## Completed Tasks (Abbreviate definition [1] [2])

- **(P1)** The state-of-the-art free Text-to-Image (T2I) model is ChatGPT. Consequently, I selected ChatGPT to fine-tune the image generation process. During prompt fine-tuning, I discovered that overly detailed or extensive contexts distract the T2I model and produce lower-quality images.
- **(P1)** I manually tried several version of T2I prompt then discovered that prompt is highly related to the quality of image produced. Thus, I created a prompt-generation model using Gemini 2.5 Pro that leverages the user story scene instead to produce an optimal T2I prompt, sparing users the need to articulate highly precise prompts themselves.
- **(P2)** I surveyed the literature and web resources on Image-to-Sound (I2S) conversion and identified two viable approaches: direct I2S or Image-to-Text (I2T) followed by Text-to-Sound (T2S). After comparison, I concluded that the latter benefits from more readily available resources.
- **(P2)** Initially, I used YOLO v12 for object detection but found its accuracy low, since the picture book styles differ greatly from YOLO's training images. I considered fine-tuning the YOLO v12 model, but this requires extensive data and annotations, which are unattainable in the short term. Fortunately, I discovered Nvidia's newly released Describe Anything Model (DAM), which can automatically detect image regions and generate textual descriptions for each.
- **(P2)** Since DAM's outputs focus on visual description, I developed another prompt-generation model using Gemini 2.5 Pro to convert these image descriptions into plausible sound effects, then fed them into a T2S model.
- **Summary** I explored the project's entire workflow and confirmed its feasibility.

## Issues and Problems

- My GPU's memory is insufficient to run the DAM model locally, requires an alternative approach or optimization in the workflow.
- Currently, the available T2S model (Elevenlabs) cannot accept sufficiently long prompts. It's output is suboptimal, frequently generating irrelevant result or sounds that diverge from the intended effect—this may stem from inherent model limitations or from semantic loss in an overly long processing chain, which hinders accurate extraction of image information and the generation of correct sound effects.
- Although the full workflow has yet to be integrated, it is already evident that system response times are excessively long (minutes even hours waiting time), as semantic data must traverse multiple generative models.

## Further Work

- Integrate all tools into a complete workflow:
  1. The user enters the desired story scene.
  2. Generate a T2I prompt and produce the scene image.
  3. Use the DAM frontend to allow the user to select image regions.
  4. Generate descriptions for selected region and convert visual descriptions into sound descriptions.
  5. Generate the corresponding sound.
- Investigate how to split the DAM model and task-specific models into smaller modules, loading only the required parts into GPU memory per invocation so that my computer can handle them.
- Explore suitable model sizes to improve response times.

## Notes

**Reference** and **AI Usage Declaration** sections are removed due to length constraint.

[1] P1: Phase 1, User Story Scene Creator
[2] P2: Phase 2, Interactable Scene Elements