Context Mask Priors via Vision-Language Model for Ergodic Search

Derek Ming Siang Tan[1], Ananya Rao[2], Abigail Breitfeld[2], Guillaume Sartoretti[1]

[1]Dept. of Mechanical Eng., National University of Singapore
[2]Robotics Institute., Carnegie Mellon University

ICRA2024
YOKOHAMA | JAPAN

Introduction

Motivation:

- Conventional segmentation techniques requires hand-crafted feature extraction rules or training for the specified target classes.
- More recent methods achieves segmentation of objects in an image described by a natural language prompt (***Referring Segmentation***).

Challenges:

- *Referring Segmentation* task often work only with simple and direct prompts.
- Conventional approaches often require **crafting of new rules or retraining** when introducing new target classes.

Proposal:

- Use of Vision-Language Model (VLM) to achieve ***Reasoning Segmentation***.
- Ability to reason complex prompt & vision inputs using world knowledge.
- Output segmentation mask can be used as **information distribution provided to downstream tasks *a priori*** (e.g., exploration, search, manipulation, information gathering for learning, surveillance, surface cleaning etc.).
- Demonstrate efficacy on the downstream task of **Ergodic Search.**

Problem Formulation

Context Mask Generation:

- Choice of VLM: **L**anguage **I**nstructed **S**egmentation **A**ssistant (LISA)[1]
- Input: Image + Task prompt, Output: Segmentation Mask + Explanation

$$(y_{\text{img}}, y_{\text{txt}}) = F(x_{\text{img}}, x_{\text{txt}})$$

Ergodic Search:

Optimization problem to obtain trajectory sets where agents spend time in each area of the domain proportional to the expected amount of information present in this area [2].

$$C^t(\mathbf{x}, \gamma_t) = \frac{1}{t}\sum_{\tau=0}^{t-1} \delta(\mathbf{x} - \gamma_i(\tau)),$$

$$\Phi(\gamma(t)) = \sum_{k=0}^{m} \lambda_k \left|c_k(\gamma(t)) - \xi_k\right|^2,$$

$$\begin{aligned} \boldsymbol{u}^*(t) &= \arg\min_{\boldsymbol{u}} \Phi(\gamma(t)), \\ \text{subject to } \dot{\boldsymbol{q}} &= f(\boldsymbol{q}(t), \boldsymbol{u}(t)), \\ \|\boldsymbol{u}(t)\| &\le u_{max} \end{aligned}$$

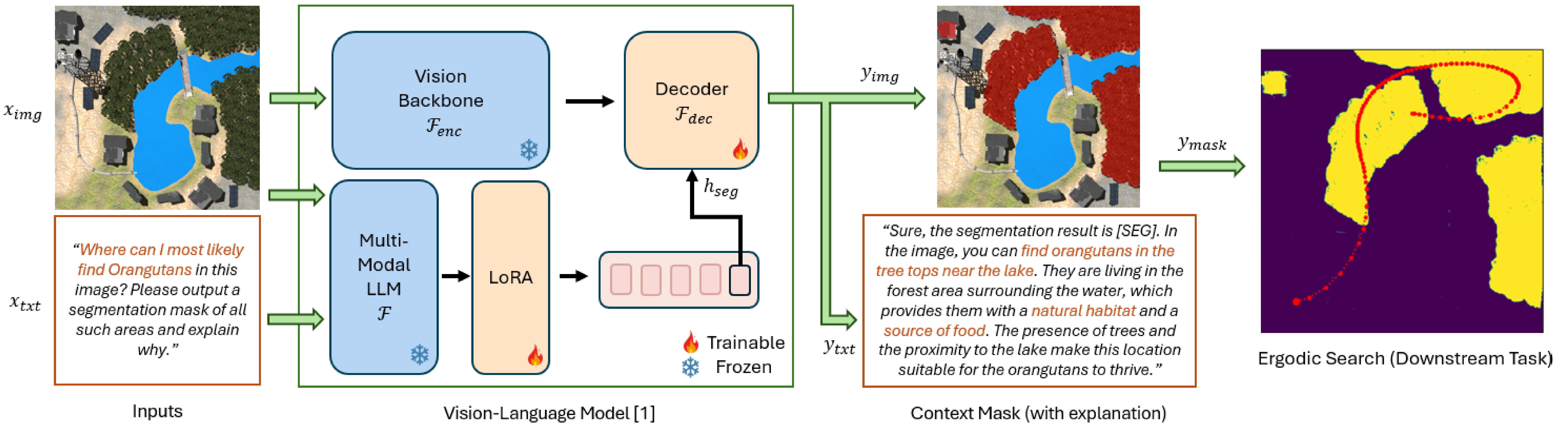
Context Mask Prior for Ergodic Search

Context Mask Generation via VLM (Upstream Task):

