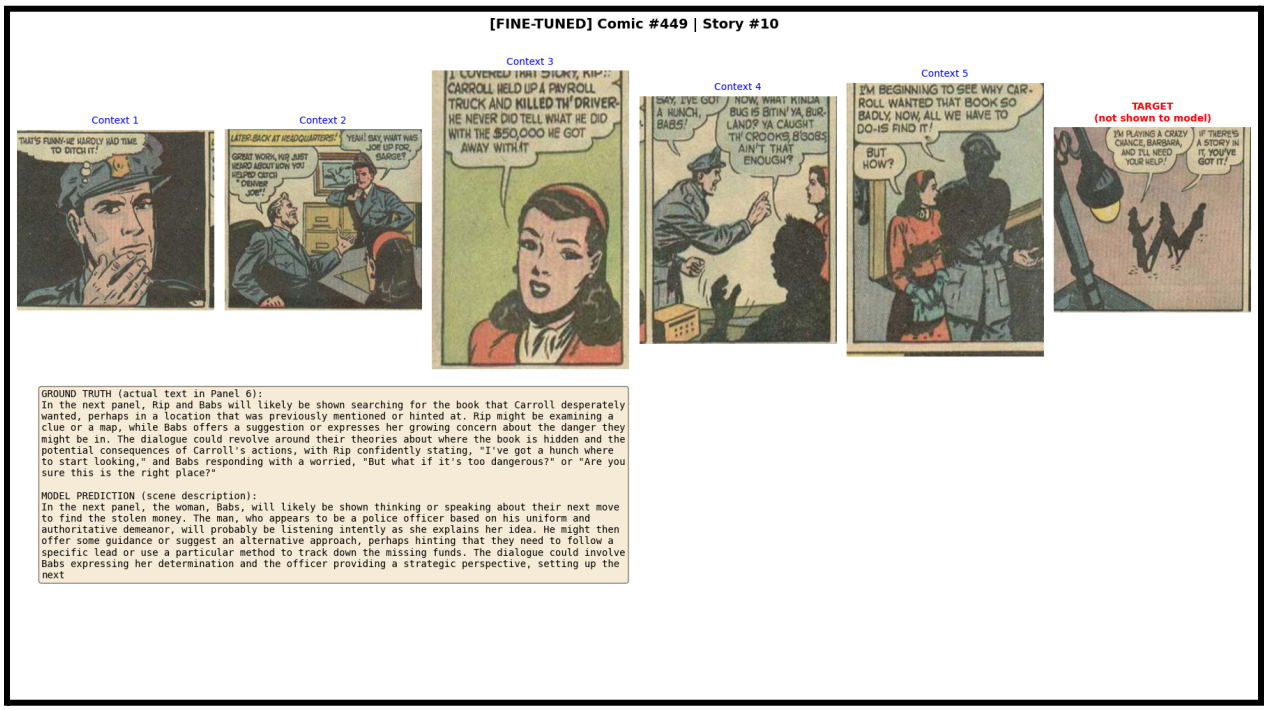
[FINE-TUNED] Comic #508 | Story #6



GROUND TRUTH (actual text in Panel 6):
Panel 6 will likely depict Wing and his companions preparing to infiltrate the Japanese camp. We might see them silhouetted against the moonlit sky, as shown in Panel 4, with Wing looking determined as he contemplates the rescue mission. He might respond to the question about rescuing his dad by saying something like "We'll have to be stealthy, but I have a plan," or perhaps he will point towards the camp and declare, "That's where we go in." The other characters, including the one who asked about rescuing their dad, will be looking towards the camp with a mixture of apprehension and resolve, ready to follow Wing's lead into what is described as their "hardest fight."

MODEL PREDICTION (scene description):
The next panel will likely depict Wing and his companions arriving at the Japanese camp, perhaps with a determined look on their faces as they survey the area. The dialogue might reveal their plan of attack or their immediate objectives, such as "We'll take them by surprise!" or "Let's get to work, Wing." They might also be seen approaching the camp, possibly under the cover of darkness or using stealth to infiltrate, setting up for a confrontation. The focus will be on their strategy and readiness to engage the enemy, emphasizing the "hardest fight lies ahead" as mentioned in Panel 3.

