LLMs Behind the Scenes: Enabling Narrative Scene Illustration

Anonymous ACL submission

Abstract

Generative AI has established the opportunity to readily transform content from one medium to another. This capability is especially powerful for storytelling, where visual illustrations can illuminate a story originally expressed in text. In this paper, we focus on the task of narrative scene illustration, which involves automatically generating an image depicting a scene in a story. Motivated by recent progress on text-to-image models, we consider a pipeline that uses LLMs as an interface for prompting text-to-image models to generate scene illustrations given raw story text. We apply variations of this pipeline to a prominent story corpus in order to synthesize a novel dataset of illustrations. We conduct a human annotation task to obtain pairwise quality judgments for these illustrations. Through our analysis of this dataset and experiments modeling illustration quality, we demonstrate that LLMs can effectively verbalize scene knowledge implicitly evoked by story text. Moreover, this capability is impactful for generating and evaluating illustrations.

1 Introduction

Observing the transformation of a story from one modality to another (e.g. from text to visual form) can make the story more compelling to its audience. Recent advances in generative AI have enabled this kind of cross-modal transformation to be performed automatically. In particular, text-to-image models allow people to create visual material using natural language alone. Current interaction with these models typically involves users envisioning a particular visual target and then crafting language that realizes that target. Many stories that currently only exist in text form would be well-suited for transfer to an image modality, but the text itself of these stories may not be naturally optimal for directly applying text-to-image models. Given their demonstrated success at meta-prompting (e.g. Zhou et al., 2023), LLMs may be able to interface with story text to synthesize suitable prompts for text-to-image models towards this end. The cooperation between these AI models would make it possible to automatically generate illustrations for any given text-based story.

In this paper, we comprehensively exemplify this approach to visual transfer of story text. Generating illustrations for stories encompasses many challenges, some of which pertain not to the relation between the text and illustrations, but the relation between the illustrations themselves for different scenes in the story (in particular, the visual consistency between depictions of story elements). In this work, because we are focused on the first set of challenges concerning the illustrations' alignment with the story text, we scope the task to focus on individual *scene illustrations*. In particular, we consider scene-level units of stories (*fragments*). We present a pipeline (outlined in Figure 1) that generates a scene illustration given a fragment in its story context. Through systematic variation and ablation of the components of this pipeline, we produce a novel dataset of scene illustrations for fragments in a notable story corpus. We conduct a human annotation task to obtain relative quality judgments for pairs of scene illustrations.

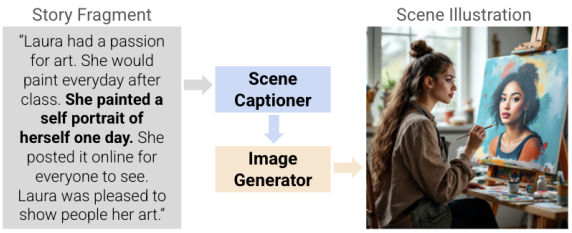

Figure 1: Scene illustration pipeline

We leverage this dataset to establish the impressive capability of LLMs to explicate visual knowledge of narrative scenes by inferring it directly from story text, without any visual input. We demonstrate this through two findings. First,

we show that LLMs are an effective interface for transforming story text into prompts that direct text-to-image models to produce illustrations. Second, we show that LLMs can verbalize scene characteristics in a way that is useful for evaluating the quality of illustrations. In particular, we demonstrate an approach to predicting human-favored illustrations among pairs in our presented dataset, through which we apply scene descriptions given by LLMs as evaluation criteria for scoring illustrations. The success of this approach relative to a criteria-ablated baseline further suggests the utility of LLMs for explicating scene knowledge that is implicitly conveyed by story text.

1.1 Contributions

This paper makes the following contributions[1]:

- We define and motivate the task of narrative scene illustration in relation to existing research on visually aligned storytelling.
- We demonstrate a pipeline for producing scene illustrations for any given story text. The pipeline components are fully interchangeable and can be used with any LLM and text-to-image models.
- We apply our pipeline to an existing story corpus in order to synthesize a dataset of scene illustrations. We elicit human judgments of the relative quality of pairs of illustrations in this dataset.
- Through analysis of these quality judgments, we show that LLMs are an effective interface between story text and text-to-image models in facilitating scene illustration.
- We assess an approach to predicting illustration quality that involves applying LLM verbalizations of scene characteristics as evaluation criteria. We discuss the evaluation results as additional evidence that LLMs can explicate visual scene knowledge inferred from story text.

2 Background and Related Work

Image-Aligned Story Data There are several datasets that pair narrative text with corresponding images. Most of these have been developed to support research on visually grounded story generation, where the task is to write a story given a sequence of images (Halperin and Lukin, 2023; Huang et al., 2016; Hong et al., 2023). The reverse-direction task of generating a sequence of images to depict a story has been termed *story visualization* (Li et al., 2019; Tao et al., 2024). While the same datasets used for visual storytelling are applicable to story visualization, most research pursuing the latter has adapted data from video captioning datasets. Distinct frames of the videos are sampled as static images, while the captions corresponding to these frames are designated as the story text (Li et al., 2019; Maharana and Bansal, 2021; Maharana et al., 2022). As pointed out in various work (Hong et al., 2023; Liu et al., 2024; Lukin et al., 2018), a limitation of these datasets is that caption text does not necessarily have narrative qualities, since the task of describing a sequence of images is not the same as telling a story. The dataset we present in this paper is unique in that it uses existing stories to derive images, rather than the reverse.

Multimodal Storytelling Systems In addition to datasets, there are increasing demonstrations of story visualization systems, as well as systems that generate story text and images in parallel, i.e. multimodal story generation (An et al., 2024; Koh et al., 2023; Singh et al., 2023; Wan et al., 2024; Yang et al., 2024). While some models applied to these use cases have been trained end-to-end on the specialized datasets described above (Feng et al., 2023; Maharana and Bansal, 2021; Tao et al., 2024), researchers have also begun to leverage generically pretrained models to expand the scope of these systems to open-domain storytelling (de Lima et al., 2024; Gong et al., 2023; Soumik Rakshit, 2024). We follow suit in leveraging a plug-and-play pipeline for scene illustration.

Meta-Prompting for Text-to-Image Models One challenge with using generic models for story visualization is that the story text itself is not necessarily an optimal prompt for text-to-image models. In particular, this text tends to be underspecified in details like the physical appearance of story elements (e.g. entities and locations), which are important characteristics specified in text-to-image prompts (Maharana et al., 2022). Users of these models who have become skilled in writing prompts have done so largely through an iterative process of observing what prompt language yields desirable images (Don-Yehiya et al., 2023). Even with this skill, significant effort is required to manually compose a prompt that captures the intended visual features of the scene corresponding to a story fragment. Following the paradigm of meta-prompting (e.g. Zhou et al., 2023), there is a variety of research on automated prompt optimization for text-to-image models (Brade et al.,

[1] All data and code available at: withheld/during/review

2023; Feng et al., 2024; Hao et al., 2023; Wang et al., 2024), some of which establishes the effectiveness of LLMs in facilitating this process (Lian et al., 2024). Accordingly, recent story visualization work has used LLMs as an interface for deriving text-to-image prompts from story text. In particular, Gong et al. (2023) and He et al. (2024) instructed GPT-4 to transform a story into a series of scene-level prompts intended as input to text-to-image models. It is presumed that these synthesized prompts are more visually descriptive than the story text and thus produce better images, but this has not been empirically validated. Thus, we address this opportunity in our work.

LLMs for Image Evaluation Assessing the degree of semantic alignment between images and text is a prominent research endeavor, which has primarily involved measuring their similarity when projected into a shared embedding space (e.g. Hessel et al., 2021). Because of their capacity for visually descriptive language, even unimodal (text-only) LLMs can contribute to this endeavor. Several works have demonstrated the utility of LLMs for zero-shot visual recognition tasks. For instance, LLM-generated visual descriptions can operate as feature or class detectors for multimodal models (Li et al., 2023; Maniparambil et al., 2023; Menon and Vondrick, 2023; Pratt et al., 2023). This line of research has recently extended to using LLMs for evaluating text-to-image model output. In particular, Lu et al. (2023) facilitated this using an LLM to compute similarity between visual descriptors detected in a generated image and the text prompt used to generate it. Following the same evaluation objective, Hu et al. (2023) and Lin et al. (2025) used LLMs to decompose the prompt into questions to be answered by applying a visual question-answering model to the generated image, with image quality indicated by the rate of answers matching information in the prompt. We similarly examine eliciting visual knowledge from LLMs as a strategy for text-to-image evaluation. Encouraged by recent demonstrations of LLM-based evaluation in multimodal story generation (An et al., 2024), we pursue this method for evaluating scene illustrations.

