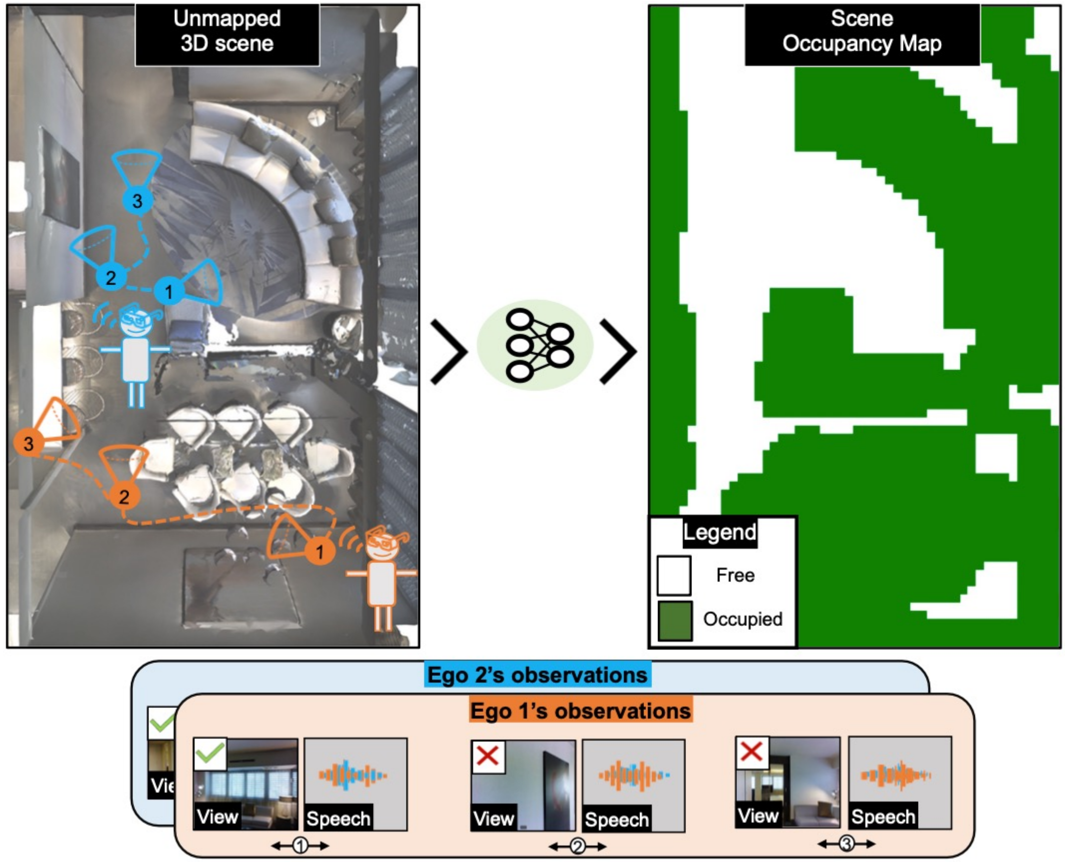
Chat2Map: Efficient Scene Mapping from Multi-Ego Conversations

Sagnik Majumder[1,2,3] Hao Jiang[2] Pierre Moulon[2] Ethan Henderson[1,2]
Paul Calamia Kristen Grauman[1,3]* Vamsi Krishna Ithapu[2]*

