Homework 1

Due: 11:59 pm on May 5

In this assignment, you will design a deep generative model for visual-textual story generation (no implementation or experiments). Given a short textual prompt describing a story idea, the model should generate (1) a full-length textual narrative and (2) a representative image illustrating the story. For example, given the input prompt “We build a Disney park on the moon”, the model should generate a detailed story expanding on this idea, as well as an image that visually captures the story's essence.

Here is an example.

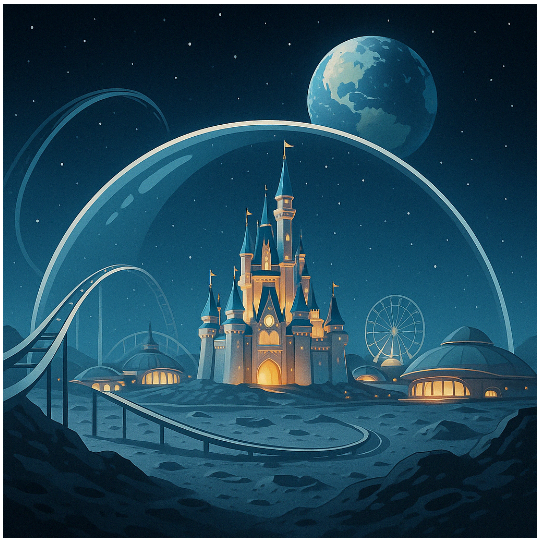
Input (Short Description):

"We build a Disney park on the moon."

Output 1 - Generated Story:

"In the year 2087, humanity completed its most ambitious theme park yet—LunaDisney, a sprawling entertainment complex built within a pressurized dome on the Sea of Tranquility. With gravity-defying roller coasters, moon-rover parades, and a Space Princess Castle that lights up the lunar sky, LunaDisney attracted families from Earth and beyond. Children floated joyfully through anti-gravity rides, while adults marveled at the Earthrise view from the Galactic Café. The park’s grand opening was broadcast across the solar system, marking a new era in interplanetary tourism."

Output 2 - Generated Image:



In summary, the model takes a short text as input and outputs a long-form story and a corresponding image. Your task is to write a 1–2 page report (in NeurIPS format) describing the design of your model and your proposed training strategy. Be sure to include: a figure illustrating the architecture of your model, and key equations that define its components and training objectives.