# Find and Replace What You Want

Chao-Hung Jeng
*Department of Biomechatronics Engineering*
*National Taiwan University*
Taipei, Taiwan
r12631013@ntu.edu.tw

Wen-Liang Chu Wang
*Department of Biomechatronics Engineering*
*National Taiwan University*
Taipei, Taiwan
r12631054@ntu.edu.tw

Pei-Chi Yang
*Department of Biomechatronics Engineering*
*National Taiwan University*
Taipei, Taiwan
r12631027@ntu.edu.tw

Chiao-Chi Hsu
*Department of Biomechatronics Engineering*
*National Taiwan University*
Taipei, Taiwan
r12631001@ntu.edu.tw

Jen-Yu Lian
*Department of Biomechatronics Engineering*
*National Taiwan University*
Taipei, Taiwan
r12631012@ntu.edu.tw

***Abstract*—Object replacement in image editing remains a challenging task, as existing tools excel at object removal but lack intuitive methods for substituting objects with user-defined alternatives. Additionally, many solutions require technical expertise and lack a user-friendly interface, limiting accessibility for general users. This paper presents a novel framework integrating Gemini, Florence-2, and Stable Diffusion XL to address these limitations. Gemini processes natural language prompts to classify tasks, while Florence-2 performs image captioning, object detection, and segmentation. Stable Diffusion XL enables high-resolution and realistic object replacement. The proposed system introduces a streamlined user interface that simplifies complex editing workflows, requiring only an input image, task category, and descriptive prompt. Experimental results demonstrate an average execution time of 8.2 seconds per task, highlighting the system's efficiency and usability. This work highlights the ability of Vision Foundation Models to connect advanced AI technologies with user-focused applications in image editing.**

***Keywords—image captioning, object detection, segmentation, inpainting***

## I. INTRODUCTION

The rise of Vision Foundation Models has transformed the field of computer vision, offering unprecedented flexibility and efficiency in handling diverse visual tasks. Florence-2, a state-of-the-art Vision Foundation Model, empowers users to perform tasks such as object detection, segmentation, and captioning with minimal input: an image, a task category, and a descriptive prompt [1]. This simplicity eliminates the need for manual annotations or extensive technical expertise, making advanced computer vision accessible to a broader audience.

Complementing Florence-2's capabilities is Gemini, a task-processing model that interprets user prompts and classifies them into specific vision tasks. Gemini's ability to bridge natural language inputs with precise vision task categorization ensures seamless communication between users and the system [2]. Together, these models form the foundation of our system, offering powerful functionality for simplifying complex editing workflows.

However, existing image editing tools often lack intuitive solutions for user-defined object replacement. While tools like Apple Intelligence excel at object removal, they fail to offer functionalities for replacing removed objects with user-specified alternatives, such as swapping a cup for a flower vase. This limitation underscores a critical gap in current systems, emphasizing the need for practical and user-centric applications of vision models to achieve more comprehensive image editing capabilities.

To address this challenge, we propose a unified system that integrates Gemini, Florence-2, Stable Diffusion XL, and a user-friendly web UI, enabling users to perform object replacement without technical expertise. This approach enhances accessibility and usability while maintaining high performance and timely feedback.

## II. RELATED WORK

### A. Limitations of Florence-2

Florence-2 employs a sophisticated data engine that integrates three stages to facilitate its multi-task learning capabilities [1]. As seen in Fig. 1, First, it utilizes various specialist models, where they work together to autonomously annotate images, ensuring high-quality and reliable data generation [1]. Following the first stage, these annotations are filtered and enhanced with specialized LLMs. A Refinement Module, then iteratively enhances these annotations using well-trained foundational models [1]. Together, these modules create the extensive FLD-5B, allowing Florence-2 to handle various tasks such as object detection, image captioning, and visual grounding [1]. Despite the rich training dataset, it lacks the ability to differentiate tasks, users must specify the target task beforehand [1]. In addition, it is best to use a standard prompt format similar to its training data annotations for better performance. These limitations present a steep learning curve for first-time users and reduce the interactivity of a multi-task computer vision tool [1].