[1]UT Austin [2]Reality Labs Research, Meta [3]FAIR

Project page: https://vision.cs.utexas.edu/projects/chat2map/

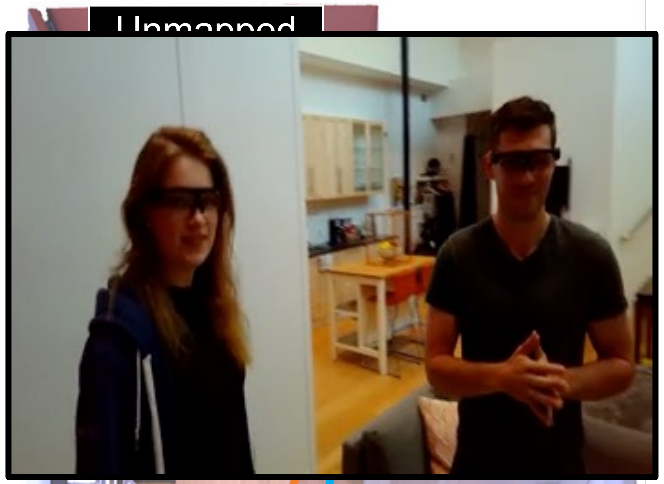
CVPR 2023 paper tag: WED-AM-222



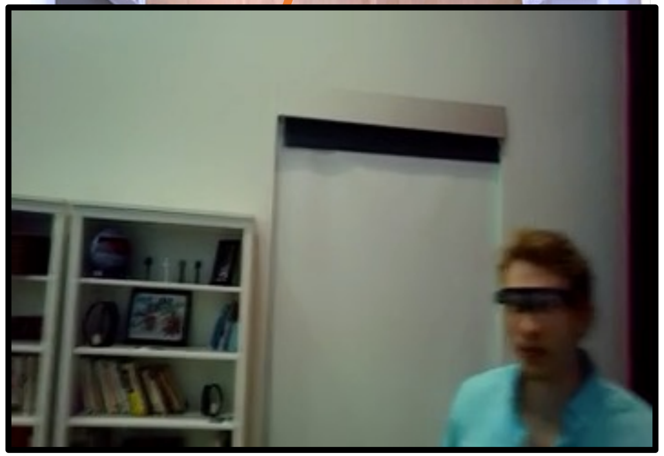
*Equal contribution

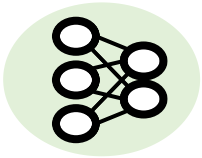
Chat2Map: Efficient Scene Mapping from Multi-Ego Conversations