Criteria-based Evaluation with LLMs In NLP, criteria is a means of anchoring evaluation to certain objectives (Yuan et al., 2024). With the rapidly expanding LLM-as-a-judge paradigm, this has evolved to the point where LLMs are not just applying human-authored criteria to assess text, but are also generating their own criteria (Cook et al., 2024). We examine LLMs' capacity to generate evaluation criteria for the scene illustration task.

3 Scene Illustration Pipeline

We first outline the high-level components[2] of the illustration pipeline in this section, before describing their application in the next section.

Story Fragmentation In our work, we consider a *scene* to be an abstract unit of a story that can be distinctly illustrated by a single image. The story text that aligns to a scene is referred to as a *fragment*. Thus, the first step of producing a scene illustration is to identify its source fragment. Recent work has validated the use of LLMs for the related task of segmenting events in narrative text (Michelmann et al., 2025). Accordingly, we utilize an LLM for this fragmentation task, by instructing it to explicitly annotate the boundaries of all fragments in a given story. Table A.14 shows the prompt we provide to the LLM to facilitate this, where the input contains the story text and the LLM is expected to generate the same text with brackets demarcating the left and right boundaries of each fragment, as demonstrated by the exemplars. We parse this output with a simple regular expression to gather the list of fragments.

Scene Descriptions Once a fragment is identified, the fragment with its story context can then be mapped to a *scene description*. A scene description is a verbalization of what should be illustrated in the image corresponding to the fragment. This verbalization serves as the input to the text-to-image model used to produce the scene illustration. In the simplest case, the scene description could be the fragment itself, or the fragment delimited within its corresponding story. However, as identified in §2, this may not specify enough information to the text-to-image model regarding the visual appearance of story elements referenced in the fragment. Thus, we assess using an LLM to transform a fragment alongside its story context into a scene description. We use the term *scene captioner* to refer to an LLM's function when prompted to generate a scene description. Table A.15 shows the prompt we use to enable this.

[2]We ran all model components using APIs, which we specify here for each model. Unless otherwise indicated, we used the default inference parameters defined by the model's API.

Image Generation As mentioned, the scene descriptions are the inputs to a text-to-image model, referred to here as an *image generator*. While we use the term 'illustration' to describe the end-to-end process that yields an image depicting a scene, the output of this process (i.e. the image generator output) is also called an *illustration*.

4 Scene Illustration Dataset

4.1 Story Text

Seeking out a story corpus suitable for the scene illustration task, we ultimately selected the well-studied ROCStories corpus (Mostafazadeh et al., 2016) based on some key considerations. In particular, these English-language stories were authored to adhere to basic narrative structure in a tightly length-constrained format. In particular, each story consists of five sentences conveying "a causally (logically) linked set of events involving some shared characters". Thus, we can expect that stories are composed of distinct fragments that are each appropriately visualized as a scene illustration. Moreover, the stories are narrations of everyday experiences that can be interpreted according to commonsense knowledge. This knowledge is general enough it is likely to be familiar to the model components of our illustration pipeline.

4.2 Phase 1 Pipeline Details

We applied the pipeline outlined in §3 to produce an initial set of scene illustrations, which we refer to as *Phase 1* data. As inputs to the pipeline, we used the first 50 stories in the ROCStories dev set.[3]

Fragmentation We divided these stories into fragments as described in §3, using CLAUDE-3.5[4] as the LLM, which has displayed notable storytelling-related capabilities (e.g. Mazur, 2025). As shown in Table A.7, this resulted in 206 total fragments across all 50 stories, an average of 4.12 per story. §A.1.1 presents some additional analysis regarding these fragments.

Scene Descriptions We ran the procedure described in §3 to generate scene descriptions for the fragments, using CLAUDE-3.5 as the scene captioner. We refer to these outputs as CAPTION scene descriptions. An example is shown in Table 1, with additional examples included in Table A.16. As outlined in Table 2, CAPTION is one of three scene description types we consider for Phase 1. We compare CAPTION with baseline scene descriptions that consist of the raw story text without any LLM transformation. In the first baseline case, we simply use the original fragment by itself as a scene description, which we refer to as FRAGMENT. The obvious limitation of FRAGMENT is that it ablates any scene-relevant information referenced by the fragment in the surrounding story context. Since this contextual information is accessible when the CAPTION is generated, we designed a comparable baseline scene description that integrates the story context, referred to as CTX-FRAGMENT. As Table 2 shows, CTX-FRAGMENT is formatted as an instruction to take the entire story into account when illustrating the target fragment.

Image Generation We then applied two image generators[5] to generate images using the scene descriptions as prompts. In particular, we used Midjourney v6.1, denoted here as MJ-6.1 (Midjourney, 2024), and FLUX-1[pro], denoted here as FLUX-1-PRO (Black Forest Labs, 2024a). We selected these image generators because they topped the Artificial Analysis Image Arena Leaderboard at the time of Phase 1 in August 2024. This leaderboard captures the relative ELO score (Boubdir et al., 2023) of text-to-image models based on pairwise human judgments regarding how well images from different models reflect the input prompt. Table A.16 includes examples of generated illustrations.

4.3 Phase 1 Annotation Task

Illustration Pairs Our primary objective for Phase 1 was to assess the effectiveness of the LLM scene captioner (i.e. CAPTION scene descriptions) in generating illustrations relative to generating them directly from the raw story text. To address this, we randomly sampled 1384 pairs of illustrations belonging to the same story fragment, where one illustration used CAPTION as the scene description, while the other used one of the baseline scene descriptions, FRAGMENT or CTX-FRAGMENT. This sampling resulted in some pairs where the illustrations used the same image gen-

[3]The dev and test items in ROCStories are actually designated as the Story Cloze Test, where items have a specific format: each story consists of four sentences plus two alternative fifth sentences, where one is the 'correct' story ending and the other is the 'incorrect' ending. For each item, we discarded the incorrect ending and appended the correct ending after the initial four sentences to form a single five-sentence story.

[4]Specifically, claude-3-5-sonnet-20240620, which we ran via the Anthropic API

[5]With exception to Midjourney, we ran all image generation models via the Replicate API.

Fragment (within story)	CAPTION
Alice was getting married in a few weeks. One night, her mother called and she forgot to call her back. Her mother left an angry message on her phone. She threatened not to come to the wedding. **Alice called her mother and apologized profusely.**	A young woman with a worried expression sits on a couch, holding a phone to her ear. She's gesticulating with her free hand, appearing to speak emphatically. In the background, a wedding dress can be seen hanging on a closet door. The room is dimly lit, suggesting it's evening, and there's a notepad with wedding plans visible on a nearby coffee table.

Table 1: An example of a CAPTION scene description generated by CLAUDE-3.5 in Phase 1

Type	**Format**
CAPTION	LLM output of scene captioning prompt (Table A.15)
FRAGMENT	`{{fragment}}`
CTX-FRAGMENT	"Consider this story: [`{{story}}`] Based on this context, illustrate this fragment of the story: [`{{fragment}}`]"

Table 2: Types of scene descriptions for Phase 1

erator and others that used different image generators. Ultimately there were 1183 unique illustrations contained in these 1384 pairs. Additional data statistics for Phase 1 including the specific distribution of pair types are provided in Table A.8.

Task Design We designed an annotation task to assess the relative quality of the two illustrations in each pair. In judging a pair, human annotators were shown the full story with the target fragment for that scene underlined, along with the two alternative images. As shown in Figure A.3, annotators were instructed to select the image that was "the better visualization of the underlined fragment". Note that scene descriptions were not shown to annotators, since their judgment of illustration quality should be anchored to the original fragment. Annotators could express uncertainty by selecting "I can't decide which is better". We implemented the UI for this task using POTATO (Pei et al., 2022).

