Multimodal Large Language Models for Visual Segment Classification and Description Generation in Digital Storybooks

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) are advanced in handling complex visual-textual tasks, but their application to narrative-driven contexts remains underexplored. In this work, we evaluate the ability of MLLMs to identify relevant visual segments and generate descriptions for segments in illustrated digital storybooks. We curate a dataset of 14,162 segments, extracted from 32 Arabic children's digital storybooks through the Segment Anything Model (SAM), with human annotations for segment relevance and descriptive labels. We evaluate five state-of-the-art MLLMs across zero-shot prompting conditions, and evaluate the two best-performing models through few-shot. Our results show that few-shot prompting of GPT-4o achieves the best results for segment relevance classification. While all models struggle with fine-grained contextual reasoning, our findings provide insights for developing AI-powered interactive digital storybooks and help advance multimodal methodologies in narrative understanding tasks.

1 Introduction

As Large Language Models (LLMs) become more adept at handling complex tasks, the recent shift towards Multimodal LLMs (MLLMs) extends their processing capabilities beyond text so that they encompass different modalities of information. This expansion widens the range of tasks they can cover and better mimics human multimodal sensing abilities (Fu et al., 2024; Huang and Zhang, 2024). However, several gaps remain. While these MLLMs can be advanced in single-image tasks, they struggle with tasks involving multiple related images (Wang et al., 2024a) in addition to fine-grained details within illustrations (Wang et al., 2024a; Fu et al., 2024). Moreover, existing MLLM evaluations primarily focus on vision-language tasks like visual question answering but neglect tasks that require deeper contextual understanding (Yang et al., 2023).

This gap in deeper contextual understanding and fine-grained visual perception is salient in applications like digital storytelling. Within this work, illustrations in digital storybooks have played a major role in children's story comprehension and engagement by providing visual reinforcement to the storybook narrative (Bus et al., 2019). Features that include interactive, clickable visual elements can especially help with introducing more interactivity beyond passive reading (Kamil et al., 2023). The use of Artificial Intelligence (AI) has been increasingly integrated into digital story-telling for immersive interactions (Sun et al., 2024); however, there is a lack of research in exploring AI-driven illustration segment (objects within a storybook page image) interaction in digital storybooks. This segment interaction requires not only object recognition but also narrative comprehension, which is under-explored in MLLM benchmarks (Yang et al., 2023).

Therefore, in this work, we introduce an approach for evaluating visual segments in digital storybooks using MLLMs to enhance interaction. We utilize and assess MLLMs (GPT-4o, GPT-4o-mini, Gemini 1.5 Pro, Gemini 2.0 Flash, and Claude 3 Opus) in identifying relevant auto-generated segments from storybook illustrations based on the narrative context. To analyze model behavior across varying levels of contextual support, we evaluate performance under both zero-shot and few-shot prompting conditions (using 2-shot examples through dynamic and fixed example selection strategies). Our goal is to advance narrative-driven multimodal evaluation by comparing model performance to human perception of segment relevance based on the story's context.

Our research question is as follows: How well do state-of-the-art multimodal language models (MLLMs) identify relevant segments from story-

book illustrations based on the storybook narrative?

We investigate whether current MLLMs can accurately determine the relevance of segment illustrations within a story's context by comparing their performance to human evaluation. Insights from this study can inform large-scale applications involving digital books, reducing the effort required to identify interactive segments that can bolster engagement and comprehension.

Our contribution to digital storytelling research and MLLM evaluation also extends to providing a benchmark comparison of MLLMs for visual segment relevance in narrative multimodal tasks and introducing a novel evaluation framework leveraging storybook context. While we applied our method to children's storybooks written in Arabic, the findings have broader implications for narrative-driven visual interaction in other domains.

2 Related Work

AI in Storybooks: Storytelling and Visuals AI in the realm of storytelling has supported the creation of interactive and personalized narratives and enhancement of readers' engagement. For instance, AI has shown potential in managing fluid narrative structures (Cavazza and Charles, 2003; Bostan and Marsh, 2012), adapting stories to user interactions in gamified storytelling and education (Riedl, 2012; Katifori et al., 2018; van Druten-Frietman et al., 2016), and fostering creativity in collaborative storytelling platforms (Garzotto et al., 2010; Burtenshaw, 2023). AI-supported tools like AI Stories and Storypark offer structured narrative assistance that has shown improvements in language skills and comprehension (Ye et al., 2024; Burtenshaw, 2023). However, many of these systems rely on rule-based mechanisms, emphasizing narrative coherence over open-ended creativity (Cavazza and Charles, 2003; Bostan and Marsh, 2012). While interactive storytelling in education enables user agency, it operates within predefined limits (Riedl, 2012; van Druten-Frietman et al., 2016). Few studies address AI's role in less structured creative domains, such as poetry-based or ethically nuanced storytelling (Świerczyńska Kaczor, 2024), and there is limited discussion on integrating automation with artistic intent. There is a gap in unstructured, context-rich storytelling settings, especially within a children's digital storybooks format, highlights the need for AI systems that interpret visual-narrative relationships dynamically, which we aim to address in this work.

Multimodal LLMs Recent advances in MLLMs target an amalgam of concetps from text, to images and structured data, extending LLMs with visual reasoning for tasks such as image-text retrieval, VQA, and document understanding (Zhang et al., 2024; Wang et al., 2024b). Vision-language models like ImageBERT strengthen cross-modal alignment through large-scale pre-training, impacting domains such as healthcare (Qi et al., 2020; Wang et al., 2023). Binary image-text relevance classification remains a core challenge for validation-intensive settings. Approaches like LLaVA-RE categorize pairs as 'relevant' vs. 'not relevant' (Sun et al., 2025), while ImageBERT employs image-text matching losses to enhance classification precision (Qi et al., 2020). Parallel work on Arabic-English cross-lingual prompting shows systematic gains in translation and multimodal retrieval, yet multilingual evaluation protocols are few (Nagi et al., 2024). In healthcare, GPT-RadScore was studied to assess MLLM's ability in fine-grained, task specific medical assessments (Zhu et al., 2024). Building on this work, we shift the focus from pairwise image-text relevance and domain-specific reporting to narrative-driven, segment-level visual relevance in children's digital storybooks, targeting the space of context-sensitive, multilingual, fine-grained illustration segmentation aligned with story narratives.

Applications of Segment Anything Model The Segment Anything Model [1] (SAM) enables flexible, zero-shot image segmentation across domains like medical imaging, agriculture, and geology (Ma et al., 2023; Zhang and Wang, 2023; Carraro et al., 2023). Its strength lies in minimizing fine-tuning while maintaining high accuracy, yet challenges remain with low-contrast images (Huang et al., 2024). This existing work focuses on static tasks to assess technical precision in various fields. However, there has not been work on SAM's applications in a user interaction design context. Our work aims to bridge the gap between automated segmentation and interactive digital storybooks design.