Fig 6 - Wing and the Escape from the Camp



Comic #449-Barbara and the Cops

SAMPLE STABLE DIFFUSION SAMPLE OUTPUT:

Fig 8 - Image generated by stable diffusion using the LlaVa - generated prediction. (Left)This was generated with the prompt: *"In the next panel, we see a dramatic moment as the ship's mast begins to collapse towards the characters. The sky above them is dark and ominous, suggesting an impending storm or the aftermath of a battle. The character who was previously speaking about the logbook seems to be in a state of panic, while another character appears more composed but equally alarmed by the situation. The sound effect \"CRACK\" echoes through the scene, emphasizing the imminent danger. The dialogue could include*

urgent warnings and pleas for action, with one character urging the other to jump overboard before it's too late. The art style remains consistent with bold lines and vibrant colors, capturing the urgency and chaos of the moment." **(Right) is the image that must have been generated**

13. Results

We evaluated four experimental conditions across 1,000 test sequences each, comparing zero-shot and fine-tuned models under both labeling paradigms (DESCP and PRED). All experiments used LLaVA-OneVision-Qwen2-7B as the base model with LoRA fine-tuning (43.2M trainable parameters, 0.54% of total).

13.1 Quantitative Results

Table 1 presents the evaluation metrics across all experimental conditions. Fine-Tuned PRED achieves the best performance across all metrics.

Table 1: Evaluation Metrics Across Experimental Conditions

Condition	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore
Zero-Shot DESCP	0.3300	0.0626	0.1726	0.0155	0.8471
Zero-Shot PRED	0.3487	0.0707	0.1785	0.0219	0.8535
Fine-Tuned DESCP	0.3355	0.1043	0.2189	0.0451	0.8602
Fine-Tuned PRED	0.4390	0.1325	0.2322	0.0558	0.8777

13.2 Relative Improvements

Table 2 shows the relative improvements between conditions, highlighting the impact of task alignment and fine-tuning.

Table 2: Relative Improvements Between Conditions

Comparison	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
ZS-PRED vs ZS-DESCP (task alignment)	+5.7%	+12.9%	+3.4%	+0.8%
FT-DESCP vs ZS-DESCP (fine-tuning)	+1.7%	+66.6%	+26.8%	+1.5%
FT-PRED vs ZS-PRED (fine-tuning)	+25.9%	+87.4%	+30.1%	+2.8%
FT-PRED vs ZS-DESCP (overall best)	+33.0%	+111.7%	+34.5%	+3.6%

13.3 Training Dynamics

Table 3 compares the training dynamics between DESCP and PRED paradigms. The PRED model starts with significantly lower loss (1.640 vs 2.226), indicating that predicting Panel 6

without visual access is inherently easier when labels are also predictions rather than descriptions.

Table 3: Training Dynamics Comparison

Component	DESCP Model	PRED Model
Initial Loss	2.226	1.640
Final Loss	1.827	1.251
Loss Reduction	18.0%	23.7%
Training Time	192.1 min	192.2 min
Total Tokens	13,503	13,576
Masked (prompt)	13,430 (99.5%)	13,430 (98.9%)
Active (response)	73 (0.5%)	146 (1.1%)

13.4 Qualitative Analysis

We conducted detailed qualitative analysis of model predictions across all four conditions. Key patterns emerged in character naming, narrative continuity, output conciseness, and error modes.

13.4.1 Character Naming

Fine-tuned models consistently identify characters by name from context dialogue, while zero-shot models rely on generic descriptions. For example, in Comic #508, Fine-Tuned PRED correctly identifies "Wing" planning a rescue mission, while Zero-Shot DESCP describes "a man in a dark coat." Similarly, Fine-Tuned PRED correctly names characters like "Hangman," "Joey," "Dagmar," "Roger Dalton," and "Tommy Owens" across various comics.

13.4.2 Narrative Continuity

Fine-Tuned PRED demonstrates superior narrative continuity by explicitly referencing earlier context. In Comic #508 (Wing rescue mission), the model prediction includes "emphasizing the 'hardest fight lies ahead' as mentioned in Panel 3," directly connecting to dialogue from an earlier panel. This contextual integration was absent in zero-shot predictions, which treated each prediction more independently.

