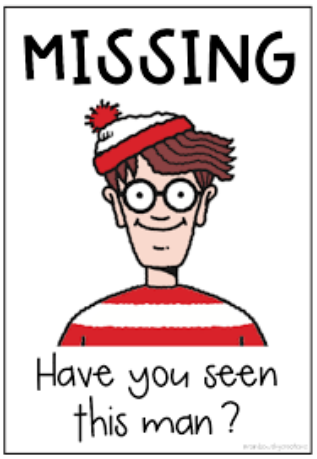
# Final Project Proposal
# Wally-like Storytelling Scene Interactor

施淙綸
110020007

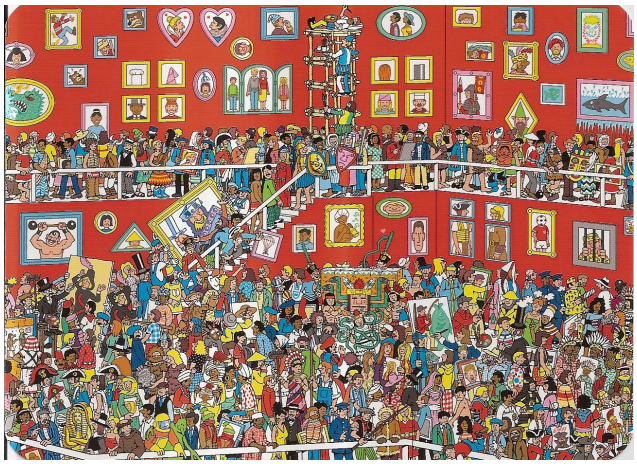
April 15, 2025

## 1 Introduction

*Where's Wally* (中譯：《威利在哪裡？》) kept me company during many after-school hours when my parents were late picking me up. I remember there was always an opened sample copy on the bookstore shelf to entertain us kids. This book series was created by English illustrator Martin Handford [1]. The objective of the book is to find Wally, who is wearing a red-striped shirt, a bobble hat, and carrying a walking stick, within a crowded scene.




(a) Charater Wally



(b) Book content: a group of people in the museum

Figure 1: The famous children's book *Where's Wally* created by Martin Handford.

### 1.1 Why *Where's Wally*?

The interactive nature of the Where's Wally series is what I find most captivating. It employs a simple presentation style, yet it possesses a unique magic that compels us to flip through the pages, searching for Wally hidden in diverse and imaginative scenes. While the moment of discovering Wally brings immense joy, the journey itself is equally rewarding, offering glimpses into the lives of a myriad of characters engaging in various activities within bustling locales like buildings, beaches, city streets, schools, and parks. For instance, there's that dog owner chasing after their runaway pet, their ice cream cone tragically tumbling to the ground mid-pursuit. And what exactly are those people gathered in a circle so intently observing? Perhaps they've found a rare coin or maybe a street performer is about to start his show.

# 2 Two Phases Plan and Goal

## 2.1 Phase 1 (P1): User Story Scene Creater

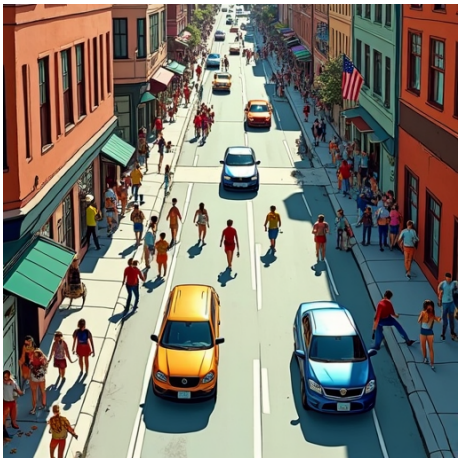
It's these hidden details interwoven within the vibrant scenes, and the unfolding stories we uncover while hunting for Wally, that contribute to the enduring success and appeal of the series. Therefore, for first phase of this assignment, I aim to leverage AI technology to reimagine the Where's Wally storytelling model. My vision is to allow users to immerse themselves in existing narratives or ignite their own creative spark, generating images brimming with life. Furthermore, I want to find possible method to integrate a diverse array of puzzles and hidden objects within these new creations, offering an even richer and more engaging experience.

