

Evaluating multi-modal LLMs on historical book illustrations

Jiaming Jiang

École Polytechnique Fédérale de Lausanne

Andrew Piper

McGill University

Abstract

Historical book illustrations are culturally significant multimodal artifacts that pose distinctive challenges for computational analysis, including sparse visual cues, historical reproduction techniques, and material degradation. We present the first task-specific evaluation of multimodal large language models (MLLMs) on historical book illustrations, testing five contemporary models on 1,969 children’s book illustrations across a range of visual labeling exercises and caption generation tasks. Frontier models performed strongly against human consensus annotations on concrete tasks, particularly illustration and setting detection, and mirrored human-level disagreement on entity, object and action identification. All model families struggled with accurate emotional intensity estimation. GPT and Gemini produced the strongest captions, with greater than 70% of images exhibiting no factual errors. Finally, frontier models closely recovered human distributions at the corpus level for major visual categories suggesting their appropriateness for use in large-scale studies.

1 Introduction

Illustrated books have played a fundamental role in storytelling for centuries. As complex multimodal artifacts, they transmit cultural values through both their textual and visual design (Nodelman, 1988; Pereira, 2008; Ionescu and Schellenberg, 2015; Kümmerling-Meibauer and Meibauer, 2017; Dowd, 2026). Despite their longstanding study across a variety of research domains, large-scale computational analysis of narrative book illustrations still remains rare.

Recent advances in MLLMs have opened new possibilities for the automated analysis of cultural heritage images (Smits and Wevers, 2023). To support this work, researchers have developed specialized datasets such as SemArt (Garcia and Vogiatis, 2018), ArtPedia (Stefanini et al., 2019),



Figure 1: Overview of the visual dimensions evaluated in the PICTURING STORIES dataset.

ArtCap (Lu et al., 2022), and DEArt (Reshetnikov et al., 2022), which pair artworks with expert descriptions, iconographic labels, or object annotations. These resources have enabled the training of captioning models that incorporate art-historical knowledge (Bai et al., 2021) and the evaluation of zero-shot classification and metadata prediction by general-purpose LLMs (Tojima and Yoshida, 2025; Rei et al., 2023).

Historical book illustrations, however, differ from the canonical artworks represented in most visual-cultural-heritage datasets (Mak, 2011; Berne, 2024). Unlike paintings, which are generally treated as self-contained visual artifacts, illustrations are embedded within larger narrative or instructional texts and often convey meaning through sparse visual cues. Many historical illustrations are also low-resolution or monochromatic, were produced using period-specific technologies such as woodcuts and engravings, and exhibit degradation such as foxing or bleed-through. Interpreting them therefore requires models to recognize subtle, noisy visual evidence and historical iconography that may be underrepresented in modern training

data and art-historical datasets.

Children’s illustrated books are an especially consequential test case. Emerging as an increasingly central cultural form during the eighteenth century, they have long promoted reading while transmitting social and moral values (Nodelman, 1988; Nikolajeva, 2002; Kümmerling-Meibauer, 2023; Grenby and Immel, 2009). Illustrated books remain important to children’s education and language development today (Montag et al., 2015; Wang and Shao, 2025). Historical children’s illustrations thus constitute both a significant cultural archive and a demanding domain in which to assess multimodal model understanding.

Against this background, we evaluate five contemporary MLLMs on PICTURING STORIES,¹ a dataset of manually annotated historical children’s book illustrations covering the visual-narrative tasks shown in Figure 1. To our knowledge, this is the first task-specific evaluation of MLLMs on historical book illustrations, an important but understudied domain of cultural heritage.

The study makes three contributions. First, it evaluates model performance across classification, multi-label, ordinal, action-recognition, and caption-generation tasks, distinguishing image-level reliability from the recovery of corpus-level visual distributions. Second, it identifies systematic, task- and category-specific patterns of error alongside qualitative examples. Third, we release a complete replication package containing the 1,969 evaluation images, human annotations, raw model outputs, exact prompts and model versions, and code for reproducing all tables and figures.²

2 Prior Work

2.1 NLP for Art Understanding

Recent research has begun to address interpretive challenges in visual art, moving beyond surface recognition to extract culturally and narratively salient information in multiple aspects (Smits and Wevers, 2023). Entity-level understanding—including character (Madhu et al., 2019), icons (Cetinic, 2021; Springstein et al., 2024), pose (Schneider and Vollmer, 2023), and affections (Achlioptas et al., 2021; Chen et al., 2024; Yan et al., 2022)—has been explored through person identification

in religious art and general object grounding in fine-grained captioning (Cetinic, 2021; Springstein et al., 2024).

Additional research considers artistic style classification and captioning (Ruta et al., 2022; Li et al., 2022), narrative and commonsense reasoning (Zellers et al., 2019), and methods for disentangling literal and symbolic meaning (Zhang et al., 2024a). Domain knowledge is incorporated through structured metadata (Bai et al., 2021; Jiang et al., 2024) or retrieval-augmented architectures, offering interpretive scaffolding in low-resource domains. Such systems have also been applied to support cultural heritage tasks, where combining visual and textual reasoning enables non-experts to access symbolically rich content in art (Zhang et al., 2024b; Arita et al., 2025).

2.2 NLP for Illustration Understanding

While research on cultural heritage images has largely focused on paintings, artifacts, or photographs, there is growing interest in book and print illustrations. Early work addressed the visual segmentation of historical page layouts—detecting elements such as footnotes, tables, and figures (Piper et al., 2020; Abuelwafa et al., 2018)—as well as the interpretation of archival newspaper illustrations (Fyfe and Ge, 2018). Recent studies have examined the sociocultural representation of race and gender in modern children’s book illustrations (Adukia et al., 2022; Szasz et al., 2022), alongside investigations into visual heritage materials including photographs (Arnold and Tilton, 2024a; Maksimova et al., 2024), maps (Mahowald and Lee, 2024), and graphical elements of scientific publishing (Lang et al., 2023). Arnold and Tilton (2024b) propose a framework for applying MLLMs to heritage collections, generating captions and embedding-based similarity scores to enable cluster-based exploration and explainable recommendations. Together, these efforts point toward the potential of multimodal models to make illustrations more searchable, interpretable, and analyzable at scale.

