Downstream Applications of Open-Vocabulary Visual Perception: Scene Understanding

Xinyu Chen, chenxy835@mail2.sysu.edu.cn

Abstract:

Advancements in deep learning and computer vision have revolutionized tasks like image recognition and object detection. Yet, traditional models are limited by fixed vocabularies, hindering their adaptability to new objects and scenes. Open-vocabulary visual perception overcomes this hurdle by enabling systems to recognize an expansive array of novel concepts. This flexibility is pivotal for scene understanding, crucial in autonomous driving, surveillance, and augmented reality. This review focuses on open-vocabulary approaches, highlighting advancements in scene graph generation, 3D instance segmentation, and vision-language navigation.

Introduction:

Open-Vocabulary Visual Perception enhances computer vision by recognizing new objects beyond fixed vocabularies. It integrates Vision-Language Models (VLMs) for semantic understanding and employs zero-shot learning to recognize unseen categories.

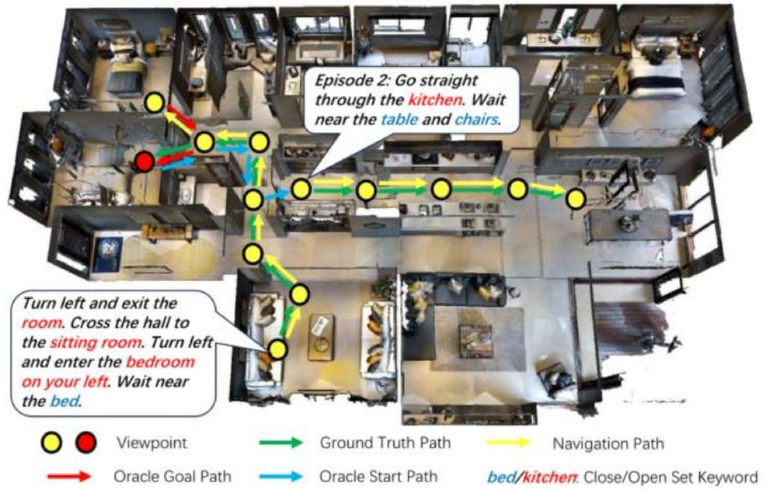
Scene understanding extracts meaning from images or videos, including object detection, relationships, and scene structure.

Scene understanding tasks: Object detection, scene classification, scene graph generation (SGG), semantic and instance segmentation.

Review OVVP techniques in scene understanding:

- From Pixels to Graphs : Open-vocabulary scene graph generation.
- OpenMask3D : Open-vocabulary 3D instance segmentation.
- OVER-NAV : Iterative vision-and-language navigation with open-vocabulary detection.

"footrest" *"side table with a flower vase on it"*



Results:

- From Pixels to Graphs: Uses generative Vision-Language Models (VLMs) for open-vocabulary Scene Graph Generation (SGG), improving efficiency and performance in vision-language tasks.
- OpenMask3D: Introduces zero-shot 3D instance segmentation, expanding object category recognition and enhancing robot interaction in new environments.
- OVER-NAV: Innovates Iterative Vision-and-Language Navigation (IVLN) with Large Language Models (LLMs) and Open-Vocabulary Detection (OVD), improving navigation in complex settings.