Experiment Report

We utilized a multimodal narrative incorporating both textual story and accompanying images. We used LLMs to annotate alternative story versions by altering crucial conditions within the narrative. Our methodology involved employing a questions and answers query method to extract underlying aspects of the alternate story, subsequently rewriting it. Each annotation consisted of six components: (1) a goal derived from the story, (2) an entity visible in the image, (3) a condition crucial for achieving the goal, (4) an altered condition contradicting the original, (5) the event leading to the altered condition (referred to as the alternate event), and (6) the rewritten story.

The annotation needs to be realistic, given the story and image. Additionally, coherence between each component of the annotation is crucial. The considerations we took to evaluate the annotation are:

- Grounding the goal within the context of the original story.
- Entities should be visually grounded in the image, crucial for goal, and not directly or indirectly mentioned in the text/story/steps
- Condition should be a circumstance pivotal to the success of the goal.
- The alternate condition contradicts the original condition, hinders the success of the goal, and involves the specified entity or its aspects.
- An alternate event is a realistic occurrence in the narrative and should be a root cause of the altered condition.
- The rewritten story should mention the alternate event, have changes that would realistically happen for that event, and imply failure of the goal.

The annotation was conducted on ARL, COIN, and V2C datasets. Each dataset was annotated using GPT-4, with the ARL dataset also annotated by Llama-7b and Mistral-7b. Llama and Mistral solely process textual inputs, whereas GPT-4 handles multimodal inputs comprising images and text.

# ARL

## GPT4V

The ARL dataset contains three sequential images and a story spanning 7-13 sentences. Annotation was completed for five different instances of image-story pairs. The **generated goals** are realistic in terms of achievability. However, there were instances where the goal lacked a focus on the agent or the person responsible for achieving it. Instead, it reads more like a summary of what needs to be done rather than a clear statement of what the individual or team intends to accomplish. For instance, “The story's goal is to highlight the appeal of a popular holiday destination.” This goal simply summarizes the story. It would have been more agent-focused if it were “people travel to the popular holiday destination.”

The second annotation is the **entity**, where GPT4 often returns an entity already mentioned in the story. There was only one instance out of five where GPT4 could find an entity visible in an image but not mentioned in the story. The cause of this problem could be because the stories are long and mention every entity in the image, or GPT4 cannot parse significant entities to the story. If the predecessors' goal and entity are excellent, annotated condition, alternate condition, and alternated event were good quality. In our pilot validation, we found that one of the rewritten stories did not imply that the goal was unsuccessful. However, changing the prompt to " rewrite the initial story with minimal changes such that the goal is not achieved" returned a new story that didn't achieve the goal.

### LLaMA

Llama annotated the ARL dataset stories using the same prompt. Llama generated more human-centered goals. For example: "the goal of the story is for people to enjoy the parade." Since there were no images and the entity is extracted from the story, Llama has a different annotation of conditions, alternate conditions, and alternate events than GPT-4. This resulted in a well-structured narrative. However, Llama had limitations in rewriting stories. Especially with longer stories, they didn't include the alternate event or were abruptly cut off.

### Mistral

The other LLM, Mistral, made appropriate annotations similar to GPT-4 annotations. In addition, the rewritten story reflected appropriate condition changes and had less text overlap with the given story. Mistral's annotation, not without an image as input, resembling those of GPT-4, raises the question of whether the narratives produced by GPT-4 predominantly rely on textual context rather than visual input.

## V2C

V2C and COIN are datasets that contain videos with short annotated captions. V2C comprises 9k videos with attribute, intention, and effect annotations. We captured frames from the videos as images and used the caption as the story and intention as a goal. We provided the goal in this annotation and didn't ask GPT4 to generate one. The other prompts were similar to questions asked when annotating the ARL dataset. This dataset has very short stories and goals, so we expected fewer issues with identifying entities. However, we encountered a problem where the entities generated weren't visible in the images instead the entity was implied in the story.

Furthermore, the brevity of the stories made rewriting them challenging. Additionally, the story only describes events in one image, resulting in other images being irrelevant to the narrative. Nevertheless, coherent conditions, alternate conditions, and alternate events were still generated.

## COIN

COIN is a database of annotated instructional videos, such as those describing lab experiments step by step. This dataset has more description than V2C but less than ARL. Despite the more descriptive stories, out of 7 annotations, we encountered 3 instances with satisfying rewritten stories. Another issue observed was during entity identification, GPT4 indicated that it could not find an entity not mentioned in the story, even if there were entities.

**In conclusion**, GPT struggles with isolating entities and understanding their significance to a story. This indicates that the revised counterfactual story depends solely on the story instead of being multimodal. Even when given a short story, **we still encounter issues with image grounding.**