3 Dataset

We evaluate model performance using the historical portion of PICTURING STORIES, which contains 27,889 images drawn from 2,827 books published between 1728 and 1932, sourced from the Internet

¹https://github.com/dot-txtlab/picturing_stories

²https://github.com/dot-txtlab/picturing_stories_MLLM_evaluation

Task	Type	Description
Illustration	Binary (2)	Is there an illustration on the page? (<i>Yes/No</i>)
Setting	Single-label (3)	Where is the image set? (<i>Indoors, outdoors, or ambiguous</i>)
Characters	Multi-label (10)	Who is present (<i>man, woman, child, animal, etc.</i>)
Objects	Multi-label (7)	Main objects (<i>nature, architecture, toys, etc.</i>)
Actions	Multi-label (21)	Behaviors (<i>sitting, standing, embracing, etc.</i>)
Emotional Intensity	Ordinal (5)	Rating from 1 (<i>calm</i>) to 5 (<i>intense</i>)
Caption	Free-text	1–2 sentence description

Table 1: Annotation tasks and label types used for evaluation. Numbers in parentheses indicate how many options were available for selection by task. Full task prompts and answer options are shown in Table A1.

Archive’s Children’s Library.³ 6,981 volunteers produced over 700,000 annotations on the data. The collection is predominantly English-language and primarily reflects British and North American publishing.

For our evaluation, we construct a random sample of 1,969 pages with illustrations from the full corpus. Each illustration is annotated for visual–narrative elements that include characters, objects, settings, actions, emotional intensity, and free-text captions (Table 1). These annotations support multiple evaluation settings with varying levels of complexity, including single-label classification, multi-label prediction, ordinal judgment, and caption generation.

4 Evaluating MLLM Performance

We evaluate five MLLMs under zero-shot and one-shot prompting, spanning proprietary frontier systems and open-weight alternatives: GPT-4.1 (gpt-4.1-2025-04-14), Gemini 2.5 Flash, Claude Sonnet 4.5 (claude-sonnet-4-5-20250929), Qwen3-VL-8B-Instruct, and LLaVA-1.6-Vicuna-13B. These models represent different contemporary multimodal architectures and levels of accessibility. Because of input-length constraints, LLaVA-1.6-Vicuna-13B was evaluated only in the zero-shot condition. Exact prompts, model identifiers, inference dates, and raw outputs are included in the replication package.

Models are scored against the human majority-vote or consensus annotation for each image (Table 2, Panel A). Separately, the dataset reports agreement among human annotators (Table 2, Panel B). We treat these quantities as empirical human-agreement references rather than as competitor performance scores.

Evaluation metrics follow the structure of each

task in Table 1. For binary and single-label classification—illustration presence and setting—we report exact-match accuracy. For the character and object tasks, we report the mean Jaccard index between the predicted and consensus label sets.

For action prediction, we ask which actions a particular entity in an image is performing. To avoid ambiguity when several instances of the same entity type were present, we restricted evaluation to images in which the human consensus annotation identified exactly one instance of the queried entity type; an image could still contain other characters or objects. One query was issued per image, with the entity type supplied to the model and the model asked to select all applicable actions from a fixed list of 21 options. The predicted action set was then compared with the human-annotated action set for that same entity using micro-averaged F1.

For emotional intensity, we report the mean absolute distance between the model rating and the human rating on the 1–5 scale, for which lower values indicate closer agreement.

Caption evaluation has three components: (1) semantic similarity to human captions, measured using both BERTScore and bge-m3 cosine similarity, with a within-image human–human baseline; (2) human validation of hallucinated and missing content; and (3) qualitative analysis of representative errors across models.

4.1 Categorical Annotations

Across categorical tasks, frontier models achieved high agreement with the human consensus on concrete visual classifications, particularly illustration presence and setting, but showed greater divergence on multi-label and subjective ordinal judgments. We examine these patterns below, with a focus on one of the best performing models, Gemini-2.5-Flash with zero-shot prompting.

³<https://archive.org/details/iacl>

Table 2: Model performance against human consensus annotations (Panel A) and inter-annotator agreement among humans (Panel B). Model values and human agreement values are reported separately because they do not always use the same metric. Human agreement is therefore an interpretive reference rather than a competitor score.

Panel A: Model performance against human consensus annotations						
Model	Illustration Exact Match (%)	Setting Exact Match (%)	Characters Jaccard	Objects Jaccard	Intensity Distance ↓	Actions Micro-F1
Claude-0	97.6	83.7	0.60	0.59	0.87	0.57
Claude-1	97.9	84.4	0.63	0.59	0.87	0.58
Gemini2.5-0	96.8	88.1	0.67	0.56	0.88	0.57
Gemini2.5-1	98.0	87.3	0.65	0.59	0.95	0.62
GPT-0	97.6	85.4	0.64	0.48	0.81	0.61
GPT-1	98.0	85.0	0.62	0.50	0.82	0.62
Qwen-0	97.8	83.0	0.69	0.53	0.83	0.49
Qwen-1	97.6	83.2	0.68	0.56	0.79	0.52
LLaVA-0	89.0	78.8	0.49	0.47	0.91	0.35

Panel B: Human inter-annotator agreement				
Task	Full Agreement (%)	Majority Agreement (%)	Jaccard	Ordinal Distance ↓
Illustration (2)	85.4	—	—	—
Characters (10)	52.5	—	72.9	—
Setting (3)	75.3	—	95.5	—
Objects (7)	33.7	—	60.8	—
Emotional intensity (1-5)	15.3	—	68.9	0.67
Actions (21)	24.7	—	52.9	—

Notes. In Panel A, models are scored against the human majority-vote or consensus annotation. Suffixes 0 and 1 denote zero-shot and one-shot prompting, respectively. Jaccard values are means across images. Intensity is the mean absolute distance from the mean human rating on the 1–5 scale; lower values indicate closer agreement. Actions are evaluated with micro-averaged F1 after restricting the analysis to queried entity types that have exactly one instance in the human annotation. Panel B parentheses refer to the number of classes per task.

4.1.1 Illustration Detection

Illustration detection was the most accurate task in the evaluation. Although binary illustration detection may initially appear straightforward, historical book pages contain decorative elements that complicate the distinction between illustrations and other forms of page design, such as book stamps, publisher symbols, photographs, ornamental page design, page headers, and small frontispiece ornaments (Figure 2). Measured against the human consensus annotation, all models frontier models achieved more than 96% exact-match accuracy, with the exception of LLaVa.