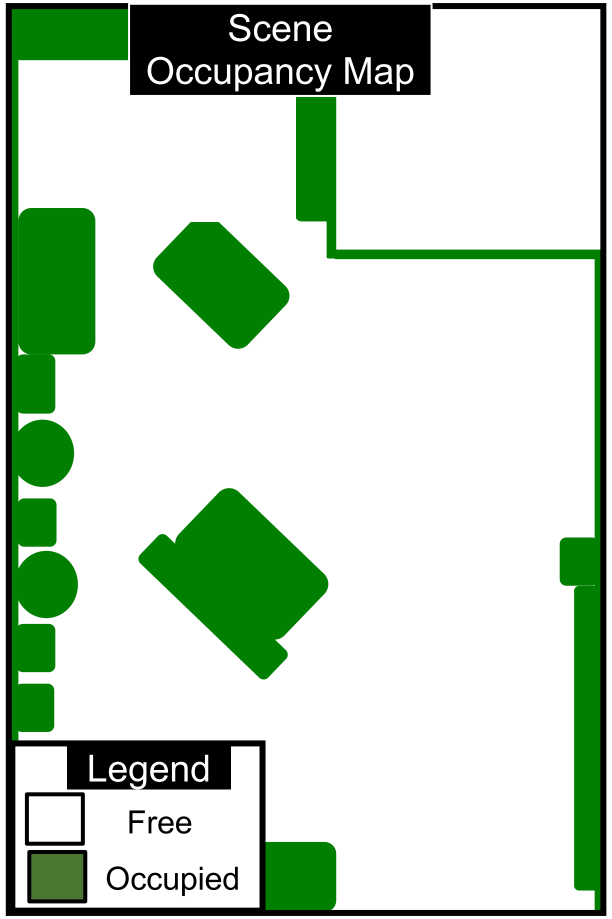
AR device wearer 1



AR device wearer 2



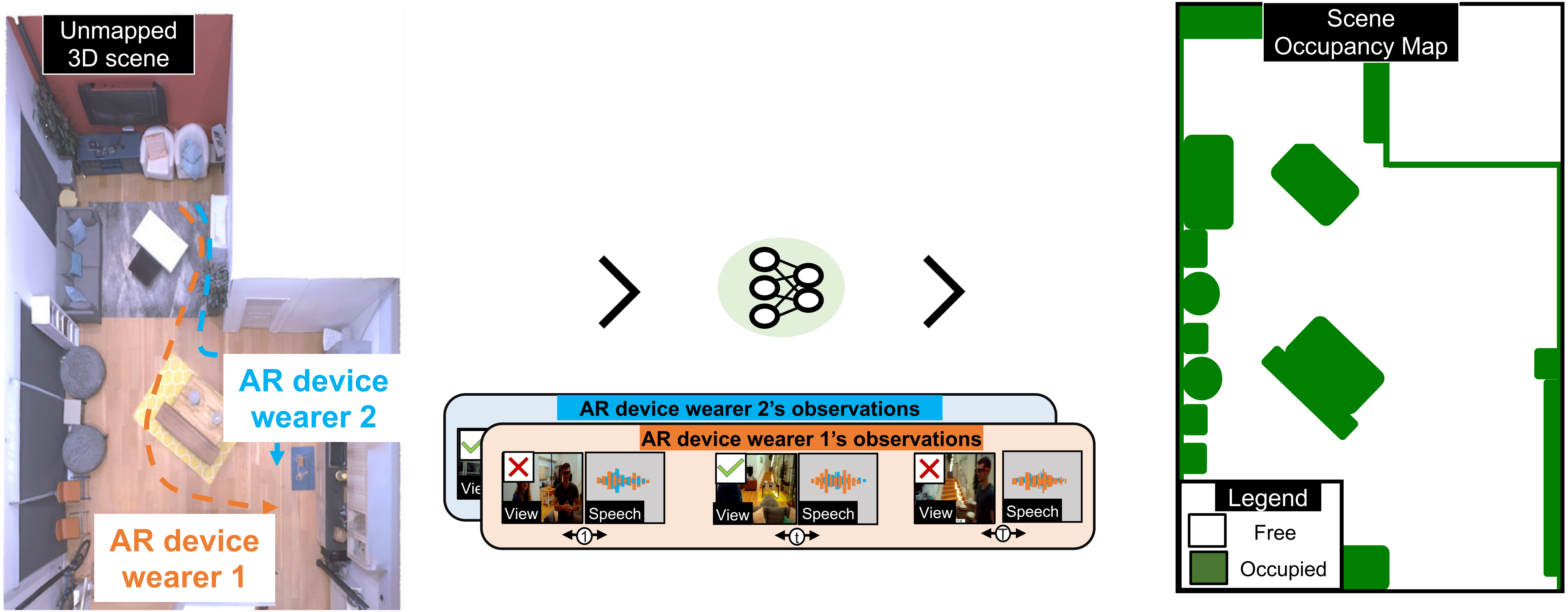




Given an **unmapped 3D scene** with two people wearing AR glasses moving about in it and conversing**,** we propose a novel task of **efficiently predicting** the **topdown occupancy map** of the scene using the **egocentric audio-visual streams** from the conversation such that **cost** of **visual capture doesn't exceed** a **fixed budget**, where the **budget** is a **very small percentage** of **all possibly sampled frames**

Data source: 1) **EgoCom.** *Northcutt et al., 2020.*, 2) **Replica.** *Straub et al., 2019.*