3 Method

3.1 Dataset Preparation

To analyze storybook illustration segments, we curate a dataset for segments from 32 illustrated

[1] https://segment-anything.com/

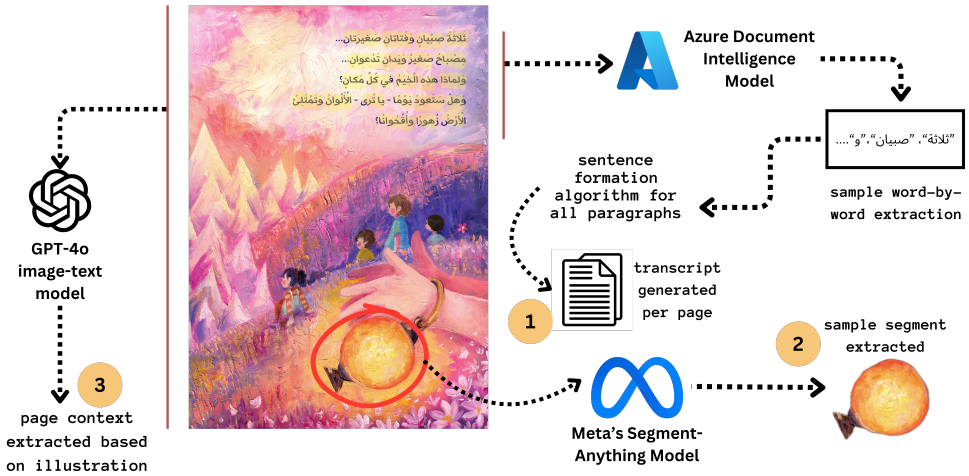


Figure 1: Dataset Preparation Process Per Storybook Page for Extracting Transcript, Context, and Segments

Arabic children's storybooks provided in digitalized PDF format and converted to images. The preparation pipeline consists of the following: text extraction, image conversion, contextual augmentation, and segmentation processing, with semi-automated techniques to handle each story. Each story's dataset is structured per page, including page-level transcripts, extracted images, context description, and segmentation outputs in order to provide a rich context for MLLM evaluation. Figure 1 summarizes the data collection process. We support MLLM's evaluation of segment relevance by both segment-level (local) and narrative-level (global) contexts for each analysis step by providing page-level transcripts as local context for the extracted image segments, along with story-extracted contextualized narrative descriptions that offer a broader understanding of the story.

Storybook Preprocessing Due to limitations of existing Arabic OCR tools in extracting illustrated Arabic text correctly with diacritics, we utilize Azure's Document Intelligence model [2] for text extraction. Since children's illustrated storybooks rely heavily on visual and textual interplay, providing robust contextual information is crucial. Thus, we use GPT-4o's vision capability to generate Arabic contextual descriptions of each page by taking into account visual elements and referencing prior pages for narrative coherence. The prompt used for this process is included in the Appendix (Section A.2). Finally, to extract individual illustrated segments from each storybook page, we employ Meta's SAM to automate the process. For each illustrated page, segments were extracted into PNG files with their polygon data in JSON format. All of this data is then organized hierarchically per page then per story.

[2] https://azure.microsoft.com/en-us/products/ai-services/ai-document-intelligence

3.2 Segment Annotation

To establish a ground truth for evaluating MLLMs on illustrated storybook segments, two bilingual annotators proficient in Arabic and English were recruited from Fiverr and were informed that their labeling work would be used for research purposes. No personal or sensitive information was collected or retained. Annotators were asked to assess the relevance of each segment to the narrative (i.e., the story's context on that specific page, as provided by the textual transcript and contextualized narrative description) on a binary scale, as well as assign a one-word description in English for relevant segments, or 'nothing' otherwise. Annotators were compensated 5 cents per segment. The annotator guidelines mirrored the instructions given to the MLLMs through our prompts.

We first measure Inter-Annotor Agreement (IAA) on a subset of three randomly selected stories from 32 stories (996 segments in total), using Cohen's kappa (κ) for binary relevance classification and SpaCy's pre-trained model `en_core_web_md`, with a threshold of 0.8, for semantic similarity of the one-word descriptions (Honnibal et al., 2020; AI, 2020a). Cohen's kappa, accounted for agreement by chance beyond simple accuracy, while

the SpaCy model was chosen because it includes tagging, parsing, lemmatization and named entity recognition (AI, 2020b), which makes it useful for calculating semantic similarity. We chose the pipeline that is trained on written web text. Given that storybook segments involve objects, actions, or concepts needing contextual understanding, the model is well-suited for our evaluation.

A preliminary Zoom session was conducted to clarify guidelines. The guidelines are given in detail in the Appendix (Section A.3). For the first round, Cohen's kappa was 0.72 (substantial agreement (Cohen, 1960; McHugh, 2012)) and average semantic similarity was 94.2%. To improve agreement, a follow-up discussion addressing annotation discrepancies was held. A second annotation round on the three stories improved Cohen's kappa to 0.97 and one-word description semantic similarity to 97.2%. The remaining disagreements for this subset were resolved by a super-annotator (the task designer), and the remaining 29 stories (13,166 segments) were evenly distributed between annotators independently. This curated dataset provides the ground truth for evaluation of MLLM performance.

3.3 Experimental Setup and Implementation

We collect results from all MLLMs through standardized API calls in a Google Colab environment through batch processing. All outputs are parsed into JSON format. We split our dataset (14,162 illustrated segments from 32 children's Arabic stories) into an 80/20 split, where 80% is used for evaluation, and 20% is used for few-shot training. The split is done on a story-level so that all segments from the selected stories are either in the training or evaluation set. The 80% part includes 22 stories (with a total of 10,921 segments). The 20% part has the remaining 10 stories (3,241 segments).

The experiments are structured into two distinct phases: zero-shot and few-shot prompting. In the zero-shot phase, we evaluate MLLMs on the 80% part of the dataset, to make results directly comparable to the few-shot's, using macro-F1 for binary classification of relevance and semantic similarity of descriptions across the following models: GPT-4o, GPT-4o-mini, Gemini 1.5 Pro, Gemini 2.0 Flash, and Claude-3 Opus. For the few-shot experiments, the 20% part of the dataset is used for extracting examples. In this setup, each prompt provides the model with two annotated examples alongside the test segment from the evaluation dataset. Each of those examples includes a segment image, its corresponding full-page image, the page-level transcript, contextual description, binary relevance (true/false), and a one-word description.

We explore two approaches for selecting few-shot examples.

- **Fixed method**: Two examples, one labeled relevant (true) and the other labeled irrelevant (false), are randomly selected from the training set and used consistently across all few-shot prompts.

- **Dynamic method**: For each test instance, select the two most semantically and visually similar segments from the example pool by generating and comparing multimodal embeddings, using the CLIP model (Radford et al., 2021). Each segment is encoded into a 1536-dimensional vector by concatenating embeddings of the segment image, the full-page illustration, and combined textual inputs (page transcript and context). Cosine similarity is used to retrieve the top two most relevant examples, thus enabling tailored few-shot prompts for each instance. The algorithm in detail is in the Appendix (Section A.4).

