Vision-Language Model for Image Composition

1 Introduction

Image composition, which involves combining specific foreground objects with background images to create realistic composites (see Figure 1), has significant applications in image editing, creative industries, and enhancing downstream vision tasks like semantic segmentation and object detection. A key challenge is the difficulty of accurately learning object placement due to the weak supervision provided by bounding box labels. Recently developed large vision-language models (VLMs) have demonstrated unprecedented capabilities in generating high-quality text based on images and prompts, leveraging strong semantic priors learned from large collections of image-caption pairs.

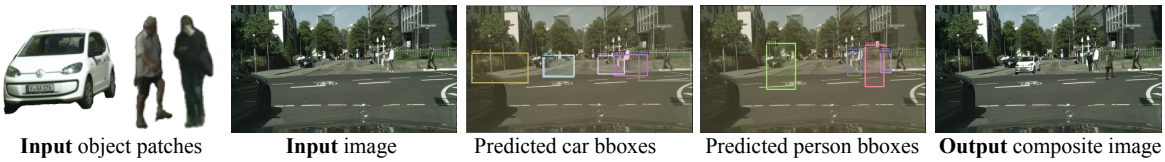

Figure 1: Visualization of image composition.

2 Task Options

2.1 Basic Task: Text-prompt VLM for Visual Grounding

Goal: Predict a bounding box for region-of-interest localization given a text and an image, see Figure 2.

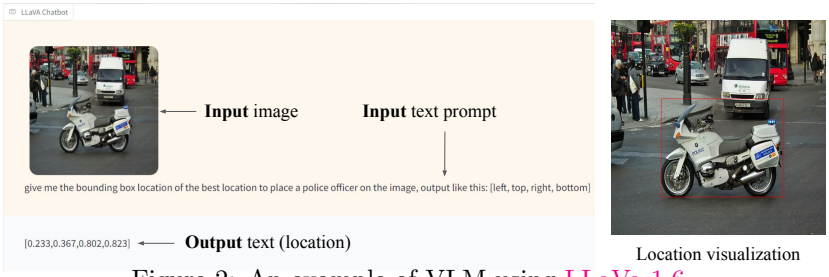

Figure 2: An example of VLM using LLaVa-1.6.

Input: An image and a text prompt, e.g., `Give me the bounding box of the best location to place a police officer on the image, output like this: [left, top, right, bottom]`
Output: A bounding box
Dataset suggestion: Cityscapes and LVIS
References: Visual instruction tuning, NeurIPS 2023
Improved baselines with visual instruction tuning, CVPR 2024
Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond, arXiv

2.2 Advanced Task: Patch-prompt VLM for Object Composition

Goal: Generate a 2D bounding box for object composition given an object patch and an image.
Input: An object patch and a scene image
Output: A composite image
Dataset suggestion: Same as Section 2.1
References:
TopNet: Transformer-based object placement network for image compositing, CVPR 2023
BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, ICML, 2023
Groma: Localized visual tokenization for grounding multimodal large language models, arXiv