Dataset Plan
Prompt Battle WebGame

Introduction

The Prompt Battle WebGame requires a dataset of image-prompt pairs where the image is generated from a known “ground-truth” prompt. Each round of the game presents players with the image while hiding the original prompt. Players attempt to reconstruct the prompt under a character limit. A scoring algorithm compares their attempt against the ground-truth prompt to calculate similarity.

The dataset is therefore central to the game: its quality directly determines fairness, trust, and learning value.

Dataset Goals

1. **Provide consistent image-prompt pairs** that are suitable for accuracy scoring.
2. **Support different difficulty levels** (easy = clear objects, medium = multiple entities, hard = abstract/complex prompts).
3. **Enable transparency** by curating prompts that contain identifiable entities, styles, or descriptors that can be highlighted in feedback.
4. **Scale with future needs** (initial curated subset, later integration of larger datasets like the Stable Diffusion 100k collection).

Data Sources

1. **Curated Starter Set (Initial Sprint)**
    - Small set (20–50 pairs) generated and reviewed manually.
    - Focus on simple prompts with clear entities (e.g., “A red sports car on a mountain road”).
    - Safe and controlled for pilot testing and classroom use.

2. **External Dataset (Future Sprint, optional)**
    - *Stable Diffusion 100k Custom Prompts and Images* (Kaggle, 46GB).
    - Contains 100,000 unique prompts and corresponding images.
    - Advantage: scale and variety for replayability.
    - Risk: unfiltered data may contain unsafe or low-quality prompts. Requires filtering pipeline.

Dataset Structure

- **Prompt text**: The original ground-truth prompt (string).
- **Image file**: The generated image (PNG).
- **Metadata fields (to be added manually or via script)**:
    - Entities: objects, characters, animals, places.
    - Style descriptors: art style, lighting, medium.
    - Complexity rating: easy, medium, hard.
    - Safety flag: safe / unsafe.

This structure ensures the scoring system can highlight overlaps and compute similarity beyond raw text.

Curation & Filtering Strategy

1. **Manual Filtering for Starter Set**
    - Remove unsafe, biased, or offensive prompts.
    - Ensure each prompt describes something clearly visible in the image.
    - Add labels for entities and styles.

2. **Automated Filtering for Larger Dataset**
    - Regex and keyword filters for banned terms.
    - Length filters (exclude extremely long or extremely short prompts).
    - Entity extraction with NLP (spaCy or similar) to prepare labels for scoring transparency.
    - Teacher mode: restrict dataset to curated safe subset.

Difficulty Levels

- **Easy**: Prompts with 1–2 clear objects (“A cat on a sofa”).
- **Medium**: Prompts with multiple elements and styles (“A futuristic city skyline at sunset, digital painting”).
- **Hard**: Prompts with abstract or artistic phrasing (“A surreal dreamscape in the style of Salvador Dalí”).

Each round can be assigned a difficulty to balance learning and competition.

Risks

1. **Ambiguous Prompts**: some dataset items may describe elements not visible in the image.
   Mitigation: manual curation for pilot dataset, auto-filtering later.

2. **Unsafe or NSFW Content**: large-scale datasets may contain inappropriate prompts.
   Mitigation: filtering pipeline and manual review for classroom mode.
3. **Over-complexity**: very long prompts may overwhelm players.
   Mitigation: apply character-length filters.

Validation Plan

- Check that every image–prompt pair loads correctly.
- Validate that extracted entities match visible objects in images.
- Pilot test with classmates to confirm fairness (players should be able to guess main elements).
- Collect feedback on whether prompts felt too easy, too hard, or unclear.

Future Integration

- *Phase 1:* 20-50 curated prompts for MVP.
- *Phase 2:* Expand to 500-1000 filtered items from Kaggle dataset for replayability.
- *Phase 3:* Develop a difficulty-balancing system using metadata.
- *Phase 4*: Optional daily challenge pulling from a rotating subset of the Kaggle dataset.

Links to Learning Outcomes

- **Creating professional IT products**: The dataset is part of realising a functional game loop.
- **Professional standard**: Filtering, labelling, and validation ensure quality and safety.
- **Orientation**: Demonstrates awareness of scaling from a curated subset to a large external dataset.