LLMS BEHIND THE SCENES: ENABLING NARRATIVE SCENE ILLUSTRATION

Melissa Roemmele[1], John Joon Young Chung[1], Taewook Kim[2], Yuqian Sun[1], Alex Calderwood[3], Max Kreminski[1]

[1]Midjourney [2]Northwestern University [3]University of California, Santa Cruz

OVERVIEW

- Transforming a story from one modality to another can enhance its appeal
- New text-to-image models suggest the potential to automatically generate visual illustrations for text-based stories
- But what is the best approach for doing this? The story text itself might not be an optimal prompt for a text-to-image model

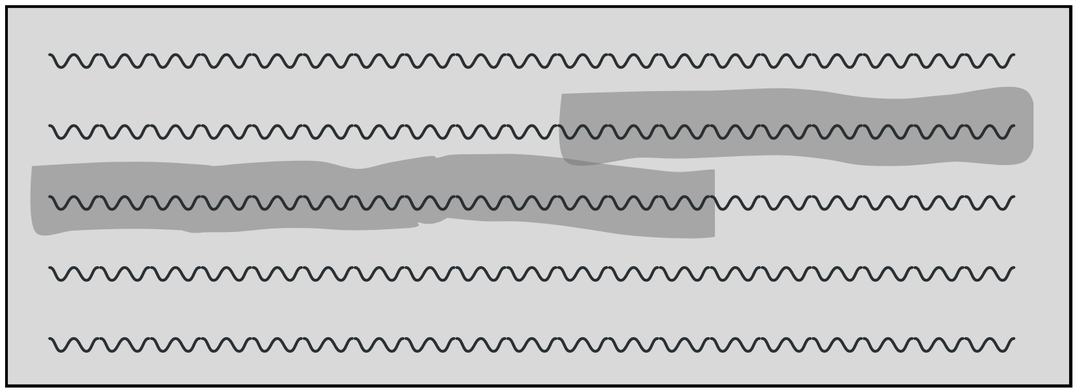
NARRATIVE SCENE ILLUSTRATION

SCENE: an abstract unit of a story that can be illustrated by a single image

NARRATIVE SCENE ILLUSTRATION

SCENE: an abstract unit of a story that can be illustrated by a single image

FRAGMENT: a unit of story text that aligns to a scene

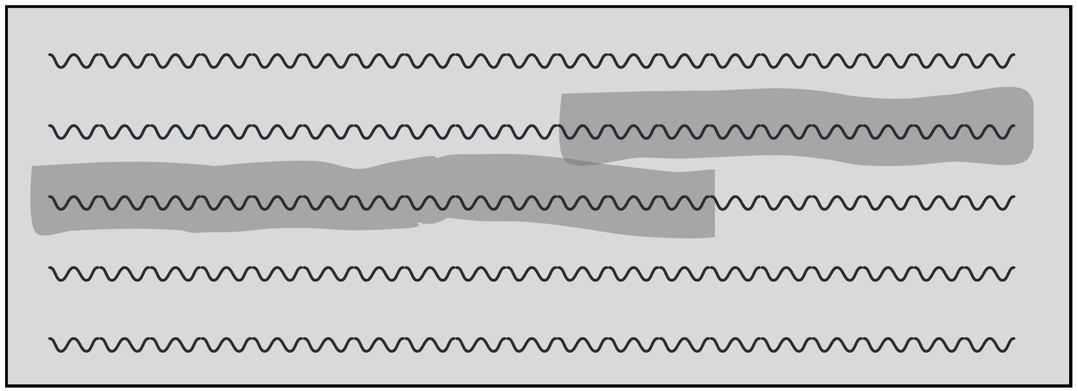


NARRATIVE SCENE ILLUSTRATION

SCENE: an abstract unit of a story that can be illustrated by a single image

SCENE ILLUSTRATION: an image depicting a fragment

FRAGMENT: a unit of story text that aligns to a scene



NARRATIVE SCENE ILLUSTRATION

FRAGMENT

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.

SCENE ILLUSTRATION

NARRATIVE SCENE ILLUSTRATION

FRAGMENT

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.