- **Possible Methods:**

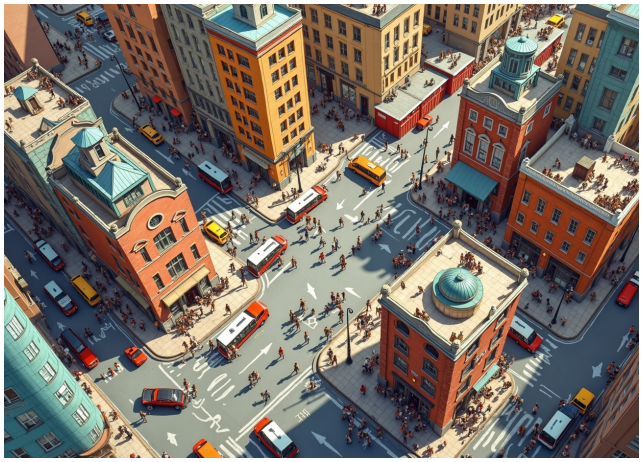
  1. Deconstruct the narrative of the story scene into descriptive text prompts for different scenes. Use existing Text-to-Image (T2I) AI models to generate corresponding scenes.
  2. Following the *Where's Wally* Storytelling Model, the generated images should fully and vividly present details unrelated to the main character, and the main character should be cleverly integrated into the image.
  3. The main character of the image can be freely set by the user, such as a specific animal, plant, portrait, etc., acting as a pre-agreed-upon "signal" or "key image" with the reader, rather than being limited to the Wally character.

- **Potential Problems and Challenges (Demo in Figure 2):**

  1. The image quality generated by free T2I AI models may not be high enough. It may be difficult to achieve the level of detail and finesse found in *Where's Wally*, and the ability to integrate the main character into the image may not be as good as expected.
  2. If the description of the main character in the prompt is incomplete or not specific enough, the model may not successfully generate that part.
  3. Many models cannot accept overly long prompts (such as Prompt 2 in Appendix), and overly detailed prompts may also distract the model.



(a) See Prompt 1 in Appendix



(b) See Prompt 2 in Appendix

Figure 2: Demo of story image generation attempt.

### 2.2 Phase 2 (P2): Interactable Scene Elements

I've also been pondering how cool it would be to add simple interactive features if we could freely create our own story worlds. For example, imagine a picture with a cute puppy sleeping. When you click on it with your mouse, it wakes up! It lifts its head, looks around to see who disturbed it, and then flops back down to sleep. Or, when you click a bell hanging from a tree, it rings with a clear, bright sound. These interactive widgets could make the story scenes much more vivid, and I think AI tools like Text-to-Sound (T2S), Image-to-Sound (I2S) or Image-to-GIF (I2GIF) could be used to achieve this.

- **Possible Methods:**

  1. Use an object detector model like YOLO to identify various objects within the image and assign text labels to them.
  2. Input the text labels from each identified object into a text-to-sound (T2S) model to generate corresponding sound effects. Then, configure the objects to play the appropriate sound effect when clicked.
  3. If needed, crop the sub-image of the identified object and input it into an I2GIF model to generate a simple animated GIF. The GIF content will then play when the mouse interacts with the object.

- **Potential Problems and Challenges:**

  1. Incomplete images generated in P1 may not be successfully identified by the object detector. For example, overly abstract depictions of birds, dogs, or trees that humans can recognize but the model cannot.
  2. Overlapping objects may create strange, distorted movements during the I2GIF process.
  3. Generating too many details might make it difficult to maintain stylistic consistency across the interactive elements.
  4. Fully implementing this phase is likely unachievable within the time constraints of the assignment.

### 2.3 Restriction and Expected Results

While realizing both phases simultaneously and creating a smooth, user-friendly application is a very attractive idea, my time on the Final Project may not be enough to achieve that scale. Therefore, I will firstly and mainly focus on the T2I implementation in P1, attempting to create an image with interesting details within a specific story. If time allows, I will move on to P2, with the main goal of creating a simple demo of the interactive scene.