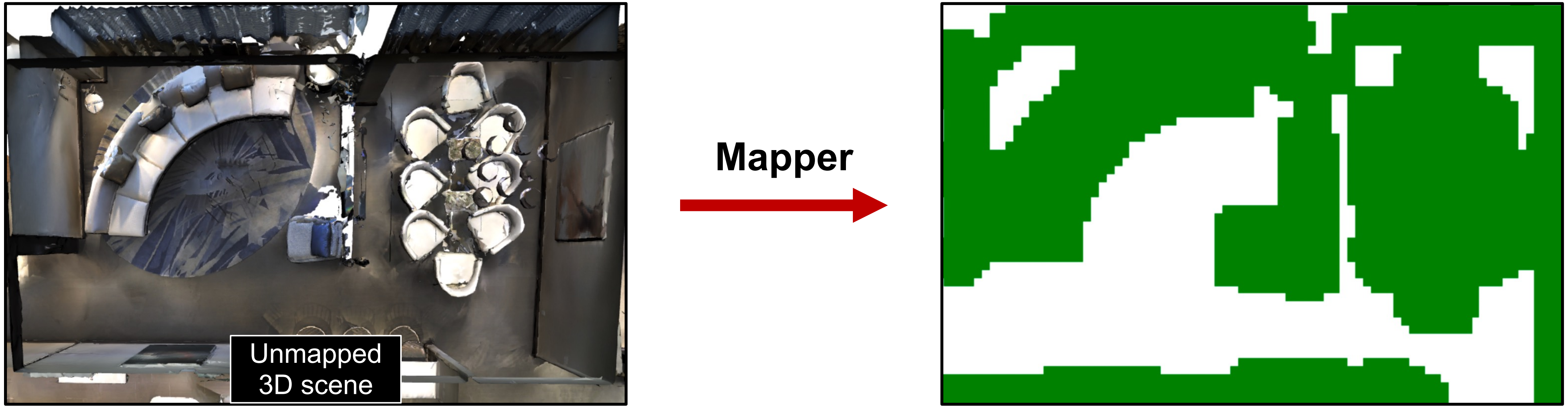
Chat2Map: Efficient Scene Mapping from Multi-Ego Conversations



Towards that goal, we learn a model that uses an **RL policy** to **decide** for each ego and at each step whether to **sample** the **current visual frame** or **skip** it, given the audio stream and sampled visual frames from the past, and a **transformer-based model** to **efficiently infer** the **scene map** using the audio and sampled frames

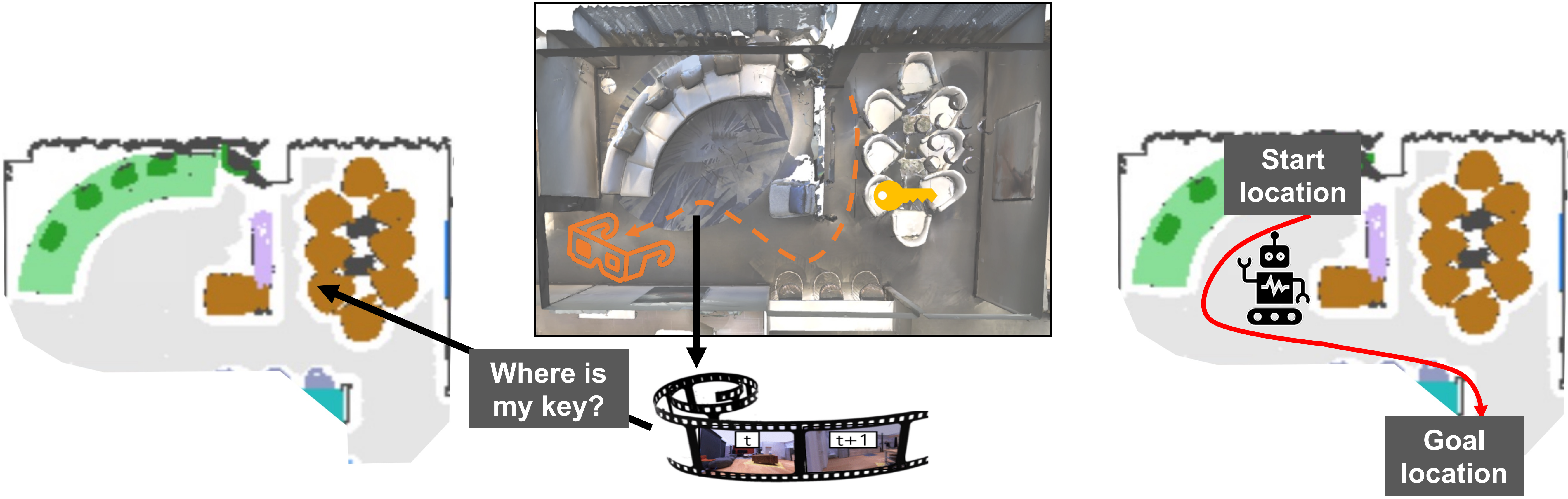
Data source: 1) **EgoCom.** *Northcutt et al., 2020.*, 2) **Replica.** *Straub et al., 2019.*