13.4.3 Output Conciseness

Fine-tuning produces more focused, panel-appropriate predictions. Zero-Shot DESCP generates verbose outputs averaging 250-300 words with excessive speculation about sound effects and background details. Fine-Tuned DESCP reduces this to approximately 80-100 words while maintaining narrative accuracy. This conciseness is reflected in inference speed: Fine-Tuned DESCP achieves 4.30 sec/example versus 6.02 sec/example for Zero-Shot DESCP (29% faster).

13.4.4 Best and Worst Examples by Condition

Table 4 summarizes the best and worst performing examples for each experimental condition, along with error analysis.

Table 4: Best and Worst Examples by Condition

Condition	Best Example	Worst Example	Primary Error Mode
ZS-DESCP	#199 Evil Queen	#449-Barbara	Scene transition failure: predicted indoor office instead of outdoor night scene
FT-DESCP	#54 Submarine	#449-Barbara	Same scene transition failure; also wrong character genders (two men vs man and woman)
ZS-PRED	#54 Submarine	#449-Barbara	Over-reliance on office context; failed to anticipate transition to investigation scene
FT-PRED	#508-Wing Mission	#449-Barbara	Persistent difficulty with this specific comic's scene transition pattern

13.4.5 Best Example Walkthrough: Comic #508 (Fine-Tuned PRED)

Comic #508 (Story #6) depicts Wing and companions preparing to rescue "Dad" from a Japanese concentration camp during WWII. This example demonstrates the strongest performance of Fine-Tuned PRED.

Ground Truth (Gemini PRED): "Panel 6 will likely depict Wing and his companions preparing to infiltrate the Japanese camp. We might see them silhouetted against the moonlit sky, as shown in Panel 4, with Wing looking determined as he contemplates the rescue mission."

Model Prediction: "The next panel will likely depict Wing and his companions arriving at the Japanese camp, perhaps with a determined look on their faces as they survey the area. The dialogue might reveal their plan of attack or their immediate objectives, such as 'We'll take them by surprise!' or 'Let's get to work, Wing.' They might also be seen approaching the camp, possibly under the cover of darkness or using stealth to infiltrate, setting up for a confrontation. The focus will be on their strategy and readiness to engage the enemy, emphasizing the 'hardest fight lies ahead' as mentioned in Panel 3."

Analysis: The model correctly identifies: (1) Wing by name, (2) nighttime/stealth setting, (3) mission objective (infiltrating camp), (4) determined emotional tone, and critically (5) directly references dialogue from Panel 3 ("hardest fight lies ahead"), demonstrating learned narrative coherence across panels**.**

13.4.6 Worst Example Walkthrough: Comic #449 (All Conditions)

Comic #449 (Story #10) represents a persistent failure case across all four conditions. The context panels show office/headquarters scenes, but Panel 6 transitions to an outdoor night scene with Rip and Barbara under a street lamp.

Ground Truth: "In Panel 6, the scene is set outdoors, possibly at night, with a large, stylized street lamp casting a warm glow on the ground. Two silhouetted figures, one taller and one shorter, are engaged in conversation. The taller figure, presumably Rip, gestures emphatically as he tells Barbara, 'I'M PLAYING A CRAZY CHANCE, BARBARA, AND I'LL NEED YOUR HELP!'"

All Models Predicted: Indoor office/headquarters continuation with investigation dialogue, often featuring a police officer character not present in the actual panel.

Error Analysis: This failure reveals a fundamental limitation: models over-rely on visual continuity from recent context panels and struggle to anticipate narrative-driven scene transitions. When context panels consistently show indoor settings, the model cannot predict an outdoor transition even when dialogue cues suggest a change in location. This represents a ceiling on current visual narrative understanding.

13.5 Interpretation and Discussion

13.5.1 Why PRED Outperforms DESCP

The PRED paradigm achieves better metrics even without fine-tuning (+5.7% ROUGE-1) because of task alignment. In DESCP, Gemini describes visual details it can see in Panel 6 (specific colors, exact character positions, facial expressions), while LLaVA must predict without this visual access. In PRED, both Gemini (generating labels) and LLaVA (at inference) perform identical forward-prediction tasks without seeing Panel 6. This alignment means the model is evaluated on its ability to perform the same task as the label generator, rather than being penalized for not describing visual details it cannot access.