We conduct prompt engineering iteratively by using held-out examples from the training data until all models achieve consistent adherence to the guidelines given to annotators. The final prompt is referenced in the Appendix(Section A.1. To prepare for evaluation, we normalize the outputs for casing and formatting. Invalid responses like multi-word descriptions or punctuated outputs are cleaned. For segments that are missed within the batch processing output, we issue single API calls to recover the missing predictions. We then run evaluation scripts to process the outputs to compute binary relevance accuracy for the segments and semantic similarity between model-predicted and annotated descriptions.

3.4 Evaluation

Relevance Classification The first evaluation task requires MLLMs to classify storybook illustration segments based on their contextual relevance within the story's narrative. Each segment would receive a binary relevance label (True/False), informed by visual and textual context provided in the prompt. Given the dataset's class imbalance (approximately 20% relevant and 80% irrelevant

Table 1: Zero-Shot Evaluation Results: F1-Score, Relevance Accuracy, and Semantic Similarity for Binary Classification of Contextual Relevance Across MLLMs

Model	Claude-3 Opus	GPT-4o-mini	GPT-4o	Gemini 1.5 Pro	Gemini 2.0 Flash
Relevance Accuracy	32.44%	73.4%	81.3%	62.16%	44.29%
F1 Score	0.3239	0.572	0.624	0.5645	0.4411
Semantic Similarity	33.8%	76.7%	83.9%	64.94%	45.69%

segments), we use the **macro-averaged F1-score** (Pedregosa et al., 2011; Opitz and Burst, 2019) as the primary metric as it considers the performance on both classes while penalizing poor performance on the minority class (Leung, 2022). We also report accuracy for completeness which is calculated as follows:

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Segments/Predictions}} \quad (1)$$

The evaluation results are computed across the 32 stories to show a comprehensive assessment of MLLM performance on the relevance classification task within a narrative-driven context.

Description Generation The second task is assigning single-word English descriptions to each segment. We compute semantic similarity between the model-generated outputs and ground truth using SpaCy's pre-trained model (`en_core_web_md`), similar to the approach used for IAA described in Section 3.2. We prioritize semantic similarity for evaluating this task since the main objective is to identify the object in the segment with a single-word description, where synonymous similarity is sufficient. The final reported measure is the average semantic similarity across all segments.

4 Results

The results from all models reveal remarkable discrepancies in the models' agreement with the ground truth, which may potentially be due to how these models handle Arabic textual input and produce English outputs in a multimodal setting.

4.1 Zero-Shot Experiments

In the first phase, we assess the model performance across the two tasks (binary classification and description generation) across the five MLLMs for the evaluation dataset. The results are encapsulated in Table 1, where relevance accuracy and F1 score are used for assessing the binary classification task while the semantic similarity metric is used to assess the descriptions provided.

Relevance Classification To establish a baseline, we include a majority baseline, which is a naive predictor that assigns the most frequent class ("not relevant") to all inputs. Since 80% of the ground truth annotations are negative (i.e., not relevant), this majority baseline achieves a relevance accuracy of 80% and a macro-averaged F1 score of 50%. We utilize these values for our baseline comparison. As summarized in Table 1, we utilize the macro-averaged F1 score as the principal evaluation metric for this task. This evaluation is conducted on 80% of the annotated dataset. Among all five models, GPT-4o ($F1 = 0.624$), GPT-4o-mini ($F1 = 0.572$), and Gemini-1.5-Pro ($F1 = 0.5645\%$) achieved performance levels that surpass the random-choice baseline of 50%, suggesting that these MLLMs are capable of capturing contextual relevance even under class imbalance. With respect to accuracy, while none of the models consistently exceeded the baseline across all conditions, GPT-4o ($Accuracy = 81.3\%$) and GPT-4o-mini ($Accuracy = 73.4\%$) demonstrated relatively stronger performance compared to the other evaluated systems. Claude-3 Opus exhibited the weakest results in this task compared to all models with an F-1 score of 0.3239 and an accuracy level of 32.44%, suggesting that the model may struggle with fine-grained relevance judgment when tasked with multimodal reasoning that involves both visual and linguistic context with multilingual content. These findings highlight challenges MLLMs face in nuanced classification scenarios especially when relevance judgments could hinge on narrative coherence and intermodal alignment.

Description Generation We utilize semantic similarity to measure alignment between model-predicted descriptions and human-annotated ground truth for identifying segments, aggregated over the evaluation dataset. As shown in Ta-

Table 2: Few-Shot Evaluation Results: F1-Score, Relevance Accuracy, and Semantic Similarity with Fixed and Dynamic Prompting Strategies Across MLLMs

Prompt Setting	Metric	GPT-4o-mini	GPT-4o
Few-Shot (Fixed)	Relevance Accuracy (%)	77.7	79.5
	Relevance F1-score	0.661	0.708
	Semantic Similarity (%)	79.1	78.5
Few-Shot (Dynamic)	Relevance Accuracy (%)	75.7	77.0
	Relevance F1-score	0.666	0.654
	Semantic Similarity (%)	76.5	78.2

ble 1, GPT-4o has the highest semantic similarity (83.9%), followed closely by GPT-4o-mini (76.7%), and Gemini-1.5-Pro (64.94%). These results indicate the models' ability to effectively generate one-word descriptions for segments in a multimodal, context-driven task. On the other hand, Claude-3-Opus (33.8%) and Gemini-2.0-Flash (45.69%) yielded lower alignment scores.

4.2 Few-Shot Experiments

In the second phase of our study, we examine the impact of few-shot prompting strategies on the performance of the two top-performing MLLMs from Phase 1: GPT-4o and GPT-4o-mini. We narrow down the models for Phase 2 to preserve computational resources while enabling a direct comparison in a few-shot setting. We evaluate two example strategies in few-shot in this phase: (1) fixed examples utilizing two pre-selected reference samples; and (2) dynamically selected examples chosen in real-time based on contextual similarity to the input segment. At this stage, we run the experiment on the evaluation dataset, while the training dataset is used for example selection.

Relevance Classification Table 2 summarizes the accuracy and F1-scores for GPT-4o and GPT-4o-mini across the two few-shot prompting strategies. GPT-4o demonstrated superior performance in the fixed few-shot setting, achieving higher accuracy ($Accuracy = 79.5\%$) and a notably stronger F1-score ($F1 = 0.708$). On the other hand, the dynamic few-shot strategy showed slightly reduced performance ($Accuracy = 77.0\%, F1 = 0.654$). This suggests that for GPT-4o, the consistent use of carefully pre-selected examples is more effective than dynamically retrieved examples, which may introduce redundancy or noise that hinders the model's generalization. As for GPT-4o-mini, the fixed few-shot setting yielded the highest accuracy

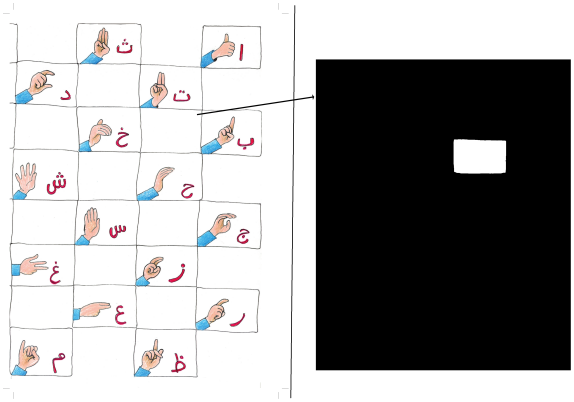


Figure 2: Example of a page containing sign language illustrations (left) and an extracted segment of white squares (right).