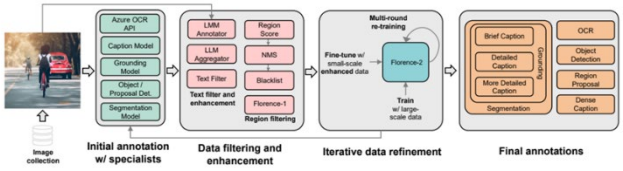


Fig. 1. Framework of Florence-2 data engine. Derived from [1].

### B. User Interface for image generating tools

To put image generating tools into full use, providing a user interface (UI) for applications like Stable Diffusion [3] is crucial for enhancing user experience and accessibility. A well-designed UI allows users, regardless of their technical expertise, to interact effectively with complex AI models, such as those used for generating and modifying images from text prompts. The Stable Diffusion Web UI [3] (Fig. 2) serves as a great example, displaying the options and parameters one can tinker with to generate better images. By studying such existing interfaces, we can gain insights into best practices for

building our own UIs, ensuring they cater to user needs while simplifying the interaction with advanced technologies. In our opinion, an ideal user interface (UI) for image generation applications like Stable Diffusion XL (SDXL) [4] is one that prioritizes simplicity and efficiency, allowing users to input instructions as prompts and receive results without the need to adjust numerous toggles and parameters. Removing the technical complexities of using an image generation model, users can freely experiment with creative ideas.

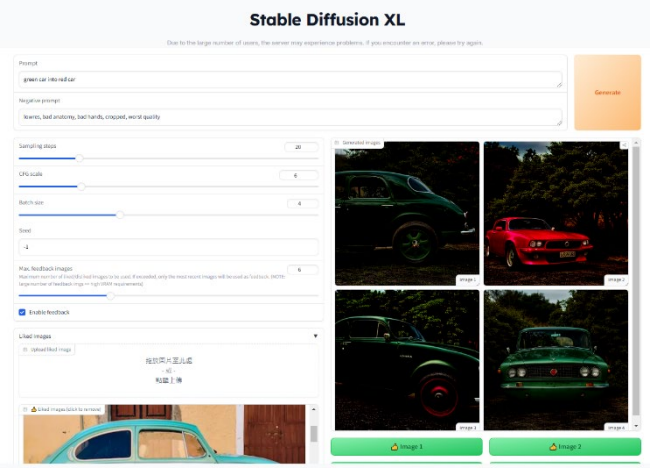


Fig. 2. User interface for SDXL. Derived from [4].

## III. METHODOLOGY

### A. Overview

In our methodology, we utilized three models - Gemini, Florence-2, and Stable Diffusion XL - along with a custom-built UI (Fig. 3). Users can upload both images and text through the UI interface. The text input is first processed by Gemini for semantic analysis, generating a refined prompt. This prompt, together with the image, is then fed into Florence-2 for computer vision tasks. For our project's object replacement objective, we implemented a two-stage process: initial preprocessing by Florence-2, followed by image generation using Stable Diffusion. Finally, all results are displayed on the UI.

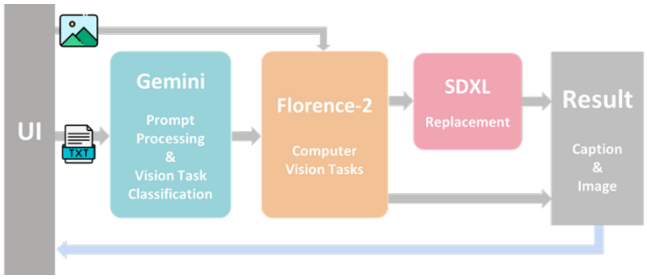


Fig. 3. Methodology overview.

### B. Gemini

In this study, we utilized Google's Gemini 1.5 Pro model, which is a multimodal large language model, to handle users' input prompts. The model demonstrates superior performance across various benchmarks and exhibits strong capabilities in understanding complex visual-linguistic relationships.