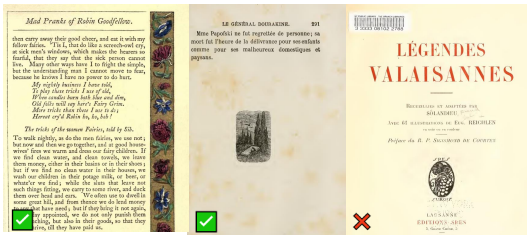


Figure 2: Examples of ambiguous illustration types from left to right: decorative ornamentation (-), small vignette (+), title-page publisher logo (-). ✓ / × indicate whether models detected the proper answer.

4.1.2 Setting Detection

Setting detection was also highly accurate, particularly for clearly indoor or outdoor scenes. For Gemini 2.5 Flash (zero-shot) only 12 of 935 evaluated images involved a direct reversal between an indoor human label and an outdoor model prediction, or vice versa. Most remaining errors involved inappropriate use of the “Both / Ambiguous” category, distributed relatively evenly across images labeled indoors and outdoors by the human consensus.

As Figure 3 illustrates, these disagreements generally involved either boundary cases, such as an image containing a doorway or window that connected indoor and outdoor spaces, or images with very sparse contextual information. These cases suggest that the primary difficulty was not recognizing clearly depicted settings but determining which contextual region should govern the label.

4.1.3 Character and Object Detection

The number of characters predicted per image was closely aligned between Gemini 2.5 Flash under zero-shot prompting ($M = 1.80$, $Mdn = 2$) and the human annotations ($M = 1.76$, $Mdn =$



Figure 3: Examples of setting ambiguity involving a boundary between indoor and outdoor spaces (left) and sparse contextual information (right).

2). At the category level, several small tendencies appeared in the uncorrected tests: Gemini over-represented Boys ($OR = 1.25$, $p = .031$, $p_{BH} = .070$) and Mythical / Imaginary Beings ($OR = 1.60$, $p = .022$, $p_{BH} = .065$), and under-represented Women ($OR = 0.80$, $p = .020$, $p_{BH} = .065$) and Men ($OR = 0.85$, $p = .042$, $p_{BH} = .075$). None of these substantive character tendencies survived Benjamini–Hochberg correction and they should therefore be interpreted as suggestive rather than statistically significant. Gemini also used the catch-all Other label substantially less often than human annotators ($OR = 0.20$, $p_{BH} < .001$), which appears to reflect a difference in label use rather than a content-specific character bias.

Object annotations showed greater divergence. Gemini predicted more object categories per image ($M = 2.05$, $Mdn = 2$) than the human annotations ($M = 1.57$, $Mdn = 1$; Wilcoxon effect size = 0.51). Gemini over-represented Toys ($OR = 2.97$, $p_{BH} = .001$), Household Items and Tools ($OR = 1.34$, $p_{BH} = .003$), and Reading and Art ($OR = 1.43$, $p_{BH} = .042$), while under-representing Architecture ($OR = 0.80$, $p_{BH} = .011$). All four effects survived Benjamini–Hochberg correction. Nature, Vehicles, and Other showed no significant differences after correction.

4.1.4 Action Detection

Gemini predicted substantially more actions per entity ($M = 3.43$, $Mdn = 3$) than human annotators ($M = 1.91$, $Mdn = 2$, Wilcoxon effect size = 0.75). Gemini over-predicted *looking*, *talking*, and *working*, and under-predicted passive states like *sleeping*, *sitting*, and *standing*. Given the prevalence of passive states like *sitting* and *standing*, we investigate these errors further. We find that a majority (> 60%) of misalignments occur when either humans or the model has outputted multiple other actions, suggesting the model is not adding these

states once other kinds of actions are identified, whereas human annotators will more often include them. In the small number of cases where there was a single action identified by both humans and the model, *looking* was used in place of *sitting* or *standing*, once again suggesting an issue of default choice behavior.

4.1.5 Emotional Intensity Estimation

Emotional intensity produced the greatest divergence between models and human annotations. For Gemini 2.5 Flash under zero-shot prompting, model ratings were significantly lower than the mean human rating (Wilcoxon signed-rank $p = 1.9 \times 10^{-56}$; matched-pairs rank-biserial $r = -0.73$). The model and human rating distributions also differed significantly in shape (Kolmogorov–Smirnov $p = 2.1 \times 10^{-45}$). Gemini assigned a rating of 1 or 2 to 64% of images, compared with 41% of human ratings, and its distribution was substantially more right-skewed toward low arousal (skewness = 0.86 versus 0.07 for humans; Figure 4). The same low-arousal tendency appeared across the other models. This direction is consistent with prior evidence that language models can underrate emotional intensity relative to human annotators (Bojić et al., 2025).

Manual inspection however provides more nuance on these results. Focusing on labels that diverged by two points or more suggests that when human labels = 3 and Gemini = 1 that the actions were more often than not serene (boys sitting around a campfire, a woman standing and drinking from a jug). Occasionally, humans identified subtle gestures that indicated some action (as in hand movements that suggested talking justifying the higher rating). However, where human labels equalled 4 or 5 and Gemini was -2 points away this almost universally indicated model failure to understand the intensity of the scene (e.g. a boat in a storm).

4.2 Caption Evaluation

We evaluate generated captions along three complementary dimensions. First, we measure semantic similarity between model- and human-generated captions using BERTScore and bge-m3 cosine similarity. Second, we use human judgments to assess hallucinated and missing visual content and, conditional on an error, its severity and visual centrality. Third, we qualitatively examine representative errors across multiple models. These analyses cap-

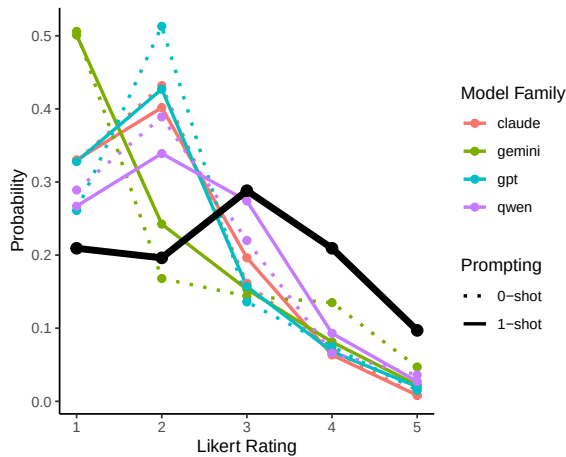


Figure 4: Distribution of emotional arousal ratings across humans and models. Black line is the observed human distribution.

ture different aspects of caption quality: semantic similarity does not necessarily imply factual correctness, while factual-error counts do not fully capture whether a caption preserves the image’s broader content.