Topdown scene maps



Topdown scene maps describe **how objects** and **structures** of various **shapes** and **sizes** (geometry), and **types** (semantics) are **distributed** in a 3D scene

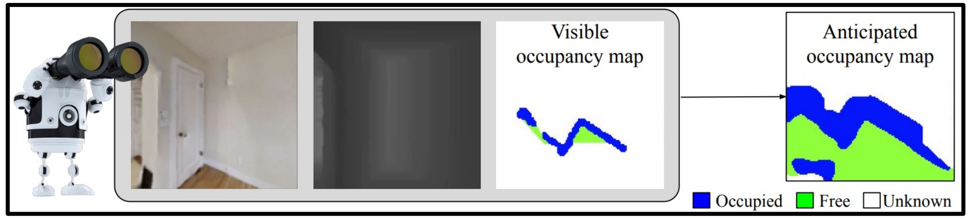
Topdown scene maps



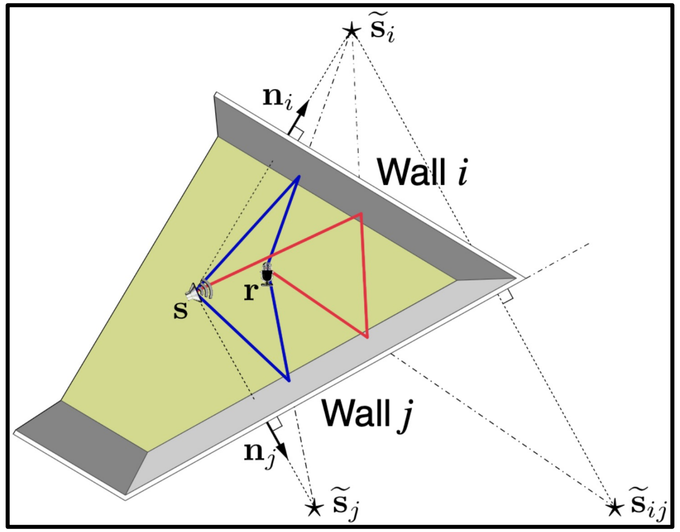
Scene layout estimation | **Episodic memory** | **Embodied navigation**

Topdown scene maps help with **scene understanding** and are important for both **AR/VR** (e.g., estimating scene layouts, episodic memory, etc.) and **robotics applications** (e.g., planning in embodied navigation)

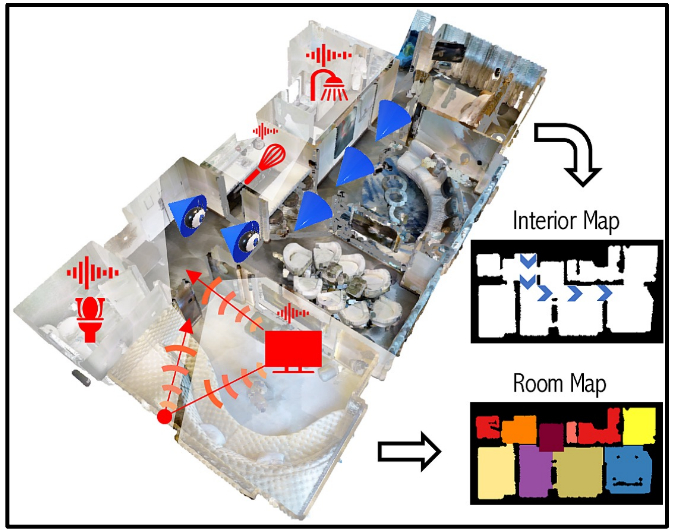
State of the art in topdown scene mapping