For our research methodology, Gemini was used to develop an intelligent prompt classification system (Fig. 4) for computer vision tasks. The model analyzes user prompts and categorizes them into three fundamental computer vision operations: image captioning, object detection, and image segmentation. According to classification results, the model generates different task prompts and description prompts for directing appropriate vision processing of further steps.


Fig. 4. Google's Gemini as a prompt classification model for computer vision tasks.

### C. Florence-2

Florence-2 is a vision foundation model introduced by Microsoft in 2023 [1]. While its predecessor, Florence, focused primarily on transfer learning, Florence-2 represents a significant advancement in its ability to perform various vision tasks (including image captioning, object detection, visual grounding, and segmentation) through simple text prompts [1]. Architecturally, Florence-2 uses a vision encoder to convert images into visual token embeddings, which are then concatenated with text embeddings and processed by a transformer-based multi-modal encoder-decoder to generate the response (Fig. 5) [1]. In this stage, we input the generated task prompt and description prompt into Florence-2 to perform vision tasks. For the replacement task, we used Florence-2 to segment the removed object, allowing us to obtain its polygons.

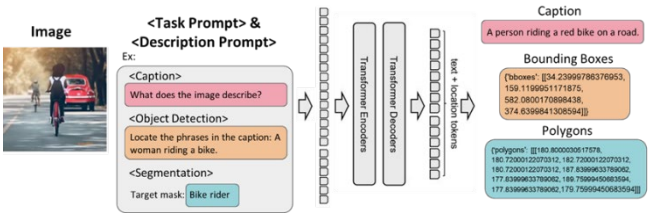


Fig. 5. Florence-2 architecture.

### D. Stable Diffusion XL

In this study, we employed Stable Diffusion XL (SDXL), a state-of-the-art text-to-image diffusion model released by Stability AI in 2023 [4]. SDXL represents a significant advancement over its predecessors, featuring an enhanced architecture with dual text encoders and a larger model size that enables higher-quality image generation and manipulation [4]. The model demonstrates superior capabilities in understanding complex text prompts and generating detailed, coherent visual outputs [4].

For our research methodology, we implemented an object replacement pipeline (Fig. 6) that combines SDXL with Florence-2's segmentation capabilities. The process begins with obtaining masks through Florence-2's semantic segmentation of the input images. These masks, along with the original images, are then fed into SDXL's inpainting pre-trained model. SDXL processes these inputs in conjunction with user-provided text prompts to generate appropriate modifications within the masked regions. This approach leverages SDXL's advanced understanding of spatial relationships and semantic content to ensure the inpainting

areas maintain visual coherence with the surrounding image context while adhering to the user's specified requirements.

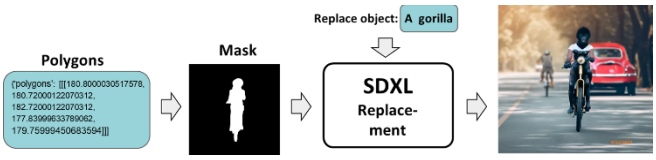


Fig. 6. Object replacement pipeline with SDXL.

## IV. Experimental Results

We conducted performance testing using an NVIDIA RTX 3090 GPU for three tasks: image captioning, object detection, and object replacement. Our experiments revealed significant variations in processing times across these tasks. Image captioning (Fig. 7) and object detection (Fig. 8) demonstrated relatively shorter processing times, with tests on 20 images yielding an average processing time of 2.5 seconds per image. In contrast, object replacement (Fig. 9) required considerably more computational time, averaging 8.2 seconds per image across the same 20-image test set. Within the object replacement pipeline, the SDXL image inpainting step proved to be the most time-consuming component, requiring an average of 4.9 seconds per operation.

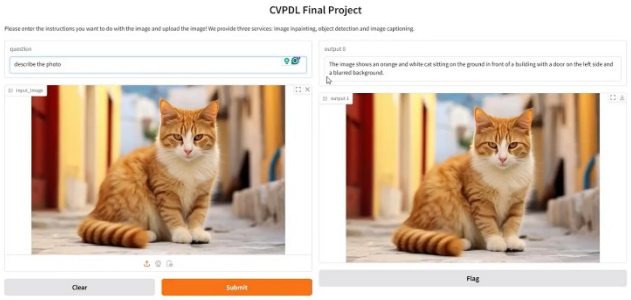


Fig. 7. Image captioning.