## 3 Schedule

| Week | Date | Event | Goal |
|---|---|---|---|
| 9 | 4/14 ~4/20 | Proposal Submit | Reference documents research |
| 10 | 4/21 ~4/27 | | P1: find proper T2I model and fine-tune prompt |
| 11 | 4/28 ~5/04 | | P1: fine-tune detail generation (1/2); Progress Report |
| 12 | 5/05 ~5/11 | Progress Report | P1: fine-tune detail generation (2/2) |
| 13 | 5/12 ~5/18 | | P2: try object detection; P1 in Final Report |
| 14 | 5/19 ~5/25 | | P2: T2S generation and alignment; Final Report |
| 15 | 5/26 ~6/01 | Final Report | (The following development...) |
| 16 | 6/02 ~6/06 | Semester End | (The following development...) |

# AI Usage Declaration

The content was initially drafted in English or Chinese by me, with subsequent refinements to content, grammar and clarity made using "Gemini 2.0 Flash" in Google AI Studio. Then, finally, I manually tune the content again before hand in to avoid off-topic and wrong imagination of GAI.

# Appendix

## Prompt 1

> Generate image: Overhead view of a bustling city street, "Where's Wally" style. Diverse people, vehicles, and objects fill the scene. Playful atmosphere with bright, saturated colors. A calico cat is subtly hidden. High detail, intricate and fun. Find the hidden cat!
>
> — Prompt 1 to Generate Figure 2a

## Prompt 2

> Detailed overhead view illustration in the style of "Where's Wally" (also known as "Where's Waldo"), without any text.
>
> Scene: A bustling and vibrant cityscape viewed from directly above. The street is teeming with a multitude of people of all ages, sizes, and ethnicities, each engaged in various activities. Include a diverse array of buildings, vehicles (cars, buses, bicycles, motorcycles, etc.), street furniture (benches, lampposts, signs, trash cans, flower planters), and other miscellaneous objects to create a rich and complex visual tapestry.
>
> Atmosphere: The overall atmosphere should be playful, whimsical, and engaging. Use bright, saturated colors and intricate details to capture the viewer's attention and invite them to explore the scene.
>
> Key elements:
>
> People: Depict hundreds of people in various poses and outfits. Some should be walking, running, talking, shopping, eating, playing, and generally interacting with their surroundings. Add some unusual characters like someone juggling, wearing a funny costume, or doing something unexpected.
>
> Objects: Scatter a wide variety of objects throughout the scene. Include everyday items like umbrellas, shopping bags, balloons, newspapers, and food carts. Also add some more quirky and unexpected objects like a misplaced traffic cone, a giant inflatable duck, or a street performer's props.
>
> The Hidden Cat: A small, calico cat is subtly hidden somewhere within the scene. The cat should blend in with its surroundings but still be visible upon close inspection. The cat's pose can be relaxed, curious, or even a little mischievous. Consider placing it in a slightly unexpected location (e.g., on a rooftop, in a window, behind a bush).
>
> Style:
>
> Inspired by the detailed and intricate style of "Where's Wally" (Where's Waldo).
>
> Use clean lines, bold colors, and a high level of detail to create a visually appealing and engaging image.
>
> The overall aesthetic should be fun, lighthearted, and accessible to a wide audience.
>
> Technical Considerations:
>
> Viewpoint: Top-down (overhead) perspective.

Color Palette: Bright, saturated, and varied.

Level of Detail: Extremely high – the more details, the better.

Resolution: High resolution for optimal clarity and detail.

file as style and content ref

— Prompt 2 to Generate Figure 2b

## References

[1] 维基百科, "威利在哪里?— 维基百科, 自由的百科全书." 2024. Available: https://zh.wikipedia.org/w/index.php?title=%E5%A8%81%E5%88%A9%E5%9C%A8%E5%93%AA%E9%87%8C%EF%BC%9F&oldid=85021366