4.2.1 Semantic Similarity Analysis

We first evaluate caption similarity on the 639 images for which human captions were collected: 364 images have two human captions and 275 have three. For each image, we compute the mean BERTScore between a model-generated caption and all available human captions. We also compute within-image human–human similarity to contextualize the natural variation among human descriptions (Figure 5).

GPT-4.1 and Gemini 2.5 Flash achieved the highest BERTScores, with one-shot prompting consistently improving similarity over zero-shot prompting. Claude, Qwen, and LLaVA formed a lower-performing group. When we supplement BERTScores with bge-m3 cosine similarity, we see similar score compression but closer human-level performance. The observed human level was 0.657 with models ranging from 0.611 (Claude) to 0.661 (GPT).

Model captions were generally longer than the human captions: human captions averaged 20.5 words, compared with 33–35 words for the evaluated models Table 4. Greater length often produced more detailed descriptions, but also created additional opportunities to introduce unsupported entities, attributes, actions, or story information.

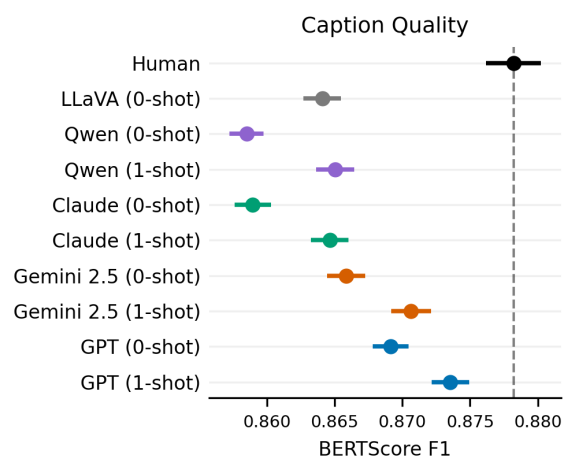


Figure 5: Caption quality measured by BERTScore F1 (mean $\pm$ 95% CI). The dashed vertical line indicates human inter-annotator agreement.

4.2.2 Factual Alignment Analysis

To evaluate factual alignment, we selected 295 images and presented multiple model-generated captions for each image to annotators using the same platform as PICTURING STORIES. The resulting dataset contains 2,950 image–caption instances and 8,587 individual judgments. Every caption instance was evaluated by at least two annotators and most were evaluated by three.

Annotators counted two forms of factual error: *hallucinations*, defined as entities or actions described in the caption but absent from the image, and *missing items*, defined as salient entities or actions visible in the image but omitted from the caption. Counts were recorded on a 0–4+ scale.

For caption instances containing an error, annotators additionally classified whether the affected content occurred in the *foreground* or *background* and whether the error was *major* or *minor*. Figure 6 shows the distribution of error severity conditional on an error having been identified.

Annotators reached majority agreement on the exact number of errors for 78.4% of hallucination judgments and 81.7% of missing-item judgments. Despite this substantial raw agreement, ordinal Krippendorff’s alpha was low ($\alpha = 0.049$ and $\alpha = 0.017$, respectively), reflecting the strong concentration of ratings at zero and the high variation in the fine-grained error-count annotations.

As shown in Table 3, Gemini and GPT produced the strongest factual-alignment results. Across their prompting conditions, 64.4–71.5% of captions received zero hallucination flags, while 69.8–78.9%

received zero missing-item flags. Nevertheless, among the errors that annotators did identify, approximately half or more were classified as major, and most involved foreground content (Figure 6).

Model	Hallucinations (% = 0)	Missing (% = 0)	Hallucinations (M)	Missing (M)
Gemini-0	68.5	77.6	0.44	0.27
Gemini-1	71.5	69.8	0.40	0.42
GPT-0	66.1	78.9	0.51	0.28
GPT-1	64.4	71.2	0.51	0.45
Qwen-0	56.3	70.2	0.70	0.44
Qwen-1	54.6	64.4	0.70	0.53
Claude-0	48.1	53.6	0.86	0.71
Claude-1	48.8	55.9	0.87	0.68
LLaVA-0	35.3	45.1	1.19	0.90

Table 3: Factual correctness by model. For each model and prompting condition, we report the percentage of captions with zero hallucinated/missing elements and the mean error count for each category.

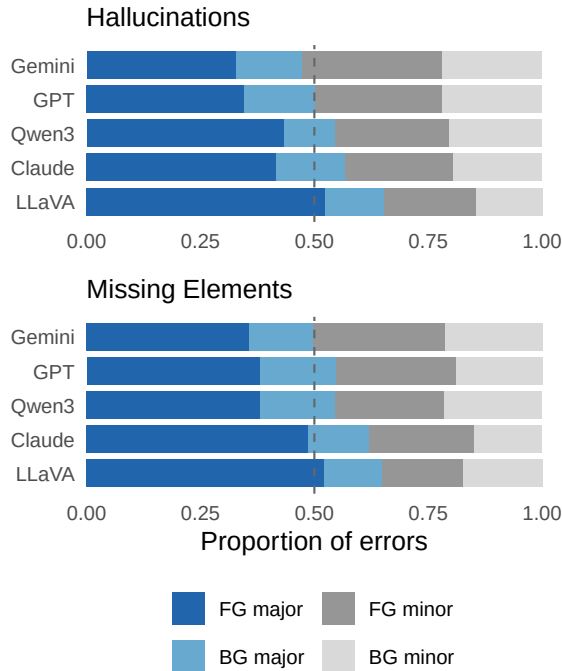


Figure 6: Distribution of error types across our background/foreground and major/minor schema for both hallucinations and missing elements

4.2.3 Qualitative Analysis of Caption Errors

For our qualitative analysis of model captions, we once again focus on Gemini 2.5-Flash, this time with one-shot prompting. We find that only 26 out of 295 illustrations (ca. 9%) were labeled as containing at least one hallucination, missing element or both and was labeled as “major” by at least two annotators. We did not find errors concentrated

in a particular style or type. Image styles in this group included both black and white and color illustrations, indoor and outdoor settings, and primary entities of animals or human beings. Some images do appear visually ambiguous to the human eye, while others were quite straightforward. In two cases, captions were completely misaligned with the image, suggesting a potential pipeline error.