Procedure We deployed the task on Prolific to obtain annotators. English proficiency was the only requirement for participation. We sought 2 annotators to judge each illustration pair. Each participant judged between 33 and 74 pairs (median=47), plus 3 "attention check" items where one illustration in the pair was replaced with one for a different story, making it trivially easy which image to select. Participants were paid $6 for an expected completion time of 30 minutes. We filtered out participants who did not pass all of the attention check items. Ultimately, 62 participants completed the task, with 59 passing all attention checks. This resulted in a total of 2768 responses for the 1384 pairs, where each item received a response from 2 annotators.

4.4 Phase 1 Annotation Results

Inter-annotator Agreement Given the annotated pairs resulting from §4.3, we computed the inter-annotator agreement of which illustration was selected as the better one in each pair. We did this using an *uncertainty-weighted* variation of Cohen's Kappa score (Cohen, 1960), which we abbreviate here as κ_u. This variation considers that response disagreements arising from one annotator expressing uncertainty (i.e. selecting "I can't decide") should be weighted less heavily than disagreements where the two annotators each select a different illustration as better. In κ_u, disagreements arising from uncertainty are down-weighted by a factor of 0.5. As indicated in Table A.8, the overall κ_u for all 1384 items was 0.447, which can be classified as moderate agreement (Landis and Koch, 1977). §A.1.2 provides a finer-grained analysis of agreement for different categories of pairs.

Win Rates for Scene Description Types To determine whether using an LLM as a scene captioner helps illustration quality, we counted how often the favored illustration was associated with each scene description type, i.e. each type's *win rate*. Table 3 shows the win rate for CAPTION illustrations when paired with FRAGMENT and CTX-FRAGMENT illustrations. This win rate is represented as the percentage of responses in which annotators selected the CAPTION illustration as better among all responses for each respective set of pairs. In both cases, the CAPTION is significantly[6] better: it has an overall win rate of ≈78% against FRAGMENT and ≈75% against CTX-FRAGMENT. Table A.17 further examines the win rates for pairs that used the same image generator, verifying that CAPTION is equally favorable regardless of which image generator was used. This validates the importance of

[6] Statistical significance was computed using a one-sample binomal test at $\alpha = 0.05$ to determine if the win rate was higher than that expected by chance, where chance is defined as $(1 - \#ties/\#responses)/2$

the scene captioner in the pipeline: it verbalizes information that is important for imagining how a story fragment should be visually illustrated as a scene. Illustrations are more successful when this information is explicated to the image generator.

Scene Description Pair	# Pairs	Caption %
CAPTION vs. FRAGMENT	680	78.1 (1360)
CAPTION vs. CTX-FRAGMENT	384	74.7 (768)

Table 3: Win rates of CAPTION over the baseline scene descriptions in Phase 1 (# of responses in parentheses)

Win Rates for Image Generators While we focus primarily on how scene descriptions affect illustration quality, we also considered whether there were quality differences based on which image generator was used. These results are given in §A.1.3.

4.5 Phase 2 Motivation and Design

After verifying that the LLM-generated scene descriptions can contribute substantially to the quality of illustrations, we then wanted to compare the impact of different LLMs as scene captioners. Phase 1 used only CLAUDE-3.5 as the scene captioner. In *Phase 2*, we included other LLMs with storytelling-relevant capabilities (e.g. Tian et al., 2024): GPT-4O[7] (OpenAI et al., 2024) and LLAMA-3.1-405B[8] (Grattafiori et al., 2024), utilizing the same scene captioning prompt (Table A.15).

We expanded the Phase 2 data to include a larger set of fragments compared with those of Phase 1. We randomly sampled 1000 stories from the ROCStories dev set, split them into fragments using the same method from Phase 1 (CLAUDE-3.5 with the Table A.14 prompt), then randomly selected one fragment per story for inclusion in the dataset.

We also considered a larger set of image generators in Phase 2. Based on the state of the Artificial Analysis Leaderboard in November 2024, we selected five image generators. This included MJ-6.1 from Phase 1, as well as FLUX1.1[pro] (referred to here as FLUX-1.1-PRO) (Black Forest Labs, 2024b), Ideogram 2.0 (IDEOGRAM-2.0) (Ideogram, 2024), Recraft V3 (RECRAFT-V3) (Recraft, 2024), and Stable Diffusion 3.5 Large (SD-3.5-LARGE) (Stability AI, 2024).

We applied the scene illustration pipeline to produce illustrations for all 1000 story fragments, varying runs of the pipeline between the three scene captioners and five image generators. Based on these variations, we sampled 1218 pairs comprised of 1582 illustrations. We sampled a roughly equal ratio of pairs where the illustrations varied by scene captioner, image generator, or both scene captioner and image generator. The exact distribution is specified in Table A.9. We repeated the same procedure described in §4.3 to obtain selections from two annotators for the better illustration in each of these pairs. There were 48 (out of 49 total) annotators on Prolific who passed the attention checks, each annotating between 46 and 109 pairs (median=50), resulting in a total of 2436 responses for 1218 pairs.

4.6 Phase 2 Annotation Results

Inter-annotator Agreement As shown in Table A.9, the overall κ_u for all 1218 pairs in Phase 2 was 0.228. This can be classified as fair agreement, which is lower than the agreement observed for Phase 1. As done with Phase 1, §A.1.2 analyzes agreement across different pair subsets.

Win Rates for Scene Captioners Table 4 shows the win rates for each LLM scene captioner against each of the others. In particular, each value is the percentage of responses where the illustration associated with the scene captioner in the row label was selected as better than the illustration associated with the scene captioner in the column label. Thus, higher values indicate more success for the scene captioner in the row against the scene captioner in the column. Statistically significant win rates are denoted with an asterisk. Recall that a response of "I can't decide" indicates a tie, which is why win rates of less than 50% may be statistically significant. These results show that CLAUDE-3.5 yields the highest win rates, followed by GPT-4O, with LLAMA-3.1 having lowest rates. The win rate for CLAUDE-3.5 against LLAMA-3.1 is statistically significant, suggesting that the former generates more descriptive captions compared with the latter.

	CLAUDE-3.5	GPT-4O	LLAMA-3.1
CLAUDE-3.5	-	46.1(532)	49.6* (536)
GPT-4O	41.2 (532)	-	47.8 (552)
LLAMA-3.1	39.7 (536)	42.9(552)	-

Table 4: Win rates (%) by scene captioner for Phase 2 (# of responses in parentheses)

Win Rates for Image Generators While not the focus of our analysis, we observed some significant differences in the win rates of different image generators. These results appear in §A.1.3.

[7] Specifically, gpt-4o-2024-05-13, ran via the OpenAI API

[8] Specifically, llama-3.1-405b-instruct, ran via the Replicate API

Fragment (within story)	**Illustration 1**	**Illustration 2**
Sophie's nana was terminally ill. Sophie visited her in the hospital to say goodbye. **Her nana gave Sophie her prized gold locket. She told Sophie to keep it to remember her by.** Sophie cried.		
Criteria	**Response**	**Response**
1. The image shows two people: an elderly woman (nana) and a younger woman (Sophie)	✓	✗
2. The setting appears to be a hospital room or medical facility	✓	✓
3. The elderly woman is in a hospital bed or medical chair	✓	✗
4. The image shows a gold locket	✓	✓
5. The locket is clearly visible and recognizable as a piece of jewelry	✓	✓
6. The elderly woman is holding or presenting the locket to the younger woman	✗	✓
7. The younger woman's hand is positioned to receive or touch the locket	✓	✗
8. The facial expressions of both women convey emotional significance	✓	✓
9. The elderly woman's expression shows love, tenderness, or sadness	✓	✓
10. The younger woman's expression shows a mix of emotions (sadness, gratitude, love)	✓	✗
11. The body language of both women suggests intimacy and connection	✓	✓
12. The composition focuses on the moment of giving/receiving the locket	✓	✓
13. The lighting adequately illuminates the locket and the faces of both women	✓	✓
14. The locket appears to be in good condition, suggesting its value as a keepsake	✓	✓
15. The elderly woman's appearance suggests illness or frailty	✓	✗
16. The younger woman's appearance and demeanor suggest she is visiting	✓	✗
17. The overall atmosphere of the image conveys a solemn and meaningful moment	✓	✓
18. The spatial relationship between the two women suggests closeness and care	✓	✓
19. Any medical equipment or hospital elements are present but not dominating the scene	✓	✓
20. The perspective allows viewers to see both the locket and the emotional exchange between the women	✓	✓
	Score=19.0	Score=14.0

Table 5: Demonstration of criterial rating approach applied to both illustrations in a given pair. In this particular example, the criteria writer is CLAUDE-3.5, and the rater providing each response is GPT-4O.