- Frozen: *LlaVa* as Multi-Modal LLM, *SAM* as Vision Backbone (Pretrained).
- Trainable: LoRA and Decoder.
- Training set: *ReasonSeg* – over one thousand image–instruction (complex) pairs.
- Extracts and reconstructs last-layer segmentation embedding into binary mask.

Ergodic Search (Downstream Task):

- Uses binary context mask as information map prior to drive search trajectory.
- Goal is to minimize the ergodic metric.
- Flexibility to adapt to other downstream tasks.


x_{img}
x_{txt}
Vision Backbone $\mathcal{F}_{enc}$
Multi-Modal LLM $\mathcal{F}$
LoRA
Decoder $\mathcal{F}_{dec}$
h_{seg}
🔥 Trainable
❄ Frozen
y_{img}
y_{txt}
y_{mask}

"Where can I most likely find Orangutans in this image? Please output a segmentation mask of all such areas and explain why."

Inputs

Vision-Language Model [1]

"Sure, the segmentation result is [SEG]. In the image, you can find orangutans in the tree tops near the lake. They are living in the forest area surrounding the water, which provides them with a natural habitat and a source of food. The presence of trees and the proximity to the lake make this location suitable for the orangutans to thrive."

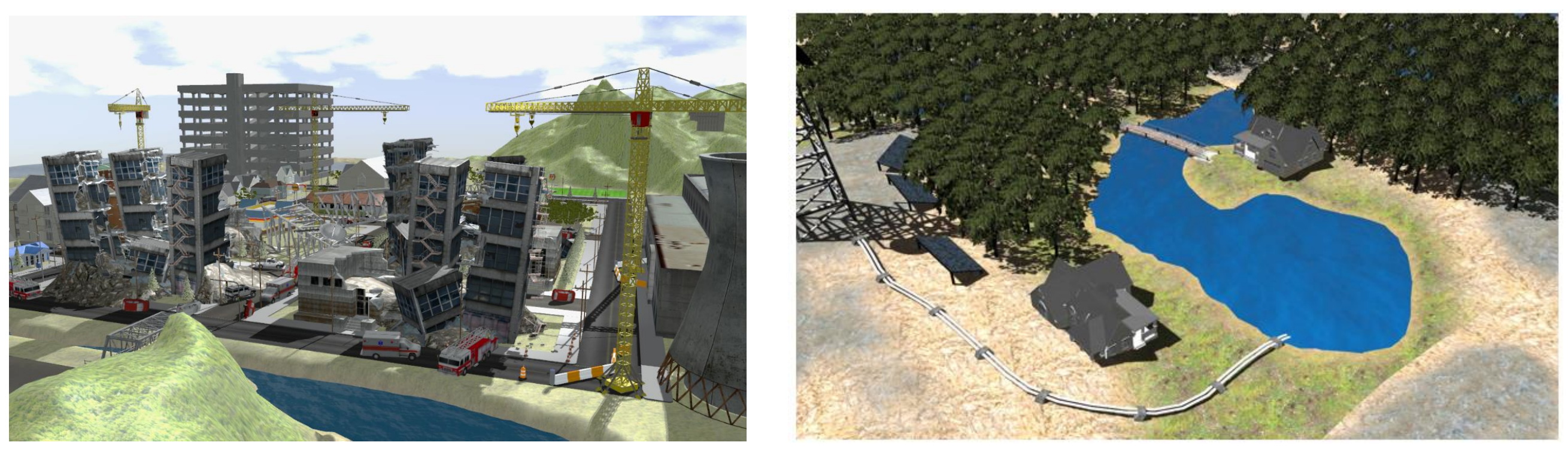
Context Mask (with explanation)

Ergodic Search (Downstream Task)

Experiments

Experimental Setup:

- Generated two environments in Gazebo: ***Earthquake***, ***Forest***.
- Three different configurations of environmental features per environment.
- Unmanned Aerial Vehicle (UAV) with gaussian sensor model.
- Context masks are available to UAV *a priori*.