In terms of **missing elements**, these could include missing persons, often children, as in a caption where a man is described as seated and looking at a woman “kneeling beside him” (*Le Général Douraki*, p. 180). In fact, there is a child kneeling before him and a woman standing beside him.

This positioning problem also occurred in several images with **hallucinations**, as for example, one in which a young girl is said to be seated at a table when in fact she is seated on the ground and the table is in the background (*The Light Hearted Girl*, p. 54). In another, a man is said to be reclined on a blue chair when it is in fact a swaddled baby lying on a woman’s lap with a blue skirt (*A Picture Book of Merry Tales*, p. 104, Figure 7, left).

Overall, we find that in just about all cases except total misalignment the errors are largely singular and do not disrupt a general understanding of the image’s overall contents. Most often, principal figures and principal backgrounds are adequately accounted for even when there are errors, suggesting a limitation to the human annotation task. In many cases, the errors are arguably minor even if labeled as major by annotators. For example, in the above-mentioned illustration with the mis-observed baby, all other elements of the scene are accurately described. Similarly, in another (from *Mother Goose’s Nursery Rhymes*) where annotators felt there was a central, foreground error, we see that a lady cat is said to be stirring a black pot when in fact she is fanning the flames of a big kettle (Figure 7, right).

4.3 Global Error Analysis

We next examine whether model error varies historically and whether noisy image-level predictions nevertheless preserve broad corpus-level patterns. We address two questions. First, using Gemini 2.5 Flash under zero-shot prompting, we test whether model–human misalignment varies with publication date. Second, we compare the zero-shot outputs of all five models with human annotations, measuring alignment in aggregate category proportions for Characters, Objects, and Setting and

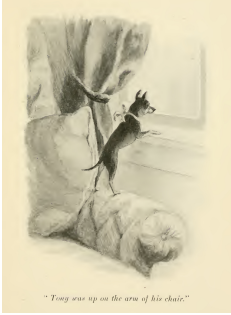
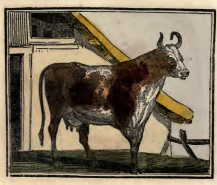
Higher Similarity (0.91)	Lower Similarity (0.87)
 <p><i>"Tony was up on the arm of his chair."</i></p> <p>Annotator: A small dog with a ribbon around its neck stands on the arm of a chair to look out of a window.</p> <p>Gemini: A small, dark-colored dog, possibly a Chihuahua named Tony, stands on the arm of a cushioned chair near a window with curtains partially drawn. The dog appears to be looking out the window.</p>	 <p><i>This is the cow with a crumpled horn, that tossed the dog, that worried the cat,</i></p> <p>Annotator: A cow with a crumpled horn standing outside a house.</p> <p>Gemini: An illustration depicts a brown and white cow with a crumpled horn standing in a barnyard next to a building. The image is a visual representation of a line from the cumulative children's rhyme "The House That Jack Built."</p>

Table 4: Examples of human-model image-caption pairs with associated BERTScores.

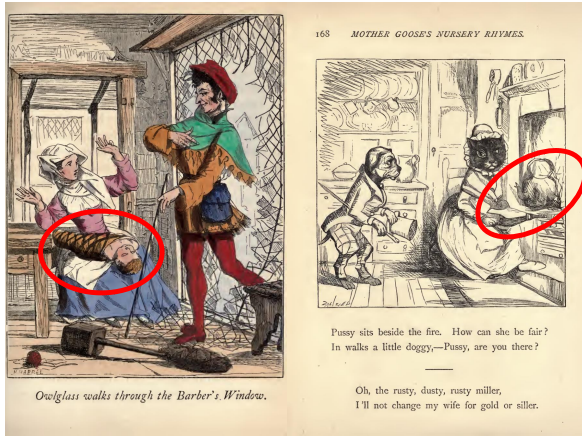


Figure 7: Example of hallucinations by Gemini 2.5-Flash on two illustrations. Left: A Picture Book of Merry Tales (1860). Right: Mother Goose's Nursery Rhymes (1870).

image-level rating correspondence for Emotional Intensity.

RQ1: Are older images more likely to contain errors? We first test whether publication date predicts model-human misalignment. Misalignment is calculated for each image as the mean Jaccard dissimilarity across the Character, Object, and Setting tasks. Contrary to our expectation that errors would be more likely with older, harder to interpret material, error was broadly stable across the nineteenth century and increased slightly in the later decades of the corpus (Figure 8). Mean misalignment was 0.24 for images published before 1850 and 0.32 for images published after 1900. Publication decade showed only a small positive association with error (Spearman's $\rho = 0.135$), and a linear model explained little of the overall variance ($R_{\text{adj}}^2 = 0.02$).

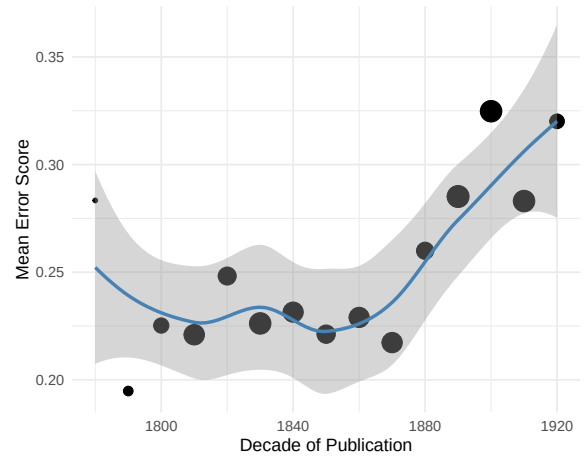


Figure 8: Mean Gemini zero-shot model-human misalignment by publication decade, with LOESS smoothing. Point size indicates the number of images in each decade. Misalignment is the mean Jaccard dissimilarity across Character, Object, and Setting predictions.

RQ2: Do model estimates approximate human estimates despite image-level error? We next test whether model outputs preserve broad corpus-level patterns even when individual predictions are noisy across all five models under zero-shot prompting. For Characters, Objects, and Setting, we calculate the proportion of all labels assigned to each category by the model and by human annotators. We then compare the resulting category distributions using Pearson's r and mean absolute difference (MAD). For Emotional Intensity, we calculate the image-level Pearson correlation between each model rating and the mean human rating (Table 5).