5 Predicting Illustration Quality

The dataset presented in §4 provides an opportunity to understand what defines the quality of a scene illustration. To initiate this line of work, we explored a particular approach to modeling annotators' judgments. Our approach leverages the finding from §4 that LLMs can effectively verbalize visual descriptions of scenes based on the story text. We consider whether these descriptions can be used as *criteria* for predicting illustration quality.

5.1 Criteria Generation

For the remaining experiments, we combined the data from Phase 1 and Phase 2, which together included illustrations for 1206 story fragments. For each fragment, we ran the prompt in Table A.18 to produce criteria articulating the expected visual characteristics of the scene illustration. We use the term *criteria writer* to refer to an LLM's role when running this prompt, and we refer to its output as a *criteria set*. An example of a criteria set is included in Table 5. Note that a criteria writer model does not require vision capabilities, since it observes only the story text as input. §A.2.1 discusses some design considerations for generating criteria.

Criteria Writer Details We examined three criteria writers, the same LLMs that operated as scene captioners in §4.5: CLAUDE-3.5, GPT-4O, and LLAMA-3.1. Applying the Table A.18 prompt with temperature=0 to facilitate deterministic output, each criteria writer generated one criteria set per fragment. We post-processed this output to identify each individual criterion according to its expected numerical label in the set. §A.2.2 gives some descriptive analysis of the criteria sets.

5.2 Criteria-based Ratings

After obtaining the criteria sets, we then enlisted visually-enabled models to assess illustrations based on this criteria. In our scheme, when applying a criteria set to score a given illustration, each criterion receives a response indicating whether or not it is satisfied by the image. The overall illustration quality is quantified by the total number of satisfied criteria. Our scoring protocol is as follows: a response conveying that the criterion is satisfied is assigned 1.0 points; a response conveying "maybe" or partial satisfaction is assigned 0.5 points; and a response conveying the criterion is not satisfied is assigned 0.0 points. The total score for an illustration is the sum of these point values.

Criteria Writer	VLM Rater							
	CLAUDE-3.5		GPT-4O		PIXTRAL		*Average*	
	Criterial	Base	Criterial	Base	Criterial	Base	Criterial	Base
CLAUDE-3.5	0.698	0.594	0.688	0.550	0.691	0.562	0.692	0.569
GPT-4O	0.684	0.588	0.668	0.566	0.687	0.560	0.680	0.571
LLAMA-3.1	0.662	0.584	0.657	0.575	0.662	0.556	0.660	0.571
Average	0.681	0.589	0.671	0.563	0.680	0.559	0.677	0.570

Table 6: Accuracy of criterial and baseline (Base) raters grouped by criteria writer and VLM

We implemented this by prompting a visually-enabled LLM (i.e. VLM) to assign responses to each criterion for a given illustration. We use the term *criterial rater* to refer to a VLM's role when running this prompt, which appears in Table A.19. As shown, the rater observes an illustration and the criteria set for the corresponding fragment. The rater is asked to respond to each criterion (where a response of '✓' means the criterion is satisfied, '✗' means not satisfied, and '?' means "maybe"). As post-processing, we parsed these response tokens and mapped them to the point values defined above to obtain the illustration score. Table 5 exemplifies this approach applied to both illustrations in a pair.

Rater Details For raters, we utilized three VLMs that have obtained notable performance on visual understanding benchmarks: CLAUDE-3.5, GPT-4O, and PIXTRAL[9] (Mistral AI, 2024). Each rater ran the prompt in Table A.19 with temperature=0. All images were resized to a height of 240 pixels with proportional width. We briefly assessed the correctness of raters' responses, which appears in §A.2.3.

Comparative Baseline To determine the impact of criteria in assessing quality, we designed a comparable rating approach that scores illustrations on the same scale as the criterial rater but without observing the criteria itself. We use the term *baseline rater* to refer to a VLM's application of the prompt for this approach, which is shown in Table A.20. The prompt presents the fragment and illustration, and instructs the VLM to assign a rating in half-point increments between 0 and a maximum that is dynamically set to the length of the given criteria set. For each criteria writer, we compare the result obtained by a particular criterial rater to the analogous result obtained by the baseline rater.

5.3 Selection Performance Results

We filtered pairs to only include those where there was a consensus between both annotators for which illustration was better. This yielded 1501 pairs comprised of 1987 illustrations. We then applied all criterial and baseline raters to score the illustrations. For a given pair, a rater's selection was the image it assigned a higher score. We measured each rater's performance in terms of proportion of pairs where the rater's selection matched the human selection. We refer to this metric as *accuracy*.

Table 6 shows the accuracy for all raters on these pairs, with the respective averages for each criteria writer and rater. For reference, always selecting the second illustration in each pair yields 50% accuracy. We observe that the criterial raters all considerably outperform the baseline raters (an average accuracy of ≈68% vs. 57%). Criteria from different writers yields comparable results, with CLAUDE-3.5 averaging the highest accuracy across raters (≈69%). The raters obtain similar accuracies when applied to the same criteria. Overall this outcome suggests that criteria are an effective strategy for modeling illustration quality, which in turn provides further evidence of LLMs' capacity to verbalize visual characteristics of narrative scenes. This still leaves room for accuracy improvements, motivating future exploration of this dataset for understanding what makes a compelling scene illustration.

6 Conclusion and Future Work

This paper details a pipeline for generating illustrations of narrative scenes, which we apply to produce a quality-annotated dataset of illustrations for a popular story corpus. We identify that LLMs can facilitate this task by distilling scene descriptions from story text. We show that this capacity to verbalize implicit scene knowledge is useful for modeling illustration quality.

Our long-term objective is to generate multi-scene illustration sequences depicting an entire story. This poses key research challenges, such as ensuring visual consistency between story elements (e.g. Liu et al., 2025) as well as progressive story development across images (e.g. Maharana et al., 2022). Our future work will expand our current illustration pipeline to address these challenges.

[9] Specifically, pixtral-large-2411, ran via the MistralAI API

Limitations

We consider the following limitations:

Proprietary Models Our scene illustration pipeline has a plug-and-play design, enabling any LLM to be used for fragmentation and scene captioning and any text-to-image model to be used for image generation. However, most of the models we assessed in this paper are proprietary (i.e. closed-weight), with exception to LLAMA-3.1 and SD-3.5-LARGE. While the gap between closed and open-weight models is narrowing (Cottier et al., 2024), currently most models with capabilities relevant to the illustration task are closed-weight. This poses a general disadvantage in accessibility and reproducibility, which applies likewise to this work.

Prompt Design Currently there is no tractable way to ensure that a particular prompt is optimal for the task it is intended to perform. Prompt optimization is fundamentally a process of iterative trial-and-error, even when automation is used to increase the number of trials. For our experiments, we primarily employed a principled approach to writing prompts, which involved adhering to general guidance on effective prompt design such as explaining instructions clearly and including representative exemplars (e.g. DAIR.AI, 2025). We iterated on this design according to qualitative subjective assessment of model outputs for inputs not included in our scene illustration dataset (i.e. "vibe-based" prompt engineering), rather than employing a quantitative optimization approach (e.g. Khattab et al., 2024) based on targets in a designated development set. There are tradeoffs to this technique: while it avoids overfitting to our presented dataset, it leaves open the possibility of further prompt optimization, which could yield a different view of model behavior compared with our observations.