($Accuracy = 77.7\%$), but the highest F1-score was achieved using dynamically selected examples ($F1 = 0.666$). Although the differences between the two few-shot strategies are subtle for this model, it might gain more balanced precision and recall from examples tailored dynamically to input context rather than predetermined examples.

Description Generation Table 2 reports the semantic similarity scores under few-shot settings. GPT-4o showed minimal variation between fixed (78.5%) and dynamic (78.2%) prompting, indicating stable performance regardless of example selection strategy. GPT-4o-mini achieved the highest similarity with fixed examples (79.1%) which dropped slightly under dynamic prompting (76.5%). This indicates that larger models like GPT-4o remain robust across few-shot configurations in terms of identifying objects within images.

4.3 Error Analysis

A key challenge encountered in this study is the misclassification of irrelevant segments as relevant by multiple MLLMs. A notable case, shown

in Figure 2, is described for the several empty white squares extracted as segments from pages, showing hand gestures of Arabic sign language letters. In the zero-shot experiment phase, Gemini 1.5 misclassified them as “hand,” while Gemini 2.0 described them as “sign”. Claude showed inconsistent descriptions like “chart,” “hands,” or “blank”, and GPT-4o-mini described them as “signal,” “hand,” or “communication". These models associated the segments with page image rather than each individual segment’s content. GPT-4 was most accurate, usually identifying those squares as irrelevant. These mistakes stemmed from models overemphasizing key objects from the full-page image without critically analyzing segments independently. Repeated exposure to relevant elements in addition to the whole page while batch processing biased the models toward assuming all segments are meaningful. In the few-shot experiment phase, both GPT-4o and GPT-4o-mini, for the same case, correctly identified these empty white squares as irrelevant in dynamic and static settings; hence, providing examples enhanced the assessment. In the dynamic setting, there was only one segment of this case where GPT-4o identified the white square as "hand" while GPT-4o-mini identified it as "signal". Other examples of discrepancies between the ground truth and MLLMs are included in the Appendix (Section A.5).

5 Discussion

In this study, we investigate the effectiveness of MLLMs in identifying relevant segments extracted from storybook illustrations throughout the contextual narrative of the storybooks, particularly in an Arabic-English multimodal context. Our findings offer insights into the current capabilities of MLLMs and the potential integration of MLLMs into digital illustrated storybook tools.

Effectiveness of MLLMs Our results reveal that, although several MLLMs surpassed the baseline approach, performance still fell short of being robust enough for the given task. Among all the models we used, GPT-4o with static few-shot prompting achieved the best results (F1=0.708) and generated descriptions that most closely aligned with the ones assigned by human annotators’ evaluation. This evaluation framework thus provides a structured benchmark for assessing how well MLLMs can handle narrative-driven multimodal tasks which is under-explored in current MLLM research. Our findings provide empirical evidence on whether state-of-the-art MLLMs can accurately identify relevant illustration segments based on a story’s context. While GPT-4o and GPT-4O-mini exhibit the strongest performance, the overall performance remains inconsistent for other MLLMs. Since Claude-3 Opus underperformed significantly, especially in semantic similarity measures, we presume that the model struggles to handle the cross-linguistic and multimodal aspects of the given task, especially where a nuanced contextual inference in Arabic storybooks is required. Similarly, Gemini 2.0 Flash performed marginally better than Claude 3 Opus but demonstrated weaker performance than its predecessor model, Gemini 1.5 Pro in both binary segment relevance classification and semantic similarity. Those findings reveal potential gaps in how these much newer MLLMs handle context-sensitive, narrative-driven tasks in addition to fine-grained reasoning, aligning with concerns about multilingual evaluation in multimodal tasks (Nagi et al., 2024). Based on our observations, most MLLMs rely on global context from entire illustrations or pages, rather than focusing on specific visual segments. This leads to overgeneralization and weak fine-grained reasoning. Using narrative-driven visual interaction as a benchmark, we expose limitations in current MLLM architectures that standard vision-language evaluations overlook. The high class imbalance in our dataset further challenges models, especially in identifying less common but crucial segments—key for applications like interactive storybooks. While some models (e.g., GPT-4o) handled minority segments better, most achieved macro F1-scores at or below 0.5, with Claude 3 dropping below the random baseline at 0.3272. These findings highlight the need for more robust, narrative-aware multimodal training and evaluation. These findings show the need to evaluate MLLMs on narrative-specific and cross-linguistic tasks that reflect real-world diversity in digital storytelling. Inconsistent results from models like Claude 3 and Gemini 2.0 Flash highlight the lack of reliable solutions for context-rich, multimodal reasoning. Current benchmarks overlook the complexities of narrative-driven tasks; thus, future MLLM development must prioritize structured, multimodal benchmarks that support continuity and fine-grained reasoning in storytelling applications.

5.1 Potential for Interactive Digital Storybooks

Our results offer implications for seamlessly integrating interactive segments into digital illustrated storybook pipelines. Before integrating interactive segments from SAM, these segments can first be evaluated through MLLMs in a context-driven process. GPT-4o and GPT-4o-mini's strong segment identification through descriptions capabilities and binary classification can streamline this process while taking the story's narrative and comprehension into account, which enhances engagement in storytelling applications as highlighted by the existing digital storytelling work (Bus et al., 2019; Kamil et al., 2023; Sun et al., 2024). This effort helps in reducing manual annotation efforts required for developing rich and immersive storybook tools.

Future research should focus on finding methods to improve visual-textual reasoning in MLLMs, refining data across different languages, and developing new ways to evaluate narratives for accuracy and context sensitivity. Expanding evaluation to other storytelling contexts, languages, or domains would also deepen our understanding of MLLM capacities.

6 Conclusion and Future work

The study evaluated the ability of MLLMs to classify segment relevance in illustrated storybooks. Among the evaluated models, GPT-4o achieved the best results, and Claude-3 Opus had the lowest performanceThe two-phase evaluation revealed that incorporating few-shot prompting significantly enhances performance over zero-shot baselines. This finding underscores the importance of contextual alignment in example selection and highlights the role of prompt design in guiding multimodal reasoning. Throughout the experiments, we observed that GPT-4o and GPT-4o-mini could integrate both visual and textual cues to make coherent judgments. This opens up opportunities for developing interactive storytelling systems powered by intelligent visual-textual understanding.

The study had limitations, including a small dataset of 32 Arabic-language storybooks, which can affect generalizability across languages and narrative styles. Using Arabic narratives to assess English outputs could also introduce cultural or linguistic bias. Future work should explore prompt engineering, fine-tuning, and larger, more diverse datasets, including English content. Further research could integrate SAM-2 for enhanced segmentation and multimodal understanding. Overall, this benchmark and evaluation framework lays the groundwork for more robust and interactive digital storytelling systems.