13.5.2 Why Fine-Tuning Benefits PRED More Than DESCP

Fine-tuning amplifies the task alignment advantage. PRED benefits substantially more from fine-tuning (+25.9% ROUGE-1) compared to DESCP (+1.7%). The DESCP model must learn to bridge the gap between prediction and description, essentially learning to hallucinate visual details. This misaligned learning signal limits improvement. In contrast, PRED fine-tuning reinforces aligned prediction patterns, leading to more efficient learning. The training dynamics confirm this: PRED starts with lower loss (1.640 vs 2.226) and achieves greater reduction (23.7% vs 18.0%)**.**

13.5.3 ROUGE-2 Shows Largest Gains

The 111.7% improvement in ROUGE-2 (bigram overlap) from Zero-Shot DESCP to Fine-Tuned PRED is particularly notable. This indicates that fine-tuning helps the model capture phrasal patterns and multi-word expressions characteristic of comic narratives, such as character name + action combinations, dialogue patterns, and scene transition phrases.

13.5.4 BERTScore Ceiling Effect

BERTScore shows the smallest relative improvement (+3.6%) despite being the highest absolute value (0.8777). This reflects a ceiling effect: BERTScore measures semantic similarity at the embedding level, and all models already capture the general semantic content of comic narratives well. The improvements from fine-tuning are primarily in lexical precision (ROUGE) rather than semantic understanding.

14. Limitation and Future Work

1. We experimented with Stable Diffusion to generate comic panels from text prompts. While the model was effective at producing images with specified style, we observed that it does not reliably generate readable or accurate text within images. This is a limitation in our work - not being able to generate an authentic comic book panel.
2. Ground truth subjectivity: Both DESCP and PRED ground truth labels are generated by Gemini, introducing the biases and limitations of that model into our evaluation. A prediction that differs from Gemini's but is equally valid would receive lower scores.

3. Metric limitations: ROUGE and BLEU measure lexical overlap, which may not fully capture narrative quality. Two semantically equivalent but lexically different predictions would receive different scores.
4. Stable Diffusion text generation: While our pipeline successfully generates panel images from scene descriptions, Stable Diffusion does not reliably generate readable or accurate text within images. This limits the end-to-end fidelity of generated comic panels, as dialogue bubbles and captions are often garbled or missing.
5. Dataset specificity: The COMICS dataset consists of Golden Age (1930s-1950s) public domain comics with specific artistic styles, dialogue patterns, and narrative conventions. Generalization to modern comics with different visual styles remains untested.

15. Division of Labour

1. **Anushree Udhayakumar:** Conducted initial experimentation with LLaVA, evaluating all model variants discussed in Section 6.2. Studied and developed the Stable Diffusion pipeline across multiple versions as presented. Led the report writing and performed proofreading and final edits.
2. **Harshvardhan Sekar:** Handled data preparation and orchestration on Google Cloud to generate silver-standard labels. Fine-tuned LLaVA models for both labeling paradigms, implemented the end-to-end pipeline, conducted evaluation, and documented the experimental results.
3. **Prisha Singhania:** Evaluated OpenFlamingo, QWEN and identified key architectural and task-specific limitations that made it unsuitable for multi-panel comic narrative prediction.
4. **Sonia Navale:** Contributed to model parameter selection and conducted the literature survey to support methodological and design decisions.

16. References

1. Iyyer, M., Manjunatha, V., Guha, A., Vyas, Y., Boyd-Graber, J., Daume, H., & Davis, L. S. (2017). 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* (pp. 7186-7195).
2. Digital Comic Museum. (n.d.)], from https://digitalcomicmuseum.com/
3. Hugging Face. (2023). Stable Diffusion [Computer software]. Hugging Face. The Stable Diffusion Guide 🎨
4.]MLFoundations. (2023). OpenFlamingo [Computer software]. GitHub mlfoundations/open_flamingo: An open-source framework for training large multimodal models.
5. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., ... & Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, *35*, 36479-36494.
6. Sapkota, R., Ahmed, D., & Karkee, M. (2024). Comparing YOLOv8 and Mask R-CNN for instance segmentation in complex orchard environments. *Artificial Intelligence in Agriculture*, *13*, 84-99.
7. Vivoli, E., Souibgui, M. A., Barsky, A., LLabres, A., Bertini, M., & Karatzas, D. (2024). One missing piece in vision and language: A survey on comics understanding. *arXiv preprint arXiv:2409.09502*.One missing piece in Vision and Language: A Survey on Comics Understanding