Story Corpus The story corpus we use, ROCStories, is popular in NLP research for some of the same reasons discussed in §4: the constrained language and structure of the text makes the narrative elements more accessible to computational modeling techniques. The stories were authored specifically for the benefit of this research. However, this corpus is distinct from "naturally" authored stories whose complexity is what makes them compelling to readers. We have not yet fully assessed whether our scene illustration pipeline generalizes to more complex narratives.

Ethical Considerations

Generative AI models, and in particular text-to-image models, pose various ethical risks (Bird et al., 2023). In this work, we were primarily concerned with the risk of exposing Prolific annotators to harmful content. We attempted to mitigate this risk by manually reviewing stories sampled for inclusion in our dataset. We flagged stories that we anticipated could yield objectionable illustrations, and re-sampled a different story to replace each of these. Ultimately, this re-sampling was triggered for 10 stories. Of course, this procedure did not eliminate the risk, so we also utilized the content warning feature on the Prolific platform, which indicated to potential annotators that the task could expose them to offensive and/or biased content.

References

Jie An, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Lijuan Wang, and Jiebo Luo. 2024. Openleaf: A novel benchmark for open-domain interleaved image-text generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 11137–11145, New York, NY, USA. Association for Computing Machinery.

Charlotte Bird, Eddie Ungless, and Atoosa Kasirzadeh. 2023. Typology of risks of generative text-to-image models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 396–410, New York, NY, USA. Association for Computing Machinery.

Black Forest Labs. 2024a. Announcing black forest labs: Flux.1 model family.

Black Forest Labs. 2024b. Announcing flux1.1 [pro] and the bfl api.

Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. 2023. Elo uncovered: Robustness and best practices in language model evaluation. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 339–352, Singapore. Association for Computational Linguistics.

Stephen Brade, Bryan Wang, Mauricio Sousa, Sageev Oore, and Tovi Grossman. 2023. Promptify: Text-to-image generation through interactive prompt exploration with large language models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

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

Jonathan Cook, Tim Rocktäschel, Jakob Foerster, Dennis Aumiller, and Alex Wang. 2024. Ticking all the boxes: Generated checklists improve llm evaluation and generation. *Preprint*, arXiv:2410.03608.

Ben Cottier, Josh You, Natalia Martemianova, and David Owen. 2024. How far behind are open models?

DAIR.AI. 2025. General tips for designing prompts.

Edirlei Soares de Lima, Marco A. Casanova, and Antonio L. Furtado. 2024. Imagining from images with an ai storytelling tool. *Preprint*, arXiv:2408.11517.

Shachar Don-Yehiya, Leshem Choshen, and Omri Abend. 2023. Human learning by model feedback: The dynamics of iterative prompting with midjourney. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4146–4161, Singapore. Association for Computational Linguistics.

Yingchaojie Feng, Xingbo Wang, Kam Kwai Wong, Sijia Wang, Yuhong Lu, Minfeng Zhu, Baicheng Wang, and Wei Chen. 2024. PromptMagician: Interactive Prompt Engineering for Text-to-Image Creation . *IEEE Transactions on Visualization & Computer Graphics*, 30(01):295–305.

Zhangyin Feng, Yuchen Ren, Xinmiao Yu, Xiaocheng Feng, Duyu Tang, Shuming Shi, and Bing Qin. 2023. Improved visual story generation with adaptive context modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4939–4955, Toronto, Canada. Association for Computational Linguistics.

Yuan Gong, Youxin Pang, Xiaodong Cun, Menghan Xia, Yingqing He, Haoxin Chen, Longyue Wang, Yong Zhang, Xintao Wang, Ying Shan, and Yujiu Yang. 2023. Interactive story visualization with multiple characters. In *SIGGRAPH Asia 2023 Conference Papers*, SA '23, New York, NY, USA. Association for Computing Machinery.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Brett A. Halperin and Stephanie M. Lukin. 2023. Envisioning narrative intelligence: A creative visual storytelling anthology. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2023. Optimizing prompts for text-to-image generation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Huiguo He, Huan Yang, Zixi Tuo, Yuan Zhou, Qiuyue Wang, Yuhang Zhang, Zeyu Liu, Wenhao Huang, Hongyang Chao, and Jian Yin. 2024. Dreamstory: Open-domain story visualization by llm-guided multi-subject consistent diffusion. *Preprint*, arXiv:2407.12899.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xudong Hong, Asad Sayeed, Khushboo Mehra, Vera Demberg, and Bernt Schiele. 2023. Visual writing prompts: Character-grounded story generation with curated image sequences. *Transactions of the Association for Computational Linguistics*, 11:565–581.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. 2023. TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering . In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20349–20360, Los Alamitos, CA, USA. IEEE Computer Society.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, San Diego, California. Association for Computational Linguistics.

Ideogram. 2024. Ideogram 2.0.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan A, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. DSPy: Compiling declarative language model calls into state-of-the-art pipelines. In *The Twelfth International Conference on Learning Representations*.

Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. 2023. Generating images with multimodal language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

JR Landis and GG Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Lin Li, Jun Xiao, Guikun Chen, Jian Shao, Yueting Zhuang, and Long Chen. 2023. Zero-shot visual relation detection via composite visual cues from large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. 2019. StoryGAN: A Sequential Conditional GAN for Story Visualization . In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6322–6331, Los Alamitos, CA, USA. IEEE Computer Society.

Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. 2024. LLM-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *Transactions on Machine Learning Research*.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2025. Evaluating text-to-visual generation with image-to-text generation. In *Computer Vision – ECCV 2024*, pages 366–384, Cham. Springer Nature Switzerland.

Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. 2024. Intelligent Grimm - Open-ended Visual Storytelling via Latent Diffusion Models . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6190–6200, Los Alamitos, CA, USA. IEEE Computer Society.

Tao Liu, Kai Wang, Senmao Li, Joost van de Weijer, Fahad Shahbaz Khan, Shiqi Yang, Yaxing Wang, Jian Yang, and Ming-Ming Cheng. 2025. One-prompt-one-story: Free-lunch consistent text-to-image generation using a single prompt. In *The Thirteenth International Conference on Learning Representations*.

Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. 2023. LLMScore: Unveiling the power of large language models in text-to-image synthesis evaluation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Stephanie Lukin, Reginald Hobbs, and Clare Voss. 2018. A pipeline for creative visual storytelling. In *Proceedings of the First Workshop on Storytelling*, pages 20–32, New Orleans, Louisiana. Association for Computational Linguistics.

Adyasha Maharana and Mohit Bansal. 2021. Integrating visuospatial, linguistic, and commonsense structure into story visualization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6772–6786, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Adyasha Maharana, Darryl Hannan, and Mohit Bansal. 2022. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In *Computer Vision – ECCV 2022*, pages 70–87, Cham. Springer Nature Switzerland.

Mayug Maniparambil, Chris Vorster, Derek Molloy, Noel Murphy, Kevin McGuinness, and Noel E. O'Connor. 2023. Enhancing CLIP with GPT-4: Harnessing Visual Descriptions as Prompts . In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 262–271, Los Alamitos, CA, USA. IEEE Computer Society.

Lech Mazur. 2025. Llm creative story-writing benchmark.

Sachit Menon and Carl Vondrick. 2023. Visual classification via description from large language models. In *The Eleventh International Conference on Learning Representations*.

Sebastian Michelmann, Manoj Kumar, Kenneth A Norman, and Mariya Toneva. 2025. Large language models can segment narrative events similarly to humans. *Behavior Research Methods*, 57(1):1–13.

Midjourney. 2024. Version 6.1.

Mistral AI. 2024. Pixtral large.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, et al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337, Abu Dhabi, UAE. Association for Computational Linguistics.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. 2023. What does a platypus look like? generating customized prompts for zero-shot image classification. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15645–15655.

Recraft. 2024. Recraft introduces a revolutionary ai model that thinks in design language.

Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L. Glassman. 2023. Where to hide a stolen elephant: Leaps in creative writing with multimodal machine intelligence. *ACM Trans. Comput.-Hum. Interact.*, 30(5).

Soumik Rakshit. 2024. Building a genai-assisted automatic story illustrator.

Stability AI. 2024. Introducing stable diffusion 3.5.