7 Limitations

Several key limitations exist in this study. First, our evaluation was conducted using Arabic-language narratives, while model outputs were in English. This cross-linguistic setup may introduce cultural or semantic mismatches that influence model comprehension. Future research could assess whether MLLMs perform better when both input and output share a consistent linguistic and cultural context, such as English-English evaluation. Second, the dataset consisted of 32 digital storybooks tailored for children, which limits generalizability across age groups and genres. Expanding the evaluation to a broader range of digital storybooks targeting different age groups could yield a more holistic understanding of MLLM capabilities. Third, while we highlight the promise of using auto-segmentation tools like SAM, this study did not directly evaluate their integration with MLLMs. Investigating how segmentation quality interacts with downstream reasoning tasks remains a valuable future direction. Additionally, future work can involves training MLLMs specifically on storybook illustrations and related data to assess their performance on complex multimodal tasks which can potentially enhance their effectiveness and usability in interactive digital storytelling applications. Moreover, we only utilized the two best performing models from the zero-shot experiments in the few-shot experiments for conserving computational resources, so the few-shot experiment results cannot be generalized to Claude-3 Opus and the Gemini models. Finally, one methodological constraint involves the instruction framework provided to both human annotators and the models. Despite careful prompt design, it remains challenging to equate human visual interpretation and contextual reasoning with the model's pattern recognition and embedding-based understanding. This mismatch can lead to inconsistencies in relevance judgments. Future research can explore the development of differentiated annotation protocols and evaluation criteria that explicitly account for the distinct cognitive and perceptual mechanisms of humans and machines.

References

Explosion AI. 2020a. spacy english core web model.

Explosion AI. 2020b. spacy models.

Barbaros Bostan and Tim Marsh. 2012. *Fundamentals of Interactive Storytelling*. Yeditepe University, Information Systems and Technologies Department.

Ben Burtenshaw. 2023. Ai stories: An interactive narrative system for children. *AAAI Proceedings*.

Adriana G. Bus, Burcu Sarı, and Zsofia K. Takacs. 2019. *The Promise of Multimedia Enhancement in Children's Digital Storybooks*, pages 45–57. Springer International Publishing, Cham.

Alberto Carraro, Marco Sozzi, and Francesco Marinello. 2023. The segment anything model (sam) for accelerating the smart farming revolution. *Smart Agricultural Technology*, 6:100367.

Marc Cavazza and Fred Charles. 2003. Interactive storytelling: From ai experiment to new media. In *ICEC Proceedings*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, Caifeng Shan, and Ran He. 2024. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *Preprint*, arXiv:2411.15296.

Franca Garzotto, Paolo Paolini, and Amalia Sabiescu. 2010. Interactive storytelling for children. *IDC 2010 Proceedings*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Jiaxing Huang and Jingyi Zhang. 2024. A survey on evaluation of multimodal large language models. *Preprint*, arXiv:2408.15769.

Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongquan Chen, Chaoyu Chen, Sijing Liu, Haozhe Chi, Xindi Hu, Kejuan Yue, Lei Li, Vicente Grau, Deng-Ping Fan, Fajin Dong, and Dong Ni. 2024. Segment anything model for medical images? *Medical Image Analysis*.

Muhammad Nazimuddin Al Kamil, Rita Eka Izzaty, and Nur Patmawati. 2023. Digital picture storybooks, can increase students' self-efficacy and interest in learning? *Jurnal Ilmiah Sekolah Dasar*, 7(1):35–45.

Akrivi Katifori, Manos Karvounis, Vassilis Kourtis, and Sara Perry. 2018. Applying interactive storytelling in cultural heritage: Opportunities, challenges and lessons learned. *ICIDS Proceedings*.

Kenneth Leung. 2022. Micro, macro weighted averages of f1 score - clearly explained.

Zhaoyang Ma, Xupeng He, Shuyu Sun, Bicheng Yan, Hyung Kwak, and Jun Gao. 2023. Zero-shot digital rock image segmentation with a fine-tuned segment anything model. *Earth Science and Engineering, Physical Science and Engineering Division, King Abdullah University of Science and Technology (KAUST)*.

Mary McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, 22:276–82.

Khalil A Nagi, Elham Alzain, and Ebrahim Naji. 2024. Informed prompts and improving chatgpt english to arabic translation. *Andalus Journal of Humanities and Social Sciences*, 11:212–213.

Juri Opitz and Sebastian Burst. 2019. Macro f1 and macro f1. *arXiv preprint arXiv:1911.03347*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Di Qi, Lin Su, Jia Song, Edward Cui, et al. 2020. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint*, 2001.07966.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pam Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*. PMLR. ArXiv:2103.00020.

Mark O. Riedl. 2012. Interactive narrative: A novel application of artificial intelligence for computer games. *AAAI Proceedings*.

Tao Sun, Oliver Liu, JinJin Li, and Lan Ma. 2025. Llava-re: Binary image-text relevancy evaluation with multimodal large language models. *Proceedings of the First Workshop on Evaluation of Multi-Modal Generation*, page 40–51.

Yuling Sun, Jiaju Chen, Bingsheng Yao, Jiali Liu, Dakuo Wang, Xiaojuan Ma, Yuxuan Lu, Ying Xu, and Liang He. 2024. Exploring parent's needs for children-centered ai to support preschoolers' interactive storytelling and reading activities. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW2).

Loes van Druten-Frietman, Heleen Strating, Eddie Denessen, and Ludo Verhoeven. 2016. Interactive

storybook-based intervention effects on kindergartners' language development. *Journal of Early Intervention*, 38(4):212–229.

Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, Yutong Zhang, Zihao Wu, Zhengliang Liu, Tianyang Zhong, Bao Ge, Tuo Zhang, Ning Qiang, Xintao Hu, Xi Jiang, Xin Zhang, Wei Zhang, Dinggang Shen, Tianming Liu, and Shu Zhang. 2024a. A comprehensive review of multimodal large language models: Performance and challenges across different tasks. *Preprint*, arXiv:2408.01319.

Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, et al. 2024b. A comprehensive review of multimodal large language models: Performance and challenges across different tasks. *arXiv preprint*, 2408.01319.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, et al. 2023. Image as a foreign language: Beit pretraining for vision and vision-language tasks. *Proceedings of CVPR 2023*.

Xiaocui Yang, Wenfang Wu, Shi Feng, Ming Wang, Daling Wang, Yang Li, Qi Sun, Yifei Zhang, Xiaoming Fu, and Soujanya Poria. 2023. Mm-bigbench: Evaluating multimodal models on multimodal content comprehension tasks. *Preprint*, arXiv:2310.09036.

Lyumanshan Ye, Jiandong Jiang, Danni Chang, and Pengfei Liu. 2024. Storypark: Leveraging large language models to enhance children story learning through child-ai collaboration storytelling. *arXiv*, 2405.06495.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *Journal of Machine Learning Research*, 24:1–39.

Peng Zhang and Yaping Wang. 2023. Segment anything model for brain tumor segmentation. *School of Electrical and Information Engineering, Zhengzhou University*.

