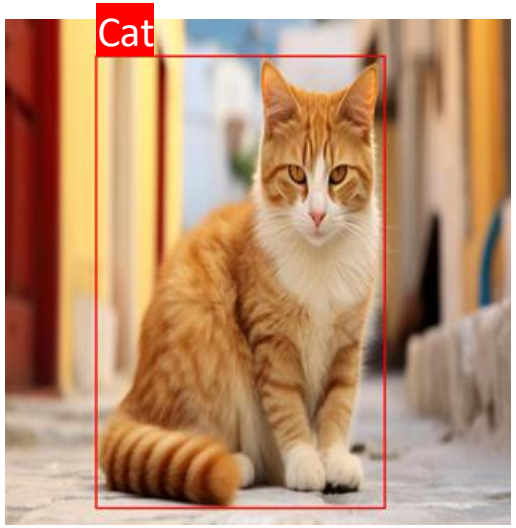
Find and Replace What You Want

R12631013 鄭朝鴻 | R12631054 朱王文亮 | R12631027 楊佩錡 | R12631001 許喬淇 | R12631012 連震宇

Overview

- We aim to develop an **object replacement tool** that allows users to **swap objects** in photos simply by providing descriptions on user interface.
- In addition to object replacement, this system can **also perform object detection and image captioning**.