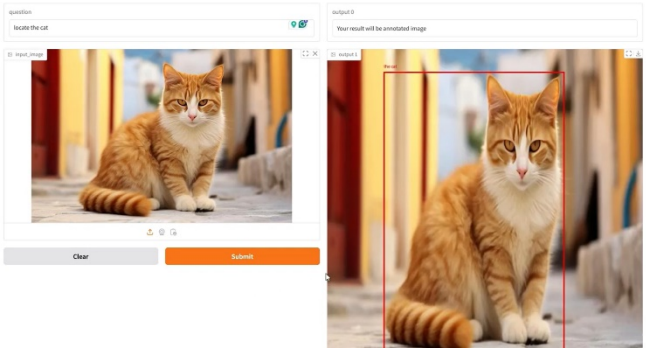


Fig. 8. Object detection.

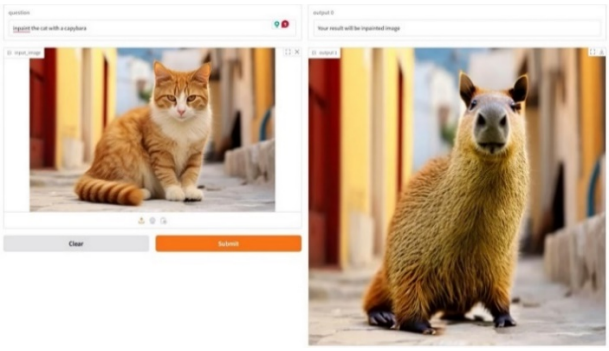


Fig. 9. Object replacement.

Furthermore, for object replacement tasks, we investigated the impact of prompt complexity on generation quality. Our experiments compared the outcomes between simple, concise prompts and detailed, comprehensive prompts. The results demonstrated that simple prompts often led to generated objects with distorted geometric structures that notably deviated from their real-world counterparts (Fig. 10). Conversely, detailed prompts produced objects that more closely aligned with user expectations and exhibited better geometric consistency with real-world objects (Fig. 11).

Fig. 10. Object replacement using a simple, concise prompt: Replace the dog with a robot.

Fig. 11.Object replacement using a detailed, comprehensive prompt: Replace the dog with a silver dog-shape robot.

## V. Conclusion

We have successfully developed an object replacement tool that can perform object replacement and many vision tasks through a user-friendly UI. 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 timely object replacement tasks.

## References


[1] B. Xiao *et al.*, "Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 4818-4829, 2023.

[2] M. Reid *et al.*, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *ArXiv,* vol. abs/2403.05530, 2024.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 10674-10685, 2021.

[4] D. Podell *et al.*, "SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis," *ArXiv,* vol. abs/2307.01952, 2023.

## APPENDIX

*Team Contribution*

| Team member | Contribution |
|---|---|
| 鄭朝鴻 | 1. Using the Gemini model to design a prompt classification system for computer vision tasks.<br>2. Slide for Gemini methods and UI example usage video.<br>3. Presentation of overview and motivation.<br>4. Methodology and experiment results section of the final report. |
| 楊佩錡 | 1. Slides for overview and system pipeline.<br>2. Video recording for overview, motivation, Gemini, Florence-2 and Stable diffusion XL.<br>3. Report writing for Abstract and Introduction. |
| 朱王文亮 | 1. Slides for SDXL model architecture.<br>2. Presenter of demo, results, limitation, conclusion.<br>3. Video recording for demo, results, limitation, conclusion.<br>4. Report writing for Related Works. |
| 許喬淇 | 1. Integrate Florence-2 and stable diffusion xl.<br>2. Slide production of method and results.<br>3. Report production of methodology, conclusion, and appendix.<br>4. Presentation of related work and methods. |
| 連震宇 | 1. Design UI, backend connection and deploy to server.<br>2. Slides for Motivation & Objective, Related Work, Limitation, and Conclusion.<br>3. Report compilation, proofreading, and formatting.<br>4. Upload video to YouTube. |