Ming Tao, Bing-Kun Bao, Hao Tang, Yaowei Wang, and Changsheng Xu. 2024. Coin: A lightweight and effective framework for story visualization and continuation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 10659–10668, New York, NY, USA. Association for Computing Machinery.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17659–17681, Miami, Florida, USA. Association for Computational Linguistics.

Qian Wan, Xin Feng, Yining Bei, Zhiqi Gao, and Zhicong Lu. 2024. Metamorpheus: Interactive, affective, and creative dream narration through metaphorical visual storytelling. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Zhijie Wang, Yuheng Huang, Da Song, Lei Ma, and Tianyi Zhang. 2024. Promptcharm: Text-to-image generation through multi-modal prompting and refinement. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Yingcong Chen. 2024. Seedstory: Multimodal long story generation with large language model. *Preprint*, arXiv:2407.08683.

Weizhe Yuan, Pengfei Liu, and Matthias Gallé. 2024. LLMCrit: Teaching large language models to use criteria. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7929–7960, Bangkok, Thailand. Association for Computational Linguistics.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.

A Appendix

A.1 Additional Dataset Statistics

A.1.1 Analysis of Phase 1 Fragments

As mentioned in §4.2, there were 206 total fragments derived from the 50 stories in Phase 1, based on applying CLAUDE-3.5 to the prompt in Table 14. As shown in Table 7, the majority consist of a single sentence, with some consisting of 2 sentences and a few having 3 sentences. An internal annotator assessed each fragment to determine if it was the correctly-sized unit for a scene illustration. A fragment was considered incorrectly-sized if it either did not include all the text in the story relevant to a single scene (i.e. the fragment was too short) or if it included text pertaining to more than one scene (i.e. the fragment was too long). The annotator considered the vast majority of fragments to be correctly-sized (≈96%).

# Total Fragments	206
# 1-Sentence Fragments	164
# 2-Sentence Fragments	40
# 3-Sentence Fragments	2
Mean # Sentences Per Fragment	1.21
Mean # Fragments Per Story	4.12
% of Correctly-Sized Fragments	96.1%

Table 7: Fragmentation statistics for stories in Phase 1

	# of Unique Illustrations
All	1183
By Scene Description	
FRAGMENT	395
CTX-FRAGMENT	384
CAPTION	404
By Image Generator	
FLUX-1-PRO	595
MJ-6.1	588

Illustration Pair Type	# Pairs	κ_u
All	1384	0.447
Different Scene Descriptions	1064	0.516
FRAGMENT vs. CAPTION	680	0.520
CTX-FRAGMENT vs. CAPTION	384	0.504
Different Image Generators		
FLUX-1-PRO vs. MJ-6.1	661	0.364

Table 8: Descriptive statistics for Phase 1 data, including inter-annotator agreement (κ_u) for different pair types

A.1.2 Inter-annotator Agreement

As presented in §4.4, the uncertainty-weighted kappa (κ_u) for all Phase 1 pairs was 0.447. We also considered whether κ_u differed based on the variable components of the illustrations in each pair. The bottom section of Table 8 shows that agreement was higher among the 1064 pairs where the illustrations used different scene descriptions (κ_u=0.516), while agreement was lower among the 661 pairs where the illustrations used different image generators (κ_u=0.364). This indicates that the scene description type was particularly influential to annotators' judgments of which illustration was better. Considered along with Table 3, we can specifically conclude that ablating the scene captioner (i.e. using the baseline FRAGMENT/CTX-FRAGMENT scene descriptions) yielded illustrations that annotators consistently judged as lower quality relative to those that used the scene captioner.

	# of Unique Illustrations
All	1582
By Scene Captioner	
CLAUDE-3.5	496
GPT-4O	532
LLAMA-3.1	554
By Image Generator	
FLUX-1.1-PRO	307
IDEOGRAM-2.0	300
MJ-6.1	321
RECRAFT-V3	323
SD-3.5-LARGE	331

Illustration Pair Type	# Pairs	κ_u
All	1218	0.228
Different Scene Captioners	810	0.236
CLAUDE-3.5 vs. GPT-4O	266	0.198
CLAUDE-3.5 vs. LLAMA-3.1	268	0.234
GPT-4O vs. LLAMA-3.1	276	0.273
Different Image Generators	813	0.228
FLUX-1.1-PRO vs. IDEOGRAM-2.0	72	0.079
FLUX-1.1-PRO vs. MJ-6.1	74	0.183
FLUX-1.1-PRO vs. RECRAFT-V3	75	0.089
FLUX-1.1-PRO vs. SD-3.5-LARGE	71	0.164
IDEOGRAM-2.0 vs. MJ-6.1	99	0.312
IDEOGRAM-2.0 vs. RECRAFT-V3	81	0.159
IDEOGRAM-2.0 vs. SD-3.5-LARGE	94	0.339
MJ-6.1 vs. RECRAFT-V3	73	0.419
MJ-6.1 vs. SD-3.5-LARGE	88	0.184
RECRAFT-V3 vs. SD-3.5-LARGE	86	0.271

Table 9: Descriptive statistics for Phase 2 data, including inter-annotator agreement (κ_u) for different pair types

For Phase 2, as reported in §4.6, the overall κ_u was 0.228 among all 1218 pairs. The bottom section of Table 9 also shows that the agreement level was similar between the 810 pairs where illustrations involved different scene captioners (κ_u=0.236) and the 813 pairs that involved different image generators (κ_u=0.228). Agreement varied especially widely based on which particular image generators were paired together (ranging from 0.079 for FLUX-1.1-PRO vs. IDEOGRAM-2.0, up to 0.419 for MJ-6.1 vs. RECRAFT-V3). This indicates that in contrast to Phase 1 where there was a

significant variable (the ablation of the scene captioner) that made the relative quality of illustrations more consistently distinguishable to annotators, the Phase 2 pairs were less reliably distinct.

A.1.3 Win Rates for Image Generators

To determine whether the choice of image generator influenced illustration quality in both Phase 1 and Phase 2, we computed the win rates for each image generator against each other among the pairs that used different image generators.

For Phase 1, there were only two image generators used to produce illustrations, FLUX-1-PRO vs. MJ-6.1. We did not find any significant difference in the win rates of these image generators. Table 10 shows these results.

FLUX-1-PRO	MJ-6.1
42.6%	41.0%

Table 10: Win rates (percentages) of FLUX-1-PRO vs MJ-6.1, for 1322 responses pertaining to Phase 1 pairs where the illustrations used different image generators.
.

The Phase 2 data utilized a larger set of image generators. Table 11 shows the win rates of these image generators, presented comparably to Table 4 where each value is the percentage of selections for the image generator in the row against the image generator in the column. According to these results, IDEOGRAM-2.0 obtains the highest win rates against the other image generators, with significant success against FLUX-1.1-PRO, MJ-6.1, and SD-3.5-LARGE. Additionally, RECRAFT-V3 is significantly favored over MJ-6.1. Further analysis of these model differences for this task is an opportunity for future work.

A.2 Criteria-based Evaluation Details

A.2.1 Criteria Design Considerations

As referenced in §5.1, two design considerations for the criteria generation prompt (Table 18) were *flexibility* and *atomicity*. Flexibility emphasizes that a scene characteristic referenced by a criterion may be depicted with multiple alternative visual details that all align equally with the story text. For example, if a criterion conveys that the scene should take place at a particular location, it should be flexible about how the location is portrayed. Regarding atomicity, we aimed for each criterion to be as atomic as possible, meaning that it should refer to only a single characteristic of the scene. This promotes concise and easy-to-parse responses when judging whether the criterion is satisfied by an image, as opposed to a criterion that conflates multiple characteristics, some of which are satisfied and others that are not. Concerning the length of the generated criteria, our prompt did not specify a particular number of criteria to return, but the exemplar and instructions indicated that the criteria should comprehensively refer to as many scene characteristics as possible without redundancy.

A.2.2 Descriptive Analysis of Criteria Sets

Regarding the generated criteria sets (§5.1), Table 12 compares the average number of criteria in the sets generated by each criteria writer, revealing that CLAUDE-3.5 generated the longest criteria sets, followed by GPT-4O, and LLAMA-3.1.