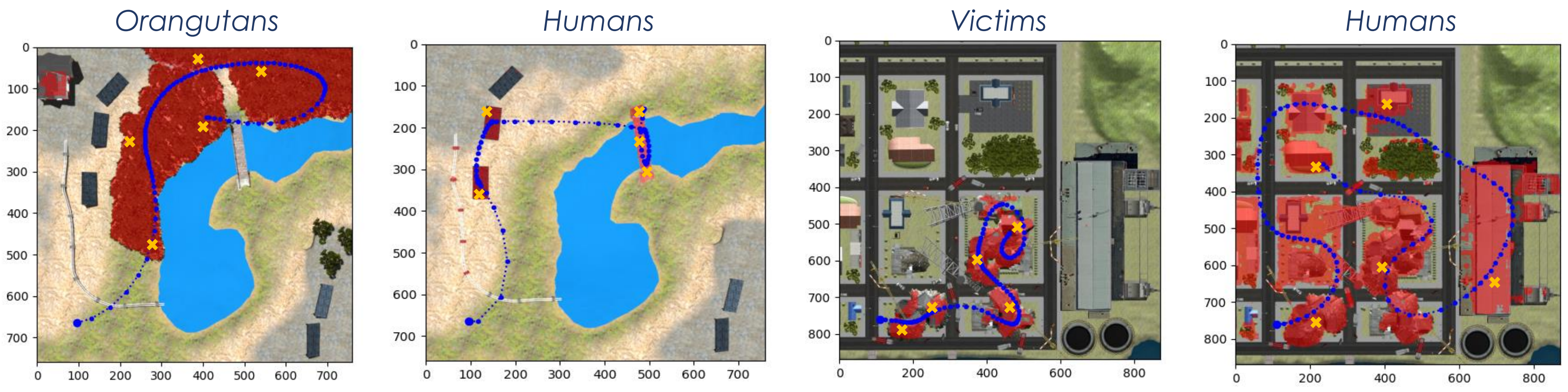


Baselines:

1. **Ergodic Search:** Trajectories balance exploration and exploitation by minimizing the ergodic metric.
2. **Lawnmower:** Trajectories uniformly cover the search region.
3. **Greedy Search:** Trajectories exploit areas of high information by picking highest information gain at each step.

Results and Discussion

Examples (Ergodic Search):



Results:

- Averaged over 50 experiments for each method.
- Varied environment configuration, robot start positions, and target positions.

Metric	Lawnmower	Greedy Search	Ergodic Search
Ergodic Metric	0.3482	0.0473	0.0023
Time Taken to Find All Targets (sec)	55.4	43.7	26.1

Future Works

1. **Extension of Validation Set:** Validation using realistic satellite maps across the world, for a larger variety of target classes.
2. **Generate Score Mask:** Fine-tune VLM to output non-binary score mask instead of binary segmentation mask.
3. **Dynamic mask refinement:** Introduce feedback loop where targets found during the search process are used to enhance prompt input into VLM.
4. **Edge Inference:** Distillation of VLM onto compute-limited embedded devices for online segmentation mask generation on real robots.

Contact Information

Derek Ming Siang Tan, Ananya Rao, Abigail Breitfeld, Guillaume Adrien Sartoretti

Email: derektan@u.nus.edu, ananyara@andrew.cmu.edu, abreitfe@andrew.cmu.edu, guillaume.sartoretti@nus.edu.sg

References

[1] X. Lai et al., "LISA: Reasoning Segmentation via Large Language Model." arXiv, May 01, 2024.
[2] A. Rao, I. Abraham, G. Sartoretti, and H. Choset, "Sparse Sensing in Ergodic Optimization," in Distributed Autonomous Robotic Systems (DARS).

NUS National University of Singapore
Carnegie Mellon University Robotics Institute