Ramakrishnan et al., ECCV '20



Dokmanic et al., PNAS '13



Purushwalkam et al., ICCV '21

Doesn't use audio → can't predict global geometry accurately

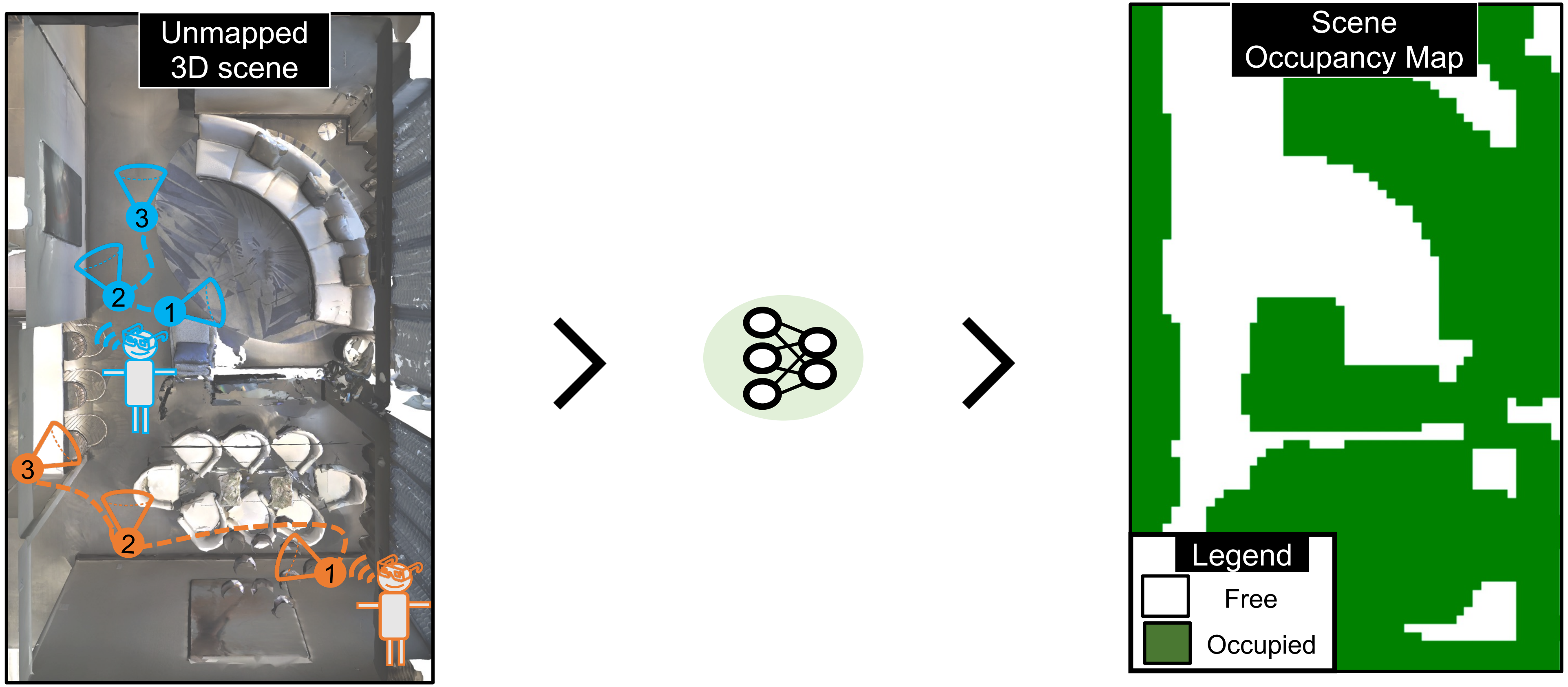
Limited to polyhedral spaces

Emits intrusive chirp signals; not sample efficient

a) **Vision-only**

b) **Audio-visual**

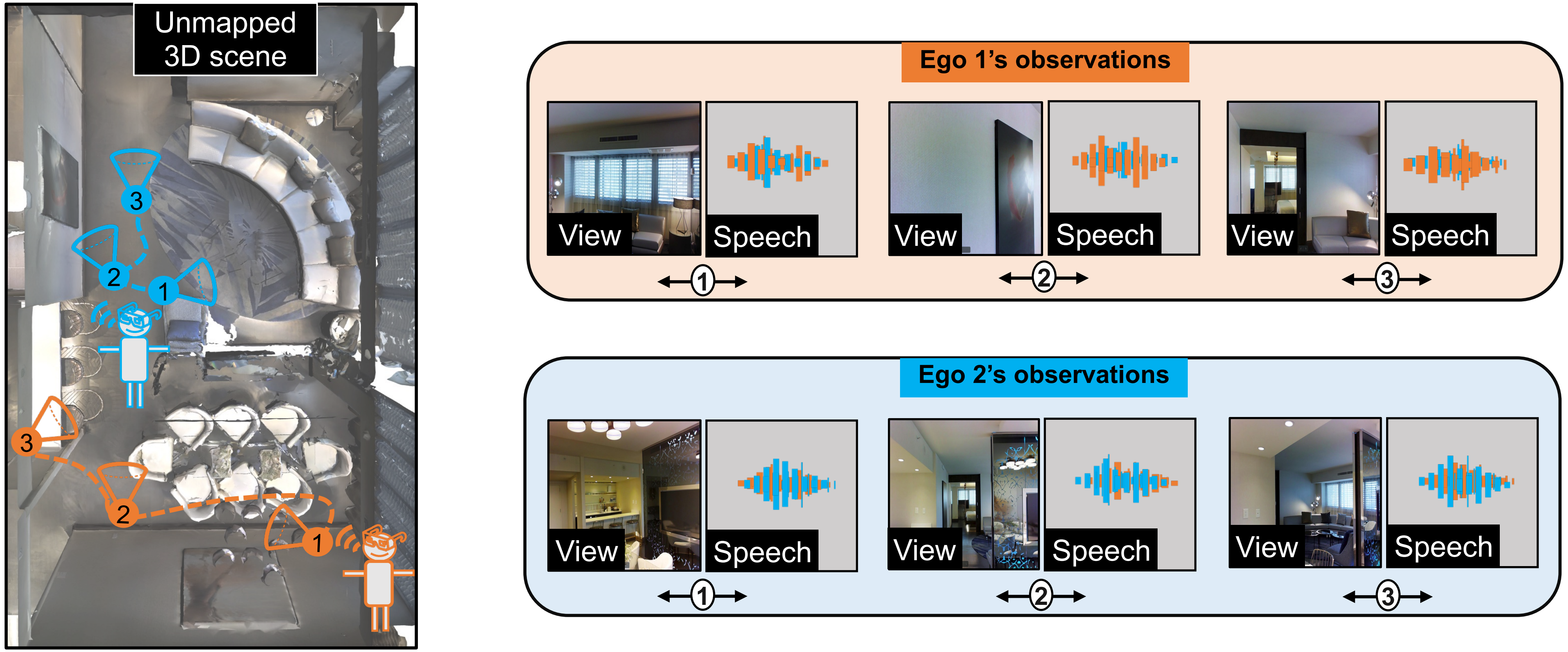
Task Description: Efficient Scene Mapping from Multi-Ego Conversations



Given an **unmapped 3D scene** where two people wearing AR glasses (egos) are moving about and conversing, the goal in this task is to **efficiently use** the **egocentric audio-visual streams** from the conversation in **real time** and **infer** the **topdown occupancy map** of the scene, such that **cost** of **visual capture doesn't exceed** a certain pre-specified **budget**.

Note: our **visual budget** is a **very small percentage** of the **total number** of **frames possibly sampled** by the egos in the episode, enabling large savings in the visual capture cost

Task Description: Efficient Scene Mapping from Multi-Ego Conversations