I want to …

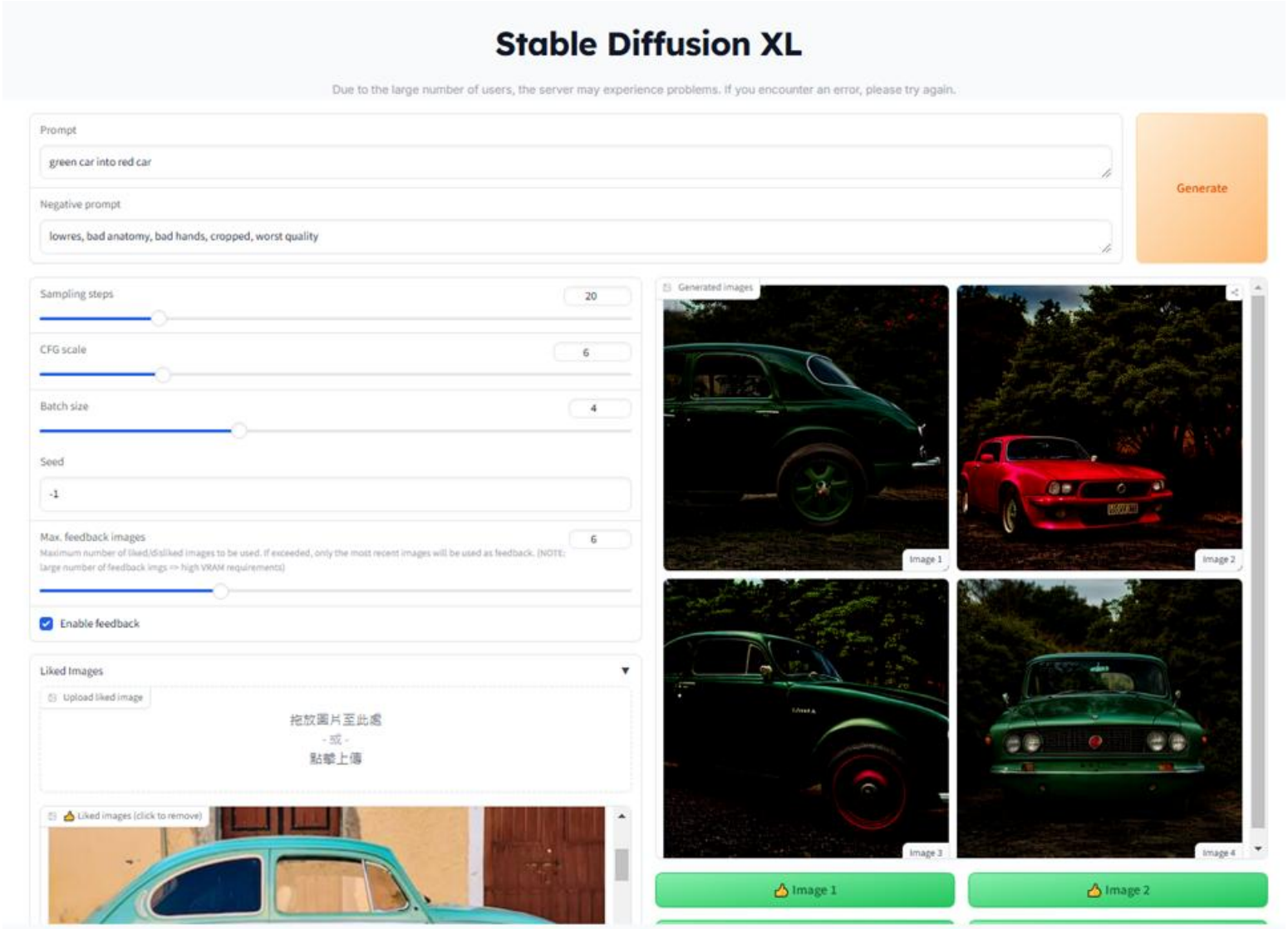


Motivation & Objective

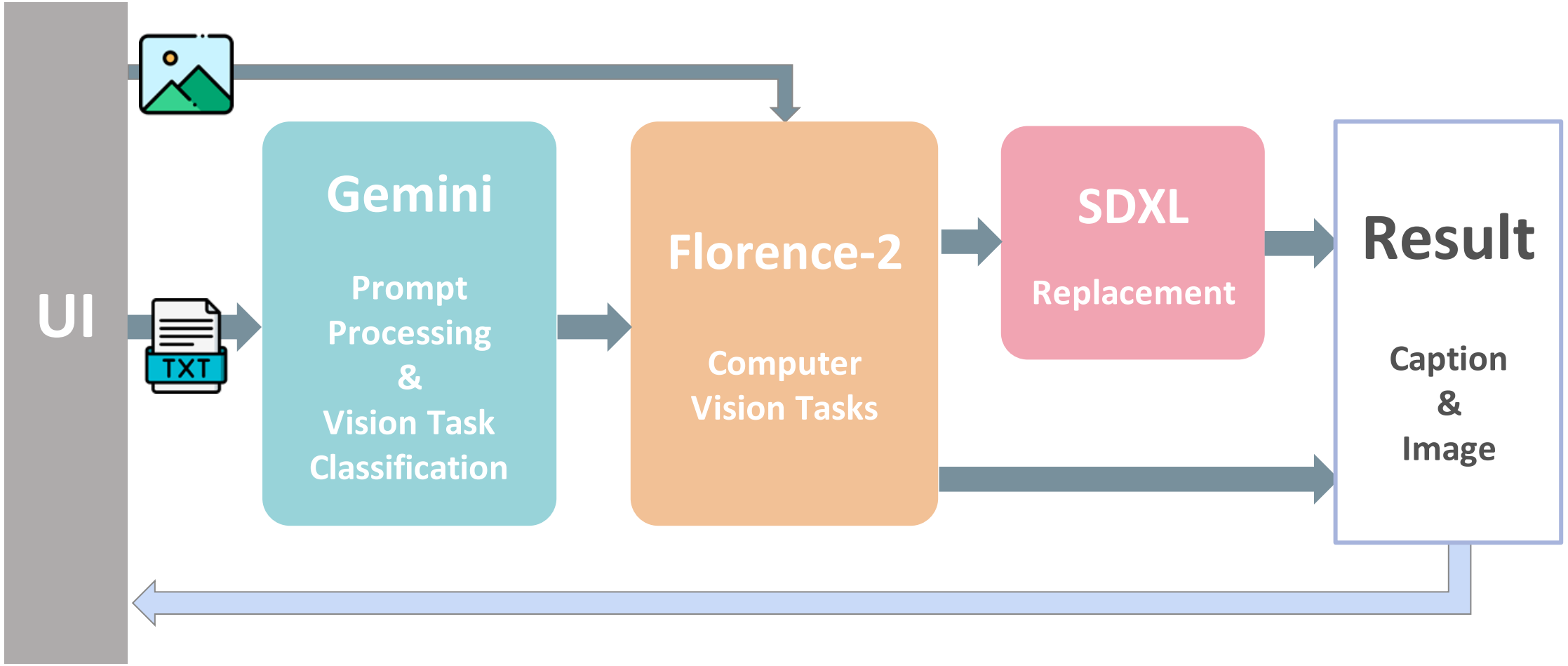
- **Motivation**
  - Object Editing
  - Simplifying Inpainting
  - Mask Generation