CLAUDE-3.5	GPT-4O	LLAMA-3.1
19.3	17.3	15.8

Table 12: Mean number of criteria per set for each writer

Additionally, Figure 2 visualizes all criteria, based on encoding each criterion with the ModernBERT embedding model (Warner et al., 2024), then running PCA + t-SNE to yield a 2D embedding. While there are no distinct clusters associated with each criteria writer, some separation can be observed between the criteria generated by CLAUDE-3.5 and GPT-4O, while those generated by LLAMA-3.1 are more distributed alongside both other writers.

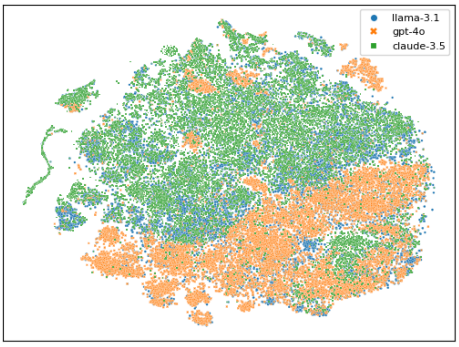

Figure 2: Visualization of criteria generated by each writer. Each point is a single criterion represented by its ModernBERT embedding. We applied PCA followed by t-SNE to plot the embedding in 2D space.

A.2.3 Criterial Rater Assessment

As referenced in §5.2, we conducted a small assessment of the correctness of the VLM raters' re-

	FLUX-1.1-PRO	IDEOGRAM-2.0	MJ-6.1	RECRAFT-V3	SD-3.5-LARGE
FLUX-1.1-PRO	-	35.4 (144)	43.2 (148)	45.3 (150)	48.6 (142)
IDEOGRAM-2.0	53.5* (144)	-	61.6* (198)	46.9 (162)	58.5* (188)
MJ-6.1	39.9 (148)	31.8 (198)	-	28.8 (146)	44.9 (176)
RECRAFT-V3	44.0 (150)	40.1 (162)	61.6* (146)	-	50.6 (172)
SD-3.5-LARGE	43.7 (142)	30.3 (188)	43.8 (176)	37.8 (172)	-

Table 11: Win rates (percentages) by image generator for Phase 2, with the number of responses in parentheses. Statistically significant win rates are denoted with an asterisk.

sponses to criteria. To do this, we randomly sampled 100 items, each with a unique image and criteria set. We then enlisted an expert human annotator to assign a response to each criterion, which we treated as the gold standard criterion response for the sampled image. We measured rater correctness in terms of linear-weighted κ agreement with the gold standard, where responses of '✗' were mapped to -1, '?' to 0, and '✓' to 1; this results in less weight assigned to disagreements involving '?' ("maybe") responses. Table 13 shows the κ on these 1699 criterion responses. It indicates that raters are all substantially aligned with the human annotator, though GPT-4O appears to have the highest human agreement, followed by CLAUDE-3.5, and then PIXTRAL.

Rater	κ
CLAUDE-3.5	0.676
GPT-4O	0.710
PIXTRAL	0.622

Table 13: Correctness of criterial rater responses (κ)

You are performing the task of story fragmentation. The task is to split a story into fragments where each fragment consists of a distinct part of the story. A fragment contains enough information to yield a visualization that is unique to that part of the story. In this version of the task, you will insert brackets (i.e. [and]) into the given story text to annotate the beginning and end of each fragment. Write the fragments without preamble. Here are some examples:

Story: Mia sat at home in her living room watching sports. Her favorite soccer team was playing their rival. To encourage her team, she began chanting positive phrases. During her chant, her favorite team scored a goal. Mia cheered loudly and thought that she helped score that goal.
Fragmented Story: [Mia sat at home in her living room watching sports. Her favorite soccer team was playing their rival.] [To encourage her team, she began chanting positive phrases.] [During her chant, her favorite team scored a goal.] [Mia cheered loudly and thought that she helped score that goal.]

[...2 more exemplars...]

Story: `{{story}}`
Fragmented Story:

Table 14: **Fragmentation prompt**. LLM prompt for annotating fragment boundaries in a story, which consists of a task instruction and exemplars demonstrating the task. The stories in the exemplars are from various corpora (ROCStories, TinyStories, and the ARL Creative Visual Storytelling Anthology).

Imagine an AI system will be used to generate illustrations for story fragments. This AI illustrator generates a single image given a caption describing what is contained in the image. Your task is to read a story fragment along with its story context and write a caption that describes how to illustrate the fragment. The caption should elaborately describe the most salient way to visualize the fragment. It should completely specify all the information the illustrator needs to generate the image. Write the caption without preamble. Here are some examples:

Story Context: Carrie had just learned how to ride a bike. She didn't have a bike of her own. Carrie would sneak rides on her sister's bike. She got nervous on a hill and crashed into a wall. The bike frame bent and Carrie got a deep gash on her leg.
Story Fragment: Carrie would sneak rides on her sister's bike.
Caption for Story Fragment: A young girl with a mischievous expression carefully wheels a bicycle that's slightly too big for her out of a garage, glancing over her shoulder as if making sure no one sees her.

[...2 more exemplars...]

Story Context: `{{story}}`
Story Fragment: `{{fragment}}`
Caption for Story Fragment:

Table 15: **Scene captioning prompt**. LLM prompt for generating scene descriptions. The prompt consists of a task instruction and exemplars demonstrating the task. The stories in the exemplars are from various corpora (ROCStories, TinyStories, and the ARL Creative Visual Storytelling Anthology).

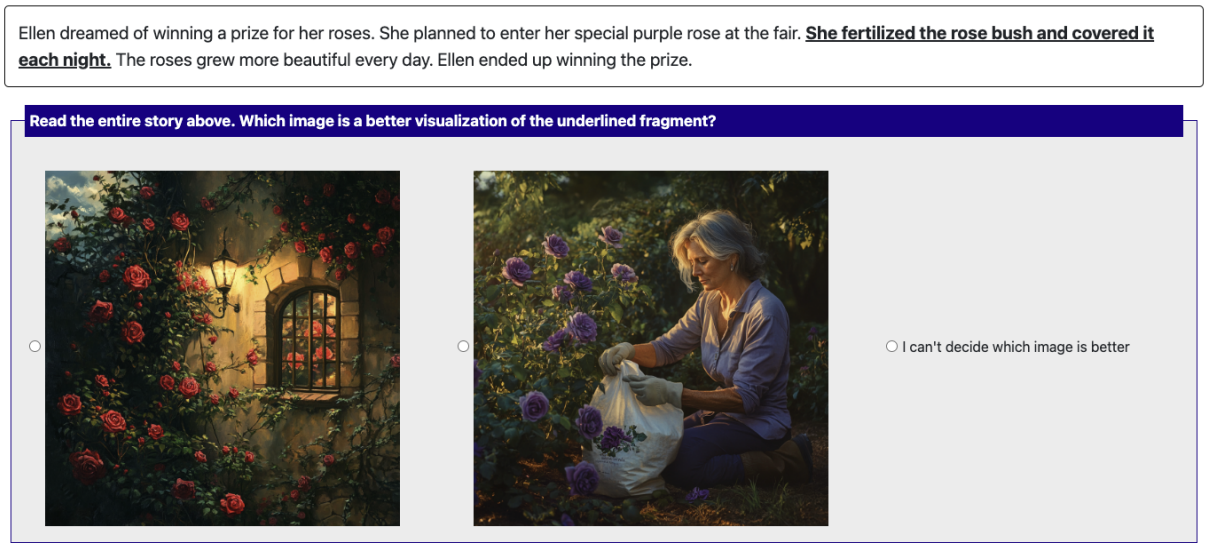

Figure 3: Example of a item shown to participants in the annotation task described in §4.3