*Demo Video*: https://www.youtube.com/watch?v=PeEOi64qpFM

*Initialize Models*

The core vision-language model Florence-2-large was obtained from Microsoft's Hugging Face repository using the transformers library. For image generation (replacement task), we employed the diffusers library, specifically implementing the Stable Diffusion XL 1.0 inpainting model. Additionally, we integrated Google's Gemini 1.5 Flash model through the Google Generative AI API for enhanced text generation capabilities.

```
def initialize_models():
    # Florence model setup
    model_id = 'microsoft/Florence-2-large'
    model = AutoModelForCausalLM.from_pretrained(
        model_id,
        trust_remote_code=True,
        torch_dtype='auto'
    ).to('cuda:0').eval()
    processor = AutoProcessor.from_pretrained(model_id, trust_remote_code=True)

    # Setup for inpainting
    pipeline_text2image = AutoPipelineForInpainting.from_pretrained(
        "diffusers/stable-diffusion-xl-1.0-inpainting-0.1",
        torch_dtype=torch.float16,
        variant="fp16"
```

```
    ).to("cuda:0")
    pipeline_inpaint = AutoPipelineForInpainting.from_pipe(pipeline_text2image).to("cuda")

    genai.configure(api_key="key_token")
    gemini = genai.GenerativeModel('gemini-1.5-flash')

    return model, processor, pipeline_inpaint, gemini
```

*Prompt Processing*

The prompt is structured to classify user queries into three specific computer vision tasks: image captioning, object detection, and replacement. While the task prompts are <DETAILED_CAPTION>, <CAPTION_TO_PHRASE_GROUNDING>, and <REFERRING_EXPRESSION_SEGMENTATION>, respectively.

```
prompt = ["""
        you are about to solve a semantic question, and you need to classify the following question to a certain
task type
        \n
        there are three different task categories: image caption, object detection, and image segmentation.
        \n
        if you think the question is referred to image captioning, your response should be '<DETAILED_CAPTION>'
without quotation marks;
        \n
        and if you think the question is referred to object detection, your response should be
'<CAPTION_TO_PHRASE_GROUNDING>' without quotation marks;
        \n
        and if you think the question refers to image segmentation, your response should be
'<REFERRING_EXPRESSION_SEGMENTATION>' without quotation marks.
        \n
        for example, if the question is 'describe what the man is doing' or 'what color is the car', your response
should be '<DETAILED_CAPTION>' without quotation marks.
        \n
        and if the question is 'locate the man', your response should be '<CAPTION_TO_PHRASE_GROUNDING>'
without quotation marks.
        \n
        here is the question:
        """]
```

For replacement tasks (<REFERRING_EXPRESSION_SEGMENTATION>), it further analyzes the input to identify both the target object to be replaced and the replacement object.

```
def get_gemini_response(model, prompt, question):
    content = [prompt[0], question]
    task = model.generate_content(content)
    time.sleep(0.3)
    # print(task)
    if task.text == '<REFERRING_EXPRESSION_SEGMENTATION>\n':

        target = model.generate_content(['what is the object this user wants to inpaint and what is the object this
user wants to inpaint with. for example, if the user said "inpaint the white dog with an orange cat" or "replace
the white dog with an orange cat" or "change the white dog with an orange cat", you only need to respond "the
white dog, an orange cat":',
                                         question])

        target, object = target.text[:-1].split(',')
        target = target.strip()
        object = object.strip()
        # object = model.generate_content(['what is the object this user wants to inpaint with. for example, if the
user said "inpaint the white dog with an orange cat", you only need to respond an orange cat:',
        #                                 question])
        return task.text[:-1], target, object
    elif task.text == '<CAPTION_TO_PHRASE_GROUNDING>\n':
```

```
    return [task.text[:-1], question, None]
  else:
    return [task.text[:-1], None, None]
```