Qingqing Zhu, Benjamin Hou, Tejas Sudarshan Mathai, Pritam Mukherjee, et al. 2024. How well do multimodal llms interpret ct scans? an auto-evaluation framework for analyses. *arXiv preprint*, 2403.05680.

Urszula Świerczyńska Kaczor. 2024. Empirical insights into traditional and ai-enhanced interactive narratives based on children's fables. *Journal of Economics Management*, 46:25–54.

A Appendix

A.1 Final Prompt

Final Prompt:
System Role: You are a helpful storybook reader. You are given context and story text for a page: page_context and page_text. *Prompt*: Based on the context and text of the given story's page, is the image segment extracted important to the context of the story where the segment is the second image and the first image is the page illustration? ONLY respond with True if yes or False if no. And then give ONLY a one ENGLISH word description of the object describing the segment object in the segment after the true or false. if the segment contains more than one object or less than an object (part of the object) consider it as a false and describe it 'nothing'. Do not respond with any explanations. If the object represents nothing important, return the description as 'nothing'. If it represents something relevant to the context, describe something in one word. So ONLY include true or false in your response along with the one word description.

A.2 Storybook Page-level Context Retrieval Prompt

Final Prompt:
System Role: You are a helpful assistant that extracts story context from an image in Arabic. *Prompt*: In Arabic, describe what is happening in this image, considering the context from the previous page: {context}.

A.3 Human Annotator Guidelines

Instructions:
You will be shown a storybook page and an image segment extracted from it. For each segment, perform the following:

Relevance: Mark as *True* if the segment shows a complete, meaningful object relevant to the story context. Mark as *False* if it:

- contains more than one object
- shows only part of an object that and does not suffice as a stand-alone segment
- is not meaningful or not relevant to the story

description: If relevant, write a single English word describing the object. If irrelevant, use the description *nothing*.

Do not include multiple words in the descriptions and format the results into CSV format

A.4 Dynamic Examples Selection

The flowchart in Figure 3, shows the dynamic few-shot example selection process using CLIP em-

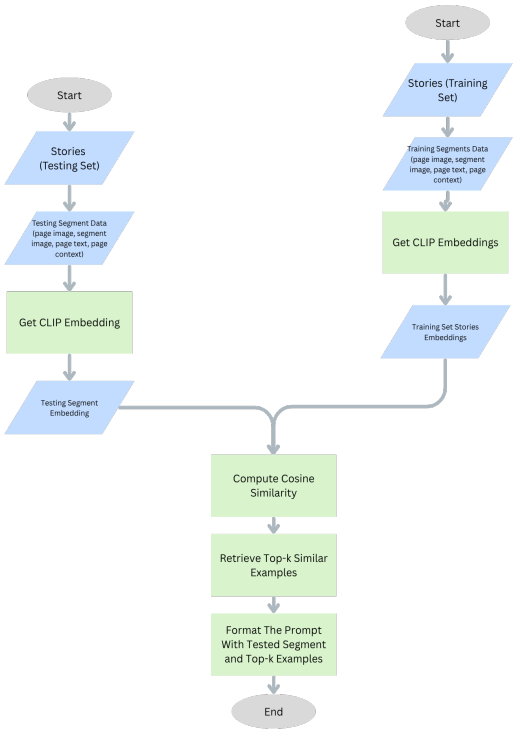


Figure 3: Dynamic Examples Selection Flowchart

beddings. First, the training set embeddings are precomputed once and stored in a JSON file. For each test segment, the pipeline begins by extracting the page image, segment image, text, and context. These inputs are embedded using the CLIP model to generate a test segment embedding. The test embedding is then compared against the precomputed training embeddings using cosine similarity. The top-k most similar training examples are retrieved. These examples, along with the test segment, are formatted into a dynamic prompt. The prompt is used to guide the model's classification. This process repeats for each test segment.

A.5 Error Analysis

To better understand the limitations and behavior of the evaluated multimodal large language models (MLLMs), we conducted a qualitative error analysis on five representative cases of misclassification. Each case highlights a unique challenge in visual reasoning, contextual interpretation, or annotation consistency.

The selected examples illustrate different types of failure modes, such as incorrect relevance judgments, misaligned desciptor predictions, and discrepancies between human and model understanding of visual segments. For each case, we present the full segment context—including the original page image, the extracted segment, textual content, and model predictions alongside an explanation of the observed error and its possible causes.

This analysis aims to shed light on the nuanced performance characteristics of MLLMs and provide insights into how such models may support, or even challenge, human annotation practices.

Case 1

In case 1, as shown in Table 3, the segment was annotated by the human annotator as relevant to the story and identified it as "planet". However, all evaluated models (*GPT-4o* and *GPT-4o-mini*) correctly predicted the segment as not relevant as shown in Table 4. Upon closer inspection, the segment corresponds to the circular window of the

Table 3: Segment Information for Case 1

Story Name:	A strange story in arabic answered
Page: 30	**Segment:** 01
Page Image	**Segment Image**
Page Text	**Page Context**
" قالَ بَيْنَ تِلْكَ النُّجومِ في السَّماءِ سَأكونُ أَعْظَمَ رائِدٍ لِلْفَضاءِ."	في هذه الصورة، نرى رسمة لصاروخ فضائي بني اللون يطير في الفضاء. الصاروخ يبدو وكأنه مصنوع من القماش أو الخشب، مع تفاصيل دقيقة مثل النوافذ المستديرة والخطوط المنحنية. من خلف الصاروخ، يتصاعد دخان أبيض، مما يشير إلى أنه في حالة انطلاق. الخلفية تحتوي على رسومات لنجوم وكواكب صغيرة، مما يعزز جو الفضاء والمغامرة. النص المكتوب في أعلى الصورة يقول" :قال علي: بين تلك النجوم في السماء ...سأكون أعظم رائد للفضاء ."...هذا النص يعبر عن حلم الطفلة بالمغامرة والاستكشاف، ولكن هذه المرة في الفضاء بدلاً من البحر. الصورة تعكس شغف الطفلة بالاستكشاف والمغامرة، سواء كان ذلك في البحار أو في الفضاء، وتظهر طموحها الكبير في أن تصبح رائدة فضاء عظيمة".
Segment Relevance: True	**Segment description:** Planet

rocket, which the annotator had mistakenly identified as a celestial body. This misinterpretation likely stemmed from the segment's shape, color, and position within the illustrated page, which visually resemble a planet, especially when viewed in isolation.

Table 4: Classification Results for Segment Contextual Relevance (Case 1)

Experiment	GPT-4o-mini	GPT-4o
Zero-shot	**FALSE**, nothing	**FALSE**, nothing
Few-Shot (Fixed)	**FALSE**, nothing	**FALSE**, nothing
Few-Shot (Dynamic)	**FALSE**, nothing	**FALSE**, nothing

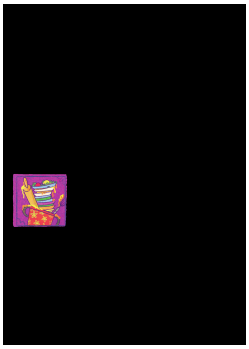
While the annotator judged this object as relevant, assuming it represented a narrative element (a planet supporting the child's dream of space travel), the models correctly identified its true semantic context, a minor structural detail (the rocket window) that is not central to the story's progression. Moreover, the segment is part of a larger object (the rocket), and the prompt explicitly instructed that any part of an object should be considered irrelevant.