For Gemini, aggregate alignment was high for Setting ($r = 0.99$), Characters ($r = 0.95$, MAD = 0.023), and Objects ($r = 0.98$, MAD = 0.023). The Character and Setting results generalized

Table 5: Model–human correspondence for aggregate category proportions under zero-shot prompting. Character, Object, and Setting values compare corpus-level category distributions. Emotional Intensity reports the image-level correlation between model ratings and mean human ratings and is therefore not an aggregate distributional correlation. Missing human labels are excluded.

Model	Characters r	Characters MAD	Objects r	Objects MAD	Setting r	Intensity r
Gemini 2.5 Flash	0.95	0.023	0.98	0.023	0.99	0.67
GPT-4.1	0.96	0.021	0.71	0.050	0.99	0.65
Claude Sonnet 4.5	0.94	0.021	0.99	0.015	0.97	0.64
Qwen3-VL-8B	0.96	0.023	0.91	0.041	0.94	0.58
LLaVA-1.6	0.87	0.031	0.94	0.032	0.90	0.15

across the other frontier models: Gemini, GPT, Claude, and Qwen all achieved correlations of at least $r = 0.94$ for these tasks. Object alignment was less consistent. Claude closely reproduced the human object distribution ($r = 0.99$), whereas GPT diverged more substantially ($r = 0.71$, $\text{MAD} = 0.050$). Emotional Intensity remained the weakest dimension. Image-level correlations ranged from $r = 0.58$ to 0.67 across the four frontier models and fell to $r = 0.15$ for LLaVA.

5 Conclusion

This paper provides the first task-specific evaluation of contemporary MLLMs on historical book illustrations. Across five models and a diverse set of visual–narrative tasks, we find that frontier models achieved high agreement with human consensus annotations on concrete classifications, particularly illustration presence and setting, while performance became less consistent for multi-label classifications while remaining relatively proximate to comparable human-level disagreement given the large choice space. The emotional intensity task exhibited the lowest overall performance, clearly diverging from human judgments. Gemini and GPT performed best overall, Claude and Qwen generally formed a second group, and LLaVA lagged on most measures.

Caption generation showed similar quality. GPT and Gemini produced the strongest semantic-similarity scores to human levels and received no factual-error flags in $>70\%$ of cases. When errors were identified, however, they were often labeled as foreground entities or actions and frequently judged to be major. Nevertheless, our qualitative analysis complicated the severity of these assessments, emphasizing that while there were often singular hallucinations or missing items these rarely impacted the overall understanding of the image.

Thus captions cannot be relied on to provide 100% accuracy for diagnostic or search purposes, but they for our best models they are reasonable descriptive proxies of the underlying images.

Despite these local error rates, we find that aggregate assessments by our best performing model (Gemini 2.5 Flash) closely reproduced human category distributions for Characters, Setting, and objects (all $r > 0.90$). Leading MLLMs may therefore be useful for estimating selected broad visual patterns across large historical collections even when individual predictions remain noisy, providing an important resource for the field of cultural analytics.

Limitations

This study has several limitations. First, our evaluation is based on a sample of illustrations from PICTURING STORIES, which is drawn primarily from the Internet Archive’s Children’s Library. The collection is therefore shaped by the holdings, digitization practices, and metadata of a specific cultural heritage repository, and is predominantly composed of English-language books from British and North American publishing contexts. As a result, our findings should not be taken as evidence of model performance on historical illustration traditions more broadly.

Second, our benchmark relies on a structured annotation schema designed to make visual–narrative features comparable across a large collection. This design enables systematic evaluation, but necessarily simplifies the richness of illustrated storytelling. Categories such as character type, object type, setting, action, and emotional intensity capture important dimensions of narrative imagery, but they do not exhaust the interpretive features that may matter to scholars of illustration, children’s literature, or book history. In particular, our tasks do not directly evaluate composition, style, sym-

bolism, interdependence between text and image, or historically specific iconographic conventions.

Third, our use of human annotations as an evaluation baseline introduces both strengths and constraints. Human agreement provides a more appropriate benchmark than assuming a single objective label for interpretive visual tasks, but consensus labels can obscure disagreement, ambiguity, or minority interpretations. This is especially important for subjective categories such as emotional intensity, where variation among annotators may reflect genuine differences in interpretation rather than annotation noise.

Fourth, our model comparison is limited to five contemporary MLLMs and two prompting conditions. The results therefore represent a snapshot of current model capabilities rather than a stable, more universal estimate of multimodal performance in this domain. Different image preprocessing pipelines, prompt designs, decoding parameters, fine-tuning strategies, or retrieval-augmented methods could produce different results. Because proprietary model systems may also change over time, exact replication of these results may be difficult in the future.

Finally, our evaluation focuses primarily on visual content visible on the page. Many book illustrations derive part of their meaning from their relation to surrounding text, page layout, sequence, genre, and publication context. By evaluating images in isolation, we test an important but partial form of visual understanding. Future evaluations that incorporate textual context and sequential relations across pages will be necessary to assess whether MLLMs can support fuller forms of narrative interpretation.

Statement on AI Use

AI use was limited to writing improvement, formatting, and figure generation. All scientific claims and conclusions represent the original intellectual work of the authors, and all AI-assisted content was carefully reviewed, verified, and, where necessary, substantially modified by the authors.

Acknowledgments

Funding was generously provided by the Social Science and Humanities Research Council of Canada.

References

- Sherif Abuelwafa, Sara Zhalepour, Ehsan Arabnejad, and Mohamed Mhiri et. al. 2018. Detecting footnotes in 32 million pages of ecco. *Journal of Cultural Analytics*, 3(1).
- Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J Guibas. 2021. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11569–11579.
- Anjali Adukia, Callista Christ, Anjali Das, and Ayush Raj. 2022. Portrayals of race and gender: Sentiment in 100 years of children’s literature. In *Proceedings of the 5th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*, pages 20–28.
- Takaya Arita, Wenxian Zheng, Reiji Suzuki, and Fuminori Akiba. 2025. [Assesing llms in art contexts: Critique generation and theory of mind evaluation](#).
- Taylor Arnold and Lauren Tilton. 2024a. [Automated image color mapping for a historic photographic collection](#). In *Proceedings of the Computational Humanities Research Conference (CHR)*, volume 3834, pages 37–47, Aarhus, Denmark. CEUR Workshop Proceedings.
- Taylor Arnold and Lauren Tilton. 2024b. Explainable Search and Discovery of Visual Cultural Heritage Collections with Multimodal Large Language Models. In *Proceedings of the Computational Humanities Research Conference*, volume 3834 of *CEUR Workshop Proceedings*, pages 559–574.
- Zechen Bai, Yuta Nakashima, and Noa Garcia. 2021. Explain me the painting: Multi-topic knowledgeable art description generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5422–5432.
- Debbie Berne. 2024. *The design of books: An explainer for authors, editors, agents, and other curious readers*. University of Chicago Press.
- Ljubiša Bojić, Olga Zagovora, Asta Zelenkauskaitė, Vuk Vuković, Milan Čabarkapa, Selma Veseljević Jerković, and Ana Jovančević. 2025. Comparing large language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm. *Scientific reports*, 15(1):11477.
- Eva Cetinic. 2021. Iconographic image captioning for artworks. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part III*, pages 502–516. Springer.
- Tianwei Chen, Noa Garcia, Liangzhi Li, and Yuta Nakashima. 2024. [Retrieving emotional stimuli in artworks](#). *Proceedings of the 2024 International Conference on Multimedia Retrieval*.