IMAGE GENERATOR:
a text-to-image model

SCENE ILLUSTRATION

NARRATIVE SCENE ILLUSTRATION

FRAGMENT

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.

SCENE DESCRIPTION: a verbalization of the visual content of the scene; serves as the prompt for the image generator

IMAGE GENERATOR

SCENE ILLUSTRATION

NARRATIVE SCENE ILLUSTRATION

FRAGMENT

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.

LLM: scene captioner

LLM PROMPT

"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 ..."

SCENE DESCRIPTION: a verbalization of the visual content of the scene; serves as the prompt for the image generator

IMAGE GENERATOR

SCENE ILLUSTRATION

NARRATIVE SCENE ILLUSTRATION

FRAGMENT

Alice was getting married in a few weeks. One night, he
called and she forgot to call her back. Her mother le
message on her phone. She threatened not to co
wedding. Alice called her mother and apologized

LLM:
scene
captioner

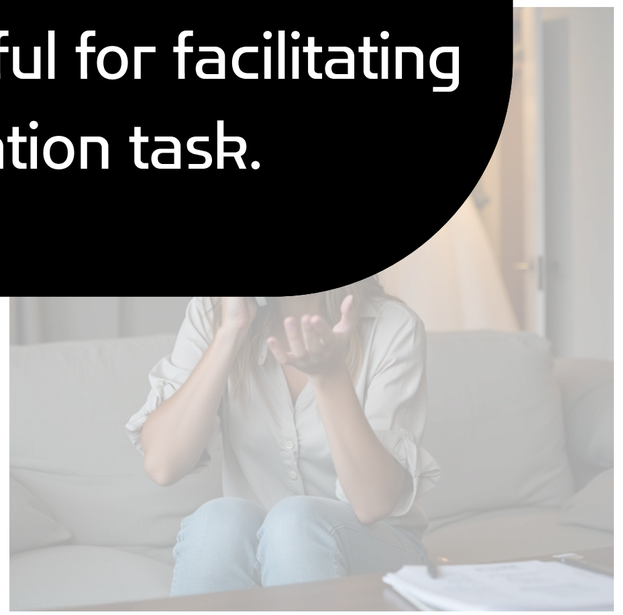
"Imagine an AI
for story fragm
image given a
image. Your tas
story context and

LLMs can articulate visual scene knowledge implicitly evoked by story text.

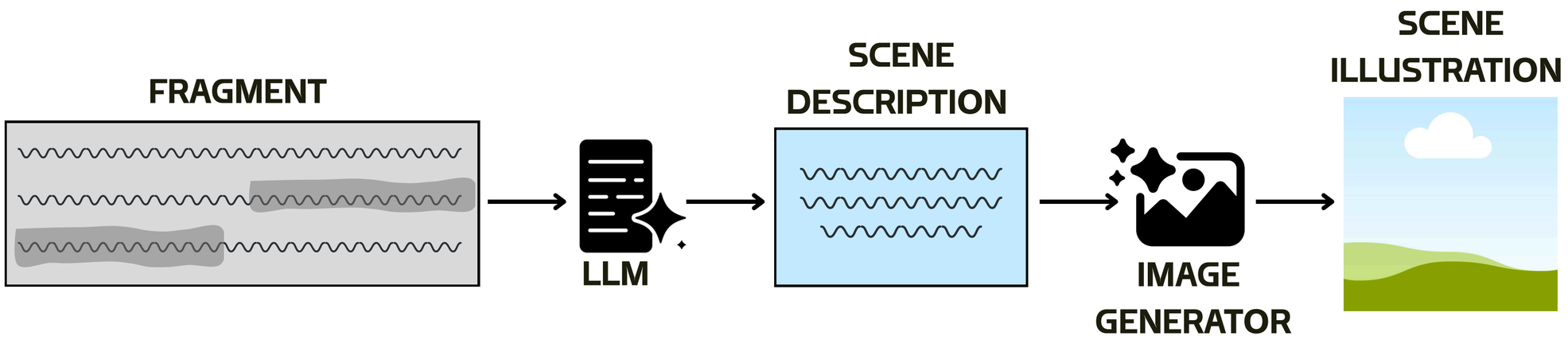
This capability is impactful for facilitating the scene illustration task.