This case highlights the sensitivity of vision-language models to object–context relationships and underscores how human biases or misinterpretations in visual descriptions can lead to mismatches in evaluation. It also suggests that model predictions, when systematically consistent, can serve as a valuable tool for flagging ambiguous or potentially incorrect annotations during dataset refinement.

Case 2

In case 2, as shown in Table 5, the human annotator identified the segment as irrelevant and assigned it the description 'nothing', according to the annota-

Table 5: Segment Information for Case 2

Story Name:	Mess in the kitchen
Page: 38	**Segment:** 05
Page Image	**Segment Image**
	
Page Text	**Page Context**
/	"في هذه الصورة، نرى مجموعة من الرسومات الملونة التي تعكس مشاهد مختلفة تتعلق بالمطبخ والأواني المنزلية. الصورة مقسمة إلى مربعات، وكل مربع يحتوي على رسم مختلف. في المربع العلوي الأيسر، نرى كومة من الأطباق المتسخة مع بقايا طعام. بجانبها، في المربع العلوي الأوسط، هناك إبريق شاي مزخرف. في المربع العلوي الأيمن، نرى مجموعة من الأطباق المتكدسة فوق بعضها البعض. في الصف الثاني، المربع الأول يحتوي على رسم لصحن مكسور. المربع الأوسط يحتوي على نمط من النقاط الملونة. المربع الثالث يحتوي على إبريق شاي آخر. في الصف الثالث، المربع الأول يحتوي على رسم لصحن أخضر مقلوب مع بقايا طعام. المربع الأوسط يحتوي على نمط من النقاط الملونة. المربع الثالث يحتوي على كوب مكسور مع سائل مسكوب. في الصف الرابع، المربع الأول يحتوي على رسم لقدر مليء بالأطباق المتسخة. المربع الأوسط يحتوي على نمط من الخطوط الملونة. المربع الثالث يحتوي على رسم لقدر مع طعام بداخله وسكين. في الصف الخامس، المربع الأول يحتوي على رسم لقدر مزخرف. المربع الأوسط يحتوي على رسم لبيضة مكسورة. المربع الثالث يحتوي على نمط من النقاط الملونة. الصورة تعكس مشاهد متنوعة من المطبخ، مع التركيز على الأواني والأطباق المتسخة والنظيفة، مما يعزز فكرة الأعمال المنزلية والتنظيف بعد تناول الطعام."
Segment Relevance: False	**Segment description:** nothing

tion guideline that any segment containing multiple objects should be considered irrelevant. However, all models (GPT-4o and GPT-4o-mini) incorrectly predicted the segment as relevant, assigning various descriptions such as plates, pot, and dishes as shown in Table 6.

Table 6: Classification Results for Segment Contextual Relevance (Case 2)

Experiment	GPT-4o-mini	GPT-4o
Zero-shot	**TRUE**, plates	**TRUE**, dishes
Few-Shot (Fixed)	**TRUE**, pot	**TRUE**, dishes
Few-Shot (Dynamic)	**TRUE**, pot	**TRUE**, pot

Upon visual inspection, the segment clearly contains a cluster of kitchen-related items, including a bowl, plate, rolling pin, whisk, and some leftovers. While these objects are semantically related to the kitchen context, the segment does not isolate a single, clearly identifiable object. The annotation prompt explicitly stated that segments with multiple overlapping objects should be treated as irrelevant, given the difficulty in assigning a specific description and their reduced narrative clarity.

This case highlights a potential failure mode in model behavior, the inability to correctly follow annotation rules that require recognizing the presence of multiple objects and adhering to a "nothing" description policy. The models appear to have focused on semantic plausibility (recognizing kitchenware relevant to the page topic) rather than structural annotation rules, suggesting that they rely more on content familiarity than task-specific constraints.

The example demonstrates the importance of including rule-based reasoning in vision-language modeling and reveals a gap between visual comprehension and annotation policy adherence. It also underscores the necessity for training or prompting strategies that explicitly reinforce domain-specific rules, especially in tasks requiring fine-grained differentiation between object count and semantic relevance.

Case 3

In case 3, as shown in Table 7, the segment was annotated as irrelevant and identified as "nothing", consistent with annotation guidelines, as it represents only the grass floor in the background of the illustrated scene. GPT-4o-mini aligned with the ground truth, correctly classifying the segment as irrelevant. However, GPT-4o misclassified the segment as relevant, assigning descriptions such as feet or butterfly across different prompting strategies as shown in Table 8.

Upon examination, the segment itself does not include any meaningful object central to the narrative. However, GPT-4o appeared to infer the presence of a human character's feet, which are partially visible in the original page but not explicitly part of the extracted segment. This suggests that the model may have leveraged contextual cues from the page image or its prior understanding of human posture and composition to extrapolate beyond the visible content.

Table 7: Segment Information for Case 3

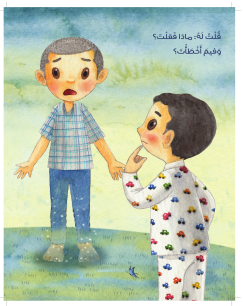
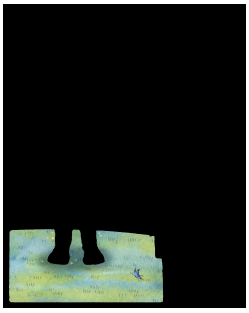
Story Name:	The secret to the world of harmony
Page: 28	**Segment:** 15
Page Image	**Segment Image**
	
Page Text	**Page Context**
"قُلْتُ ماذا فَعَلْتَ؟ وَفِيمَ أَخْطَأْتَ؟"	"في هذه الصورة، نرى طفلين يقفان في مشهد طبيعي. الطفل الأول، الذي يبدو عليه الخوف والقلق، يرتدي قميصًا مخططًا باللون الأزرق وسروالًا أزرق. تعابير وجهه تعكس حالة من الذعر، وفمه مفتوح كما لو كان يصرخ أو يطلب المساعدة. يمكن ملاحظة أن قدميه تبدوان وكأنهما تختفيان أو تتلاشى، حيث تظهر عليهما نقاط لامعة. الطفل الثاني، الذي يرتدي بيجامة بيضاء مزينة برسومات سيارات ملونة، يقف بجانبه ويبدو عليه التفكير أو الحيرة. يضع إصبعه على ذقنه وكأنه يسأل أو يحاول فهم ما يحدث . النص الموجود في أعلى الصورة يقول: "قُلْتُ لَهُ: مَاذَا فَعَلْتَ؟ وَفِيمَ أَخْطَأْتَ؟" مما يشير إلى أن الطفل الثاني يسأل الطفل الأول عن سبب حالته وما الذي فعله ليصل إلى هذه الحالة. الخلفية الطبيعية للمشهد تضفي جوًا من التوتر والتشويق، مما يعزز الإحساس بالخوف والذعر الذي يعيشه الطفل الأول."
Segment Relevance: False	**Segment description:** nothing

Table 8: Classification Results for Segment Contextual Relevance (Case 3)

Experiment	GPT-4o-mini	GPT-4o
Zero-shot	**FALSE**, nothing	**TRUE**, feet
Few-Shot (Fixed)	**FALSE**, nothing	**TRUE**, butterfly
Few-Shot (Dynamic)	**FALSE**, nothing	**TRUE**, feet

This behavior points to a form of overextension in visual reasoning, where the model attempts to "complete" the visual scene based on what it expects rather than what is actually visible. While such inference can be powerful in some applications, it poses a challenge in annotation-driven tasks that require strict attention to segmentation boundaries and adherence to explicit description rules.