- **Objective**
  - Create a user-friendly web ui for inpainting that requires no technical expertise.

Related Work – Stable Diffusion WebUI



Methodology



Gemini

- Identifying task type from user's input prompt using Gemini.

Vision Foundation Model: Florence-2

2024 CVPR - Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks

Image

<Task Prompt> & <Description Prompt>

Ex:

<Caption>

What does the image describe?

<Object Detection>

Locate the phrases in the caption: A woman riding a bike.

<Segmentation>

Target mask: Bike rider

Transformer Encoders

Transformer Decoders

text + location tokens

Caption

A person riding a red bike on a road.

Bounding Boxes

{'bboxes': [[34.23999786376953, 159.1199951171875, 582.0800170898438, 374.6399841308594]]}

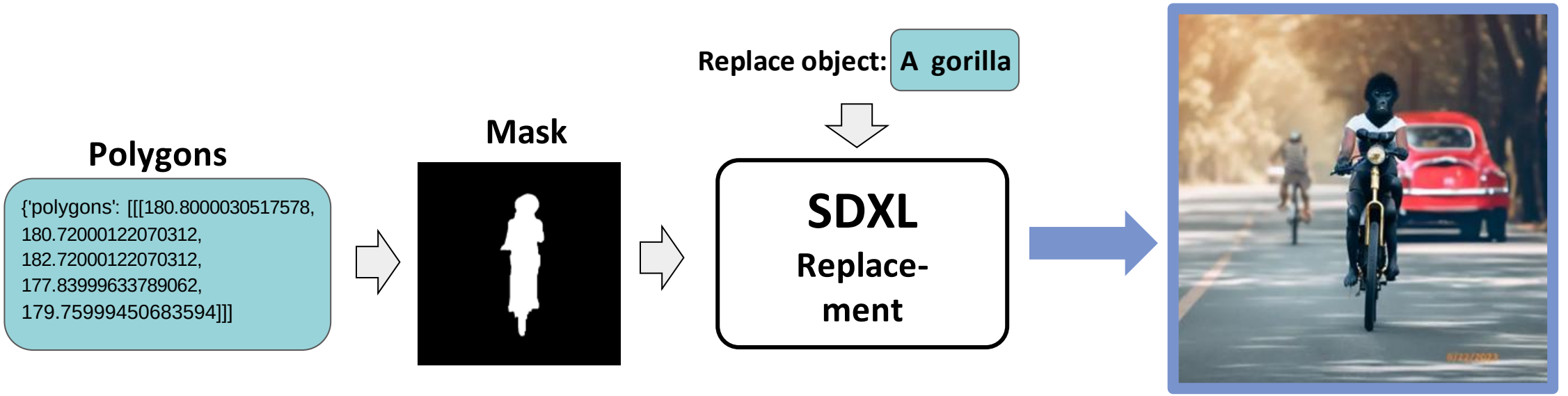
Polygons

{'polygons': [[[180.8000030517578, 180.72000122070312, 182.72000122070312, 180.72000122070312, 187.83999633789062, 177.83999633789062, 189.75999450683594, 177.83999633789062,179.75999450683594]]]