SCENE DESCRIPTION: a verbalization of the visual content of the scene; serves as the prompt for the image generator

IMAGE GENERATOR

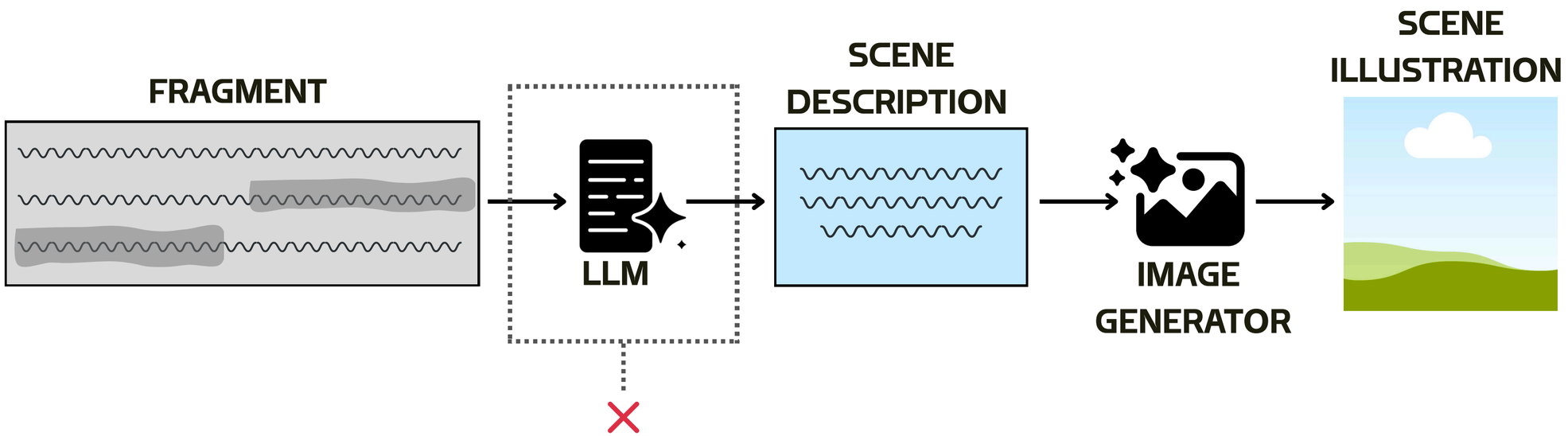


SCENE ILLUSTRATION PIPELINE



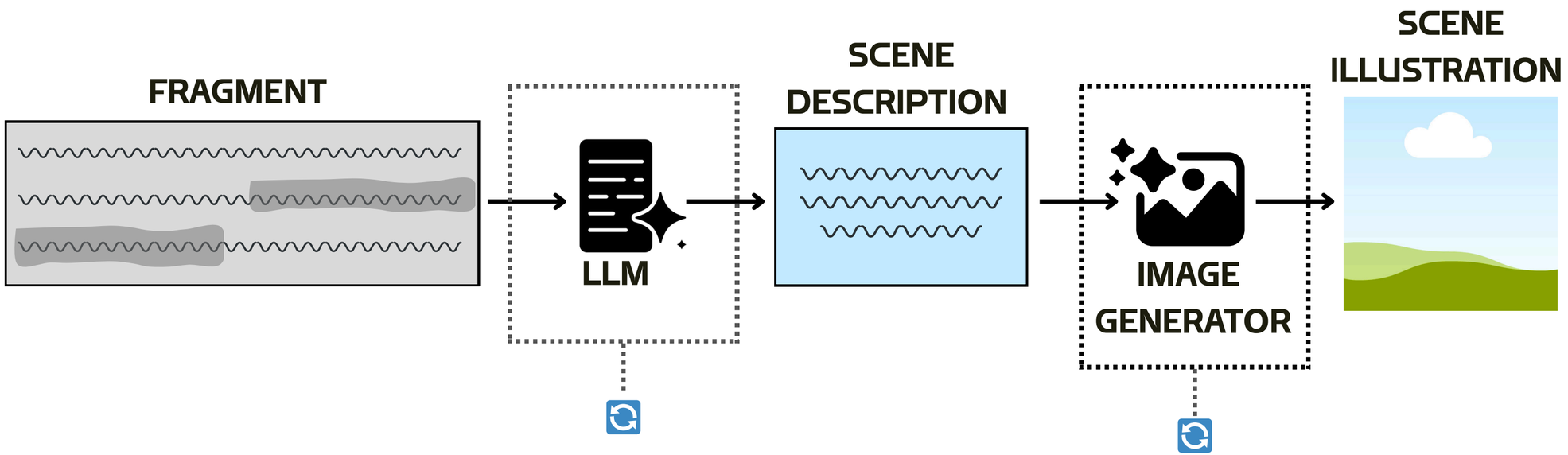
- We apply this pipeline to fragments in an existing story corpus (ROCStories)