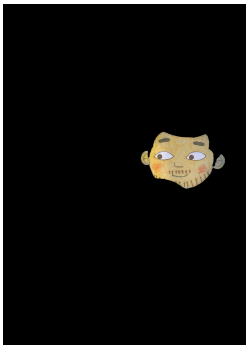
The case reveals an important limitation in evaluation: even highly capable models like GPT-4o may introduce false positives by detecting plausible content that lies outside the designated segment. It underscores the importance of reinforcing spatial precision in prompt design and training, especially in settings where models must operate under localized input constraints.

Case 4

In case 4, as shown in Table 9, the segment was annotated by the human annotator as relevant to the story and identified as "face". During evaluation, both GPT-4o and GPT-4o-mini initially misclassified the segment as irrelevant in the zero-shot and fixed few-shot settings. However, under the dynamic few-shot prompting strategy, both models successfully recognized the segment as relevant and described it more precisely 'father' as shown in Table 10.

The segment visually represents the father's face,

Table 9: Segment Information for Case 4

Story Name:	Tomorrow I will be
Page: 42	**Segment:** 09
Page Image	**Segment Image**
	
Page Text	**Page Context**
"في المَساءِ، جَلَسْتُ في حِضْنِ أبي، وَفي حِضْنِ أمّي جَلَسَتْ عَبِيرُ."	"في هذه الصورة، نرى مشهدًا دافئًا يجمع بين أب وابنه في لحظة حميمية. يجلس الطفل في حضن والده، ويبدو عليهما السعادة والراحة. الطفل يرتدي بيجامة مخططة باللونين الأزرق والأبيض، بينما الأب يرتدي قميصًا أخضر. في الزاوية اليسرى من الصورة، يوجد مصباح مضاء على طاولة صغيرة وردية اللون، وعلى الطاولة يوجد مجهر وزجاجة تحتوي على سائل أزرق. خلفية الصورة مزينة بنمط من الأزهار والأوراق الخضراء، مما يضفي جوًا مريحًا ودافئًا على المشهد. النص المكتوب في أعلى الصورة يقول: "في المَساء، جَلَسْتُ في حِضنِ أبي، وفي حِضنِ أمي جَلَسَتْ عَبِير."، مما يشير إلى أن الطفل يشعر بالأمان والراحة في حضن والديه، وأن هناك أختًا تدعى عبير تجلس في حضن الأم. الصورة تعكس جوًا من الحب والدفء العائلي، حيث يبدو أن الطفل يستمتع بوقته مع والده في المساء، وربما يستعد للنوم بعد يوم طويل ".
Segment Relevance: True	**Segment description:** face

a meaningful narrative element within the page, where the story shows the child nestled in his father's arms. While the human annotator conservatively described the segment as face, the models in the dynamic setting went beyond pure visual identification and inferred the character role using the surrounding visual and textual context, effectively linking the segment with its narrative identity.

Table 10: Classification Results for Segment Contextual Relevance (Case 4)

Experiment	GPT-4o-mini	GPT-4o
Zero-shot	**FALSE**, nothing	**FALSE**, nothing
Few-Shot (Fixed)	**FALSE**, nothing	**FALSE**, nothing
Few-Shot (Dynamic)	**TRUE**, father	**TRUE**, father

This case underlines the semantic sensitivity of large vision-language models when given proper prompting, and emphasizes the importance of aligning model objectives with annotation criteria, especially in tasks where the description traction may vary between annotators and models.

Case 5

In case 5, as shown in Table 11, the human annotator classified the segment as irrelevant and identified it as "nothing". This judgment was consistent with most of the model predictions in all prompt strategies, except GPT-4o-mini, which classified the segment as relevant in both fixed and dynamic few-shot settings and described it as "smoke" as shown in Table 12.

Notably, this description is semantically accurate. The segment is part of a larger illustrated smoke cloud depicted in the full-page image. Although the segment in isolation may not exhibit strong visual features typically associated with smoke, the model's prediction appears to rely on contextual and spatial cues derived from its training or from its ability to reason across the full visual-

Table 11: Segment Information for Case 5

Story Name:	Lily
Page: 26	**Segment:** 15
Page Image	**Segment Image**
Page Text	**Page Context**
"صاحَ ماهذِهِ الرّائِحَةُ؟ كح كح كح إنَّها تَخْنُقُ! أوووووف ما كُلُّ هذا الدُّخانِ؟ ماذا يَحْصُلُ؟"	"في هذه الصورة، نرى طفلاً في حالة من الفزع والدهشة بسبب انفجار مفاجئ. الطفل يبدو عليه القلق والخوف، حيث يضع يده على وجهه ويبدو عليه الارتباك. في الخلفية، نرى غيمة كبيرة من الدخان الأخضر تنتشر في الغرفة، مع بعض العناصر الغريبة والألوان الزاهية التي تملأ المشهد. يبدو أن هناك تفاعلاً كيميائياً أو انفجاراً صغيراً قد حدث، مما تسبب في هذه الفوضى والقلق. النص المكتوب في أعلى الصورة يصف حالة الفزع التي أصابت الطفل من صوت الفَرْقَعة والانفجار المدوي، والذي شكل غيمة دخان كبيرة خضراء، وملأت الغرفة برائحة كريهة. الصورة تعكس حالة من الفوضى والارتباك، وربما تكون نتيجة لخطأ ما في تجربة علمية، مما يتماشى مع السياق السابق حيث كانت ديما تعتذر عن خطأ ارتكبته. يبدو أن الطفل في الصورة قد تأثر بما حدث، مما يعزز جو التوتر والقلق في المشهد."
Segment Relevance: False	**Segment description:** nothing

Table 12: Classification Results for Segment Contextual Relevance (Case 5)

Experiment	GPT-4o-mini	GPT-4o
Zero-shot	**FALSE**, nothing	**FALSE**, nothing
Few-Shot (Fixed)	**TRUE**, smoke	**FALSE**, nothing
Few-Shot (Dynamic)	**TRUE**, smoke	**FALSE**, nothing

textual context.

This case demonstrates the importance of refining annotation guidelines to accommodate fragmented but semantically valid visual elements, and suggests that model outputs in such situations should not be penalized without considering contextual correctness.