Fragment (within story)	**CAPTION**	**CAPTION Illustration**	**CTX-FRAGMENT Illustration**
Joe went to the junkyard to find treasures. He dug through a stack of old clothes. At the bottom was a vintage shirt with a rock band on it. Joe put the shirt on to try it out. Joe was happy the shirt fit.	A young man with an excited expression stands at the entrance of a sprawling junkyard. He's wearing casual clothes and holding a small backpack. The junkyard is filled with piles of discarded items, including old appliances, furniture, and scrap metal. The sky is bright and sunny, creating an adventurous atmosphere. In the foreground, a rusty sign reads "Junkyard" with operating hours listed below.		
Trish hated the outdoors. Her friends convinced her to go camping. **She wasn't having a good time.** They showed her how to fish and showed her the stars. Eventually Trish grudgingly came to accept nature.	A young woman with a scowling expression sits on a log by a campfire, arms crossed and looking miserable. She's surrounded by cheerful friends setting up tents and unpacking camping gear in a forest clearing. Her clean, urban clothing contrasts with the rugged outdoor setting, emphasizing her discomfort with nature.		
Sammy's coffee grinder was broken. He needed something to crush up his coffee beans. He put his coffee beans in a plastic bag. **He tried crushing them with a hammer.** It worked for Sammy.	A man in casual clothing stands at a kitchen counter, holding a hammer above a clear plastic bag filled with whole coffee beans. The hammer is poised mid-swing, about to strike the bag. The man's face shows a mix of determination and uncertainty. Scattered around the counter are a few stray coffee beans and an unplugged, visibly broken coffee grinder.		
I decided to go on a bike ride with my brother. We both headed out in the morning. We were having a lot of fun. Suddenly, he hit a rock and broke his wheel! **I felt very badly for my brother**.	A concerned young person stands next to their brother, who sits dejectedly on the ground next to a fallen bicycle with a visibly bent front wheel. The scene takes place on a sunny morning on a bike path, with trees and nature in the background. The standing sibling has a sympathetic expression, reaching out to comfort their brother, who looks disappointed and upset about the broken bike.		

Table 16: Scene illustrations in Phase 1. For each story fragment, we show an illustration resulting from the LLM-generated CAPTION scene description and one resulting from the baseline CTX-FRAGMENT scene description. The image generator for all illustrations is FLUX-1-PRO.

	CAPTION Win %		
Scene Description Pair	MJ-6.1 & FLUX-1-PRO	MJ-6.1 Only	FLUX-1-PRO Only
CAPTION vs. FRAGMENT	78.1 (1360)	79.2 (342)	77.7 (336)
CAPTION vs. CTX-FRAGMENT	74.7 (768)	74.5 (372)	75.0 (396)

Table 17: Extended view of Table 3. Here, the win rates for CAPTION vs. baseline scene descriptions in Phase 1 are split out by pairs where both illustrations used the same image generator (the MJ-6.1 Only and FLUX-1-PRO Only columns). This shows that the CAPTION win rate is similar regardless of which image generator is used. The number of responses is in parentheses.

Imagine an AI system will be used to judge the quality of images intended to illustrate story fragments. This AI judge scores the images given some criteria about what should be depicted in the images. Your task involves writing the criteria for this AI judge. In particular, you will read a story and focus on a fragment at a specific position in the story. You will write the criteria defining the characteristics the image for that fragment should satisfy in order to be considered a good illustration of the fragment. There are a few things to consider when writing the criteria. First, while the criteria should define the fundamental characteristics depicted in the image, the visual details of these characteristics may vary across images, and alternative details may be similarly effective in illustrating the fragment. Each criterion should be written in a way that accommodates these potential variations in detail, without assuming specific information that is not defined explicitly in the story. Additionally, each criterion should refer to only a single atomic characteristic of the image. If a criterion references multiple characteristics such that an image might satisfy some but not others, it should be further split into multiple separate criteria. For example, instead of writing "the image shows a sapphire ring on the bathroom floor" as one criterion, you should write "the image shows a ring", "the ring contains a sapphire", and "the ring is on the bathroom floor" as separate criteria. The criteria should not only consider the presence of certain elements in the image, but also the visual quality of their depiction. Write the criteria without preamble, with a number header (e.g. '1.') for each criterion. Try to write as many criteria as possible, but avoid specifying extraneous or redundant criteria. Here is an example:

Story Context: Lisa has a beautiful sapphire ring. She always takes it off to wash her hands. One afternoon, she noticed it was missing from her finger! Lisa searched everywhere she had been that day. She was elated when she found it on the bathroom floor!
Story Fragment: She was elated when she found it on the bathroom floor!
Image Criteria for Story Fragment:
1. The image shows a clearly visible ring
2. The image portrays a bathroom setting recognizable through typical bathroom elements (tiles, fixtures, etc.)
3. The ring contains a blue gemstone recognizable as a sapphire
4. The ring is on the bathroom floor
5. The ring appears to be positioned naturally as if it had fallen or been dropped
6. A female figure (Lisa) is present in the image
7. The woman's facial expression clearly conveys joy or elation
8. The woman's body language demonstrates excitement or relief
9. The woman's positioning suggests she has just discovered or is reaching for the ring
10. The lighting adequately illuminates the ring to make it visible as the focal point
11. The perspective of the image allows viewers to see both the ring and the woman's emotional reaction
12. The composition draws attention to the moment of discovery
13. The spatial relationship between the woman and ring suggests imminent retrieval
14. The overall scene composition captures the spontaneous nature of the discovery
15. The woman's appearance suggests this is taking place during daytime/afternoon
16. The ring appears intact and undamaged, justifying the woman's relief
17. The bathroom setting appears residential rather than public

Story Context: `{{story}}`
Story Fragment: `{{fragment}}`
Image Criteria for Story Fragment:

Table 18: **Criteria generation prompt**. LLM prompt used to generate evaluation criteria for assessing the quality of scene illustrations.

You will observe an image along with a list of criteria, where each criterion describes a characteristic or quality that may or may not be depicted in the image. Your task is to determine whether or not each criterion is satisfied by the image. For each criterion, if the image fully satisfies that criterion, write a checkmark ('✓') after it. If the image only partially satisfies the criterion but not completely, write a question mark ('?') after it. Otherwise, if the image does not satisfy that criterion, write an X mark ('✗') after it. Reiterate each criterion before giving your assessment for it, but do not provide additional preamble in your response. Here is an example:

Criteria:
1. The image shows a young woman (Laura) in an apartment setting
2. The woman's facial expression conveys happiness or contentment
3. The apartment appears to be newly moved into, with some visible unpacked items
4. There are visible windows in the apartment
5. The view through the windows shows recognizable California scenery (palm trees, ocean, mountains, or urban landscape)
6. The lighting suggests natural daylight entering the apartment
7. The apartment appears residential and suitable for a recent college graduate
Image: <IMAGE WILL APPEAR HERE>
Criteria Responses:
1. The image shows a young woman (Laura) in an apartment setting ✓
2. The woman's facial expression conveys happiness or contentment ✗
3. The apartment appears to be newly moved into, with some visible unpacked items ?
4. There are visible windows in the apartment ✓
5. The view through the windows shows recognizable California scenery (palm trees, ocean, mountains, or urban landscape) ✗
6. The lighting suggests natural daylight entering the apartment ✓
7. The apartment appears residential and suitable for a recent college graduate ✓

Criteria:
`{{criteria}}`
Image: `{{image}}`
Criteria Responses:

Table 19: **Criteria-based rating prompt**. VLM prompt used to score the quality of a scene illustration by assigning responses to each criterion in a provided criteria set

Your task is to rate how well a particular image illustrates a fragment of a story. You will observe the fragment with its story context, alongside the image depicting the fragment. Provide a rating on a scale ranging from 0.0 to `{{len(criteria)}}` in half-point increments, where 0.0 indicates the image is unrelated to the fragment and `{{len(criteria)}}` indicates the image is a perfect illustration of the fragment. Do not provide additional preamble before the rating. Here is an example:

Story: Laura had just graduated college. She was planning on moving on California. She packed all her belongings in her car and drove 18 hours. When she arrived at her new apartment she unpacked all her things. Laura loved the new change of scenery at her new place.
Fragment: Laura loved the new change of scenery at her new place.
Image: <IMAGE WILL APPEAR HERE>
Rating: 4.5

Story: `{{story}}`
Fragment: `{{fragment}}`
Image: `{{image}}`
Rating:

Table 20: **Baseline rating prompt**. VLM prompt used to score the quality of a scene illustration by directly assigning a rating between 0 and a maximum. This maximum is dynamically set to the total number of criteria in a provided criteria set (`{{len(criteria)}}`), even though the criteria themselves are not referenced in the prompt.