- DB Dowd. 2026. *Reading Pictures: A History of Illustration*. Princeton University Press.
- Paul Fyfe and Qian Ge. 2018. [Image analytics and the nineteenth-century illustrated newspaper](#). *Journal of Cultural Analytics*, 3(1).
- Noa Garcia and George Vogiatzis. 2018. How to read paintings: semantic art understanding with multi-modal retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0.
- Matthew Orville Grenby and Andrea Immel. 2009. *The Cambridge companion to children’s literature*. Cambridge University Press.
- Christina Ionescu and Renata Schellenberg. 2015. *Book Illustration in the Long Eighteenth Century: Reconfiguring the Visual Periphery of the Text*. Cambridge Scholars Publishing.
- Yanbei Jiang, Krista A Ehinger, and Jey Han Lau. 2024. Kale: an artwork image captioning system augmented with heterogeneous graph. *arXiv preprint arXiv:2409.10921*.
- Bettina Kümmerling-Meibauer. 2023. Picturebooks. In *The Routledge Companion to Children’s Literature and Culture*, pages 95–105. Routledge.
- Bettina Kümmerling-Meibauer and Jörg Meibauer. 2017. Picturebooks and cognitive studies. In *The Routledge companion to picturebooks*, pages 391–400. Routledge.
- Sarah Lang, Bernhard Liebl, and Manuel Burghardt. 2023. [Toward a computational historiography of alchemy: Challenges and obstacles of object detection for historical illustrations of mining, metallurgy and distillation in 16th–17th century print](#). In *Proceedings of the Computational Humanities Research Conference (CHR)*, volume 3558, pages 29–48, Paris, France. CEUR Workshop Proceedings.
- Haotang Li, Sheng Jun Guo, Kailin Lyu, Xiao Yang, Tianchen Chen, Jianqing Zhu, and Huanqiang Zeng. 2022. [A challenging benchmark of anime style recognition](#). *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4720–4729.
- Yue Lu, Chao Guo, Xingyuan Dai, and Fei-Yue Wang. 2022. Artcap: A dataset for image captioning of fine art paintings. *IEEE Transactions on Computational Social Systems*, 11(1):576–587.
- Prathmesh Madhu, Ronak Kosti, Lara Mührenberg, Peter Bell, Andreas K. Maier, and Vincent Christlein. 2019. [Recognizing characters in art history using deep learning](#). *Proceedings of the 1st Workshop on Structuring and Understanding of Multimedia Heritage Contents*.
- Jamie Mahowald and Benjamin Charles Germain Lee. 2024. [Integrating visual and textual inputs for searching large-scale map collections with clip](#). In *Proceedings of the Computational Humanities Research Conference (CHR)*, pages 528–547, Aarhus, Denmark.
- Bonnie Mak. 2011. *How the page matters*. University of Toronto Press.
- Erika Maksimova, Mari-Anna Meimer, Mari Piirsalu, and Priit Järv. 2024. Viability of Zero-shot Classification and Search of Historical Photos. In *Proceedings of the Computational Humanities Research Conference*, volume 3834 of *CEUR Workshop Proceedings*, pages 1242–1258.
- Jessica L Montag, Michael N Jones, and Linda B Smith. 2015. The words children hear: Picture books and the statistics for language learning. *Psychological science*, 26(9):1489–1496.
- Maria Nikolajeva. 2002. *The rhetoric of character in children’s literature*. Scarecrow Press.
- Perry Nodelman. 1988. *Words about pictures: The narrative art of children’s picture books*. University of Georgia Press.
- Nilce M Pereira. 2008. Book illustration as (intersemiotic) translation: Pictures translating words. *Meta*, 53(1):104–119.
- Andrew Piper, Chad Wellmon, and Mohamed Cheriet. 2020. The page image: Towards a visual history of digital documents. *Book History*, 23(1):365–397.
- Luis Rei, Dunja Mladenec, Mareike Dorozynski, Franz Rottensteiner, Thomas Schleider, Raphaël Troncy, Jorge Sebastián Lozano, and Mar Gaitán Salvatella. 2023. Multimodal metadata assignment for cultural heritage artifacts. *Multimedia Systems*, 29(2):847–869.
- Artem Reshetnikov, Maria-Cristina Marinescu, and Joaquim More Lopez. 2022. Deart: Dataset of european art. In *European conference on computer vision*, pages 218–233. Springer.
- Dan Ruta, Andrew Gilbert, Pranav Aggarwal, Naveen Marri, Ajinkya Kale, John Briggs, Chris Speed, Hailin Jin, Baldo Faieta, Alex Filipkowski, Zhe Lin, and John P. Collomosse. 2022. [Stylelabel: Artistic style tagging and captioning](#). In *European Conference on Computer Vision*.
- Stefanie Schneider and Ricarda Vollmer. 2023. [Poses of people in art: A data set for human pose estimation in digital art history](#). *ArXiv*, abs/2301.05124.
- Thomas Smits and Melvin Wevers. 2023. A multimodal turn in digital humanities. using contrastive machine learning models to explore, enrich, and analyze digital visual historical collections. *Digital Scholarship in the Humanities*, 38(3):1267–1280.