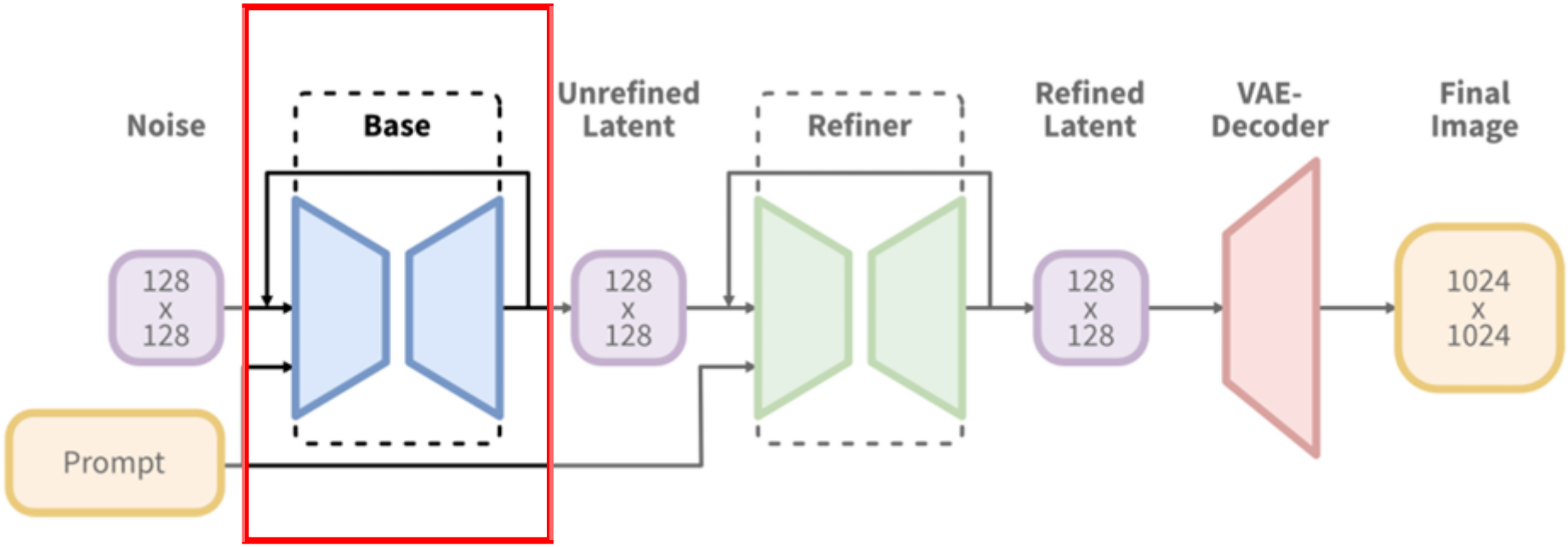
Inpainting: Stable-Diffusion XL

2024 ICLR - SDXL: Improving Latent Diffusion Models For High-resolution Image Synthesis

- SDXL (Stable Diffusion XL) is an advanced diffusion model designed for generating high-resolution, detailed images.
- If Gemini classifies the task as segmentation, SDXL processes segmented regions.