SCENE ILLUSTRATION PIPELINE



- We apply this pipeline to fragments in an existing story corpus (ROCStories)
- When doing so, we systematically vary components of the pipeline:
  - Ablate the LLM scene captioner

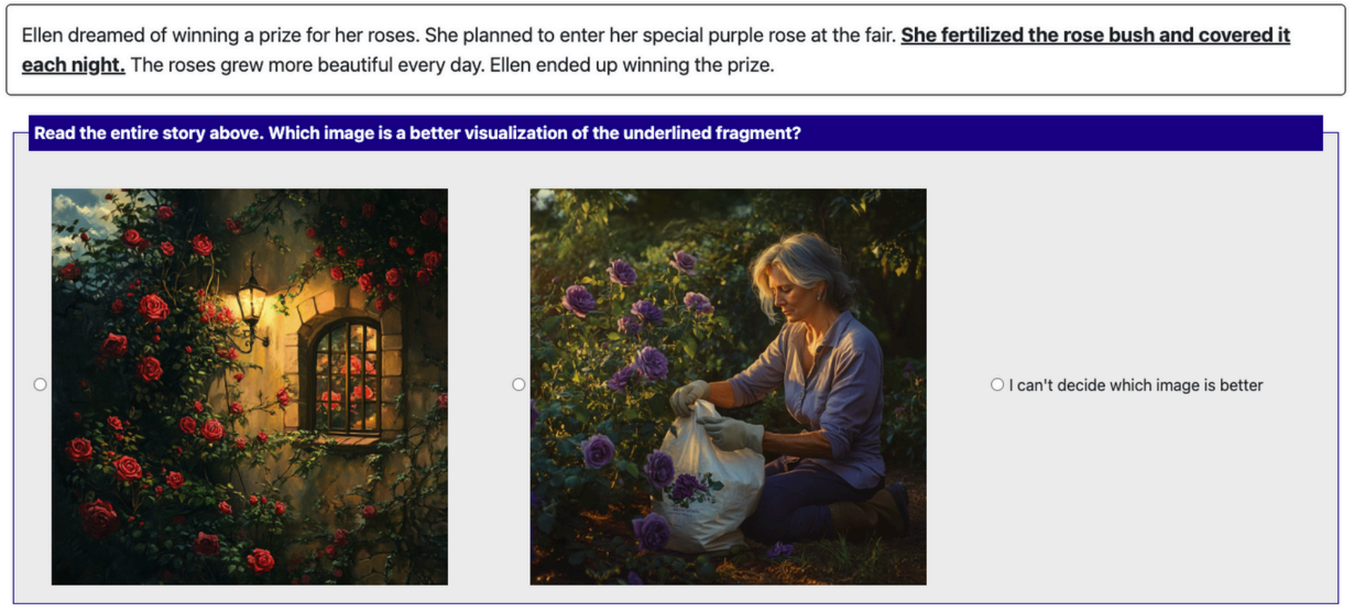
SCENE ILLUSTRATION PIPELINE



- We apply this pipeline to fragments in an existing story corpus (ROCStories)
- When doing so, we systematically vary components of the pipeline:
  - Ablate the LLM scene captioner
  - Alternate which particular models are used for the LLM scene captioner and image generator

SCENE ILLUSTRATIONS DATASET

- ~3000 items, where each item consists of:
  - A fragment (with story context)
  - 2 illustrations depicting the fragment
    - which vary based on their scene description and/or the image generator
  - Annotator judgments of which of the 2 illustrations is better (from 2 annotators)



COMPARING SCENE DESCRIPTIONS

FRAGMENT (within story context)

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.

COMPARING SCENE DESCRIPTIONS

FRAGMENT (within story context)

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.

- Baseline scene descriptions using raw story text:
  - **NO-CONTEXT**: fragment without any story context → “Alice called her mother and apologized profusely.”

COMPARING SCENE DESCRIPTIONS

FRAGMENT (within story context)

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.

- Baseline scene descriptions using raw story text:
  - **NO-CONTEXT**: fragment without any story context → "Alice called her mother and apologized profusely."
  - **VERBOSE-CONTEXT**: fragment with the entire story prepended as context → "Consider this story: {{story}}. Based on this context, illustrate this fragment of the story: [Alice called her mother and apologized profusely.]"

COMPARING SCENE DESCRIPTIONS

FRAGMENT (within story context)

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.

- Baseline scene descriptions using raw story text:
  - **NO-CONTEXT**: fragment without any story context → "Alice called her mother and apologized profusely."
  - **VERBOSE-CONTEXT**: fragment with the entire story prepended as context → "Consider this story: {{story}}. Based on this context, illustrate this fragment of the story: [Alice called her mother and apologized profusely.]"
  - **SUCCINCT-CONTEXT:** fragment that has been rewritten to summarize context → "The bride-to-be called her mother and apologized profusely for forgetting to return her call and for the resulting angry message threatening not to attend the wedding."

COMPARING SCENE DESCRIPTIONS

FRAGMENT (within story context)

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.

- Baseline scene descriptions using raw story text:
  - **NO-CONTEXT**: fragment without any story context → "Alice called her mother and apologized profusely."
  - **VERBOSE-CONTEXT**: fragment with the entire story prepended as context → "Consider this story: {{story}}. Based on this context, illustrate this fragment of the story: [Alice called her mother and apologized profusely.]"
  - **SUCCINCT-CONTEXT:** fragment that has been rewritten to summarize context → "The bride-to-be called her mother and apologized profusely for forgetting to return her call and for the resulting angry message threatening not to attend the wedding."
- Compared baselines against:
  - **CAPTION**: LLM-generated scene description → "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."

LLM-GENERATED SCENE DESCRIPTIONS YIELD BETTER ILLUSTRATIONS

- Among dataset items where one illustration used CAPTION* and the other used a baseline scene description, annotators selected the CAPTION illustration as better a significant majority of the time

Scene Description Pair	Win Rate for CAPTION (%)
CAPTION vs. NO-CONTEXT	78.1
CAPTION vs. VERBOSE-CONTEXT	74.7
CAPTION vs. SUCCINCT-CONTEXT	72.5

*using claude-3.5-sonnet as the LLM to generate CAPTION

LLM-GENERATED SCENE CRITERIA

FRAGMENT (within story context)

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.

LLM-GENERATED CRITERIA	ILLUSTRATION 1 	ILLUSTRATION 2 
The image shows a female figure (Alice).	✓	✓
Alice is holding a phone to her ear.	✓	✗
There are no other people visible near Alice.	✓	✗
Alice’s facial expression conveys regret or apology.	?	?
	= 3.5	= 1.5

LLM-GENERATED SCENE CRITERIA PREDICT ILLUSTRATION QUALITY

- Task is to predict which illustration is better, among dataset items where annotators agreed in their selection
- Prediction accuracy is higher when using LLM-generated criteria to judge illustration quality, compared to equivalent approach that doesn't use criteria

LLM Criteria Writer	Accuracy w/ Criteria (%)	Baseline Accuracy w/o Criteria (%)
CLAUDE-3.5	71.7	60.6
GPT-4O	70.1	60.2
LLAMA-3.1	68.4	59.7

*using claude-3.5-sonnet as the criteria rater VLM

SUMMARY

- LLMs can articulate visual scenes implicitly evoked by story text
- This makes them an effective interface between story text and text-to-image models in generating scene illustrations
- We release the SceneIllustrations dataset to support future work on cross-modal narrative transformation

huggingface.co/datasets/roemmele/SceneIllustrations