- Matthias Springstein, Stefanie Schneider, Javad Rahnama, Julian Stalter, Maximilian Kristen, Eric Müller-Budack, and Ralph Ewerth. 2024. Visual narratives: Large-scale hierarchical classification of art-historical images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7220–7230.
- Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Massimiliano Corsini, and Rita Cucchiara. 2019. Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain. In *Image Analysis and Processing-ICIAP 2019: 20th International Conference, Trento, Italy, September 9–13, 2019, Proceedings, Part II 20*, pages 729–740. Springer.
- Teodora Szasz, Emileigh Harrison, Ping-Jung Liu, Ping-Chang Lin, Hakizumwami Birali Runesha, and Anjali Adukia. 2022. Measuring representation of race, gender, and age in children’s books: Face detection and feature classification in illustrated images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 462–471.
- Tatsuya Tojima and Mitsuo Yoshida. 2025. Zero-shot classification of art with large language models. *IEEE Access*.
- Zhenlin Wang and Yihan Shao. 2025. Picture book reading improves children’s learning understanding. *British Journal of Developmental Psychology*, 43(1):12–35.
- Jianhao Yan, Wenmin Wang, and Cheng Yu. 2022. Affective word embedding in affective explanation generation for fine art paintings. *Pattern Recognition Letters*, 161:24–29.
- Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731.
- Mingliang Zhang, Xia Hou, Yujing Yan, and Meng Sun. 2024a. [Decoupleclip: A novel cross-modality decouple model for painting captioning](#). *Electronics*.
- Wei Zhang, Wong Kam-Kwai, Biying Xu, Yiwen Ren, Yuhuai Li, Minfeng Zhu, Yingchaojie Feng, and Wei Chen. 2024b. Cultiverse: Towards cross-cultural understanding for paintings with large language model. *arXiv preprint arXiv:2405.00435*.

Appendix A Annotation Task Workflow

Task	Category	Question	Possible Answers
T0	Illustration	Is there an illustration on this page?	<i>Yes, No</i>
T1	Characters	Who are the characters on the page? Check all that apply.	<i>Man, Woman, Girl, Boy, Animal, Animate Object (e.g., a teddy bear or spoon who talks), Mythical / Imaginary Being, Other, None</i>
T2	Setting	Where is the scene set?	<i>Indoors, Outdoors, Both / Ambiguous</i>
T3	Objects	What are the main objects in the scene? Check all that apply.	<i>Nature (trees, flowers, lakes), Architecture (buildings, walls, furniture), Vehicles (cars, boats, planes), Toys, Reading and Art (books, musical instruments, paintings), Household items and tools (pots, pans, hammers, sticks), Other, None</i>
T4	Intensity	How emotionally intense is this scene? What is the energy like?	<i>1 - Very calm (still, contemplative, restful), 2 - Mildly calm, 3 - Neutral, 4 - Mildly intense, 5 - Very intense (frantic, tense, active)</i>
T5	Actions	Choose an entity type in the image. What behaviors do you see?	<i>sitting, standing, walking, talking, looking, pointing, gesturing, holding hands or embracing (someone), being embraced, holding (something not someone), giving, playing, working, eating/drinking, reading, listening, crying, sleeping, moving actively (running, jumping, dancing, climbing, flying, etc), animal noise, violent action</i>
T6	Captions	Please write a caption for this illustration. Make sure to include key characters, objects, actions, and aspects of the setting. Write 1-2 full sentences and be as objective as possible.	<i>"Three small children rest on hay: a boy holding a rake stands between a seated girl holding a flower and another boy lying down." "Three children play in the field, one is laying, one is sitting and holding a flower, the third is standing and holding a stick."</i>

Table A1: The Picturing Stories annotation framework across seven separate tasks including questions posed to annotators and possible answers provided.

Appendix B MLLM Prompts

This appendix provides the exact prompts used in our experiments. For each task we report both the zero-shot prompt and the one-shot example provided to the model. For one-shot prompting, a fixed demonstration pair (example image and output) was shown before the target image.

Task 1: Scene Analysis — Zero-shot Prompt

You are given a children’s book illustration. Analyze the image and answer the following 5 questions based on the visual content only. Respond using the exact JSON format provided.

1. Is there an illustration on this page? Answer 'Yes' or 'No'.
2. Who are the characters on the page? Choose all that apply from: ['Man', 'Woman', 'Girl', 'Boy', 'Animal', 'Animate Object', 'Mythical / Imaginary Being', 'Other', 'None']
3. Where is the scene set? Choose one: 'Indoors', 'Outdoors', or 'Both / Ambiguous'
4. What are the main objects in the scene? Choose all that apply from: ['Nature', 'Architecture', 'Vehicles', 'Toys', 'Reading and Art', 'Household items and tools', 'Other', 'None']
5. How emotionally intense is the scene? Rate from 1 (very calm) to 5 (very intense).

Respond in this exact JSON format (no explanations or extra commentary):

```
{
  "illustration_present": "",
  "characters": [],
  "scene_setting": "",
  "objects": [],
  "emotional_intensity":
}
```

Task 1: Scene Analysis — One-shot Example

Prompt: same as the zero-shot prompt above.

{Example image provided.}

Correct JSON output for the example image:

```
{
  "illustration_present": "Yes",
  "characters": ["Woman", "Boy"],
  "scene_setting": "Indoors",
  "objects": ["Household items and tools"],
  "emotional_intensity": 2
}
```

The model was then asked to perform the same task for the target image.

Task 2: Action Classification — Zero-shot Prompt

You are given a children’s book illustration.

The entity type is: <ENTITY_TYPE>.

There is exactly ONE such entity in the image. Your task is to identify ONLY the behaviors performed by that entity.

What behaviors do you see?

Choose ALL that apply from the list below. Use the option text exactly (spelling).

Options:

- <action option 1>
- <action option 2>
- <action option 3>
- ...

Return VALID JSON only in this format:

```
{  
  "actions": []  
}
```

Task 2: Action Classification — One-shot Example

Prompt: same as the zero-shot prompt above.

{Example image provided}

Correct JSON output for the example image:

```
{  
  "actions": ["Holding something (not someone)", "Sitting", "Working"]  
}
```

The model was then asked to identify actions for the target image.

Task 3: Caption Generation — Zero-shot Prompt

Please write a caption for this children's book illustration. Include the key characters, important objects, actions, and aspects of the setting. Write 1–2 full sentences, be descriptive and as objective as possible, and do not invent story details that are not clearly visible in the image.

Task 3: Caption Generation — One-shot Example

Prompt: same as the zero-shot prompt above.

{Example image provided}

Example caption:

Inside a room with a table and fireplace, a woman holds a broom while a boy stands in front of her with one arm raised. Another child looks into the room from behind an open door; and a birdcage hangs on the wall.

The model was then asked to generate a caption for the target image.