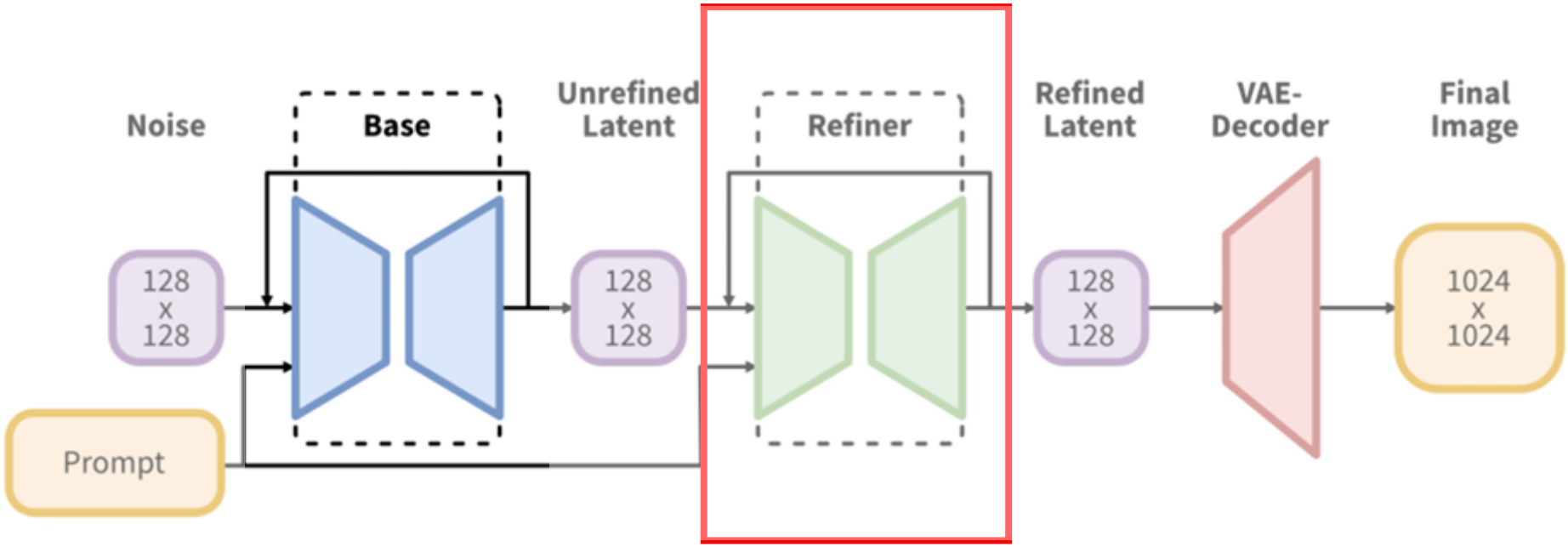


Inpainting: Stable-Diffusion XL



- **Increase UNet size:** Add attention blocks and second text encoder
- **Utilize Conditioning techniques:** More flexible image synthesis

Inpainting: Stable-Diffusion XL



- **Refinement Model:** Increase local image quality

Experiment Results

- **Demo Video**
- **Comparing Inpaint Results Given Different Prompts**
  - **Short & Simple Prompt vs. Detailed Prompt**
- **Execution Time**

CVPDL Final Project

Please enter the instructions you want to do with the image and upload the image! We provide three services: image inpainting, object detection and image captioning.

question

output 0

input_image

拖放圖片至此處

- 或 -

點擊上傳

output 1

Clear

Submit

Flag

Simple Prompt vs. Detailed Prompt

Short and Simple Prompt: Inpaint the dog with a robot

Simple Prompt vs. Detailed Prompt

Detailed Prompt: Inpaint the dog with a silver dog-shape robot

Execution time

- Testing execution time of segment and inpainting task on 20 samples.

Prompt Processing

1.3 s

Image Segmentation

2.0 s

Image Inpainting

4.9 s

Average Execution Time: 8.2 s

Limitation

- **Mask Accuracy:**
  - The segmentation produced by Florence struggles to capture precise object edges and contours.
- **Unsatisfactory Generation:**
  - Generation diversity is limited by the training data, making it impossible for users to generate objects that weren't included in the training set.
- **Resource Constraints:**
  - Limited computational resources prevent us from using larger models, which leads to slower UI response times.

Conclusion

- We have successfully developed a **object replacement tool** that can perform **object replacement, identify objects, and predict captions** through text descriptions.

- This tool uses **Gemini** to analyze the provided descriptions and feeds the distilled information into **Florence-2 and SDXL** to accomplish multiple vision tasks.

- The proposed tool takes only **8.2 seconds** on average to perform object replacement tasks.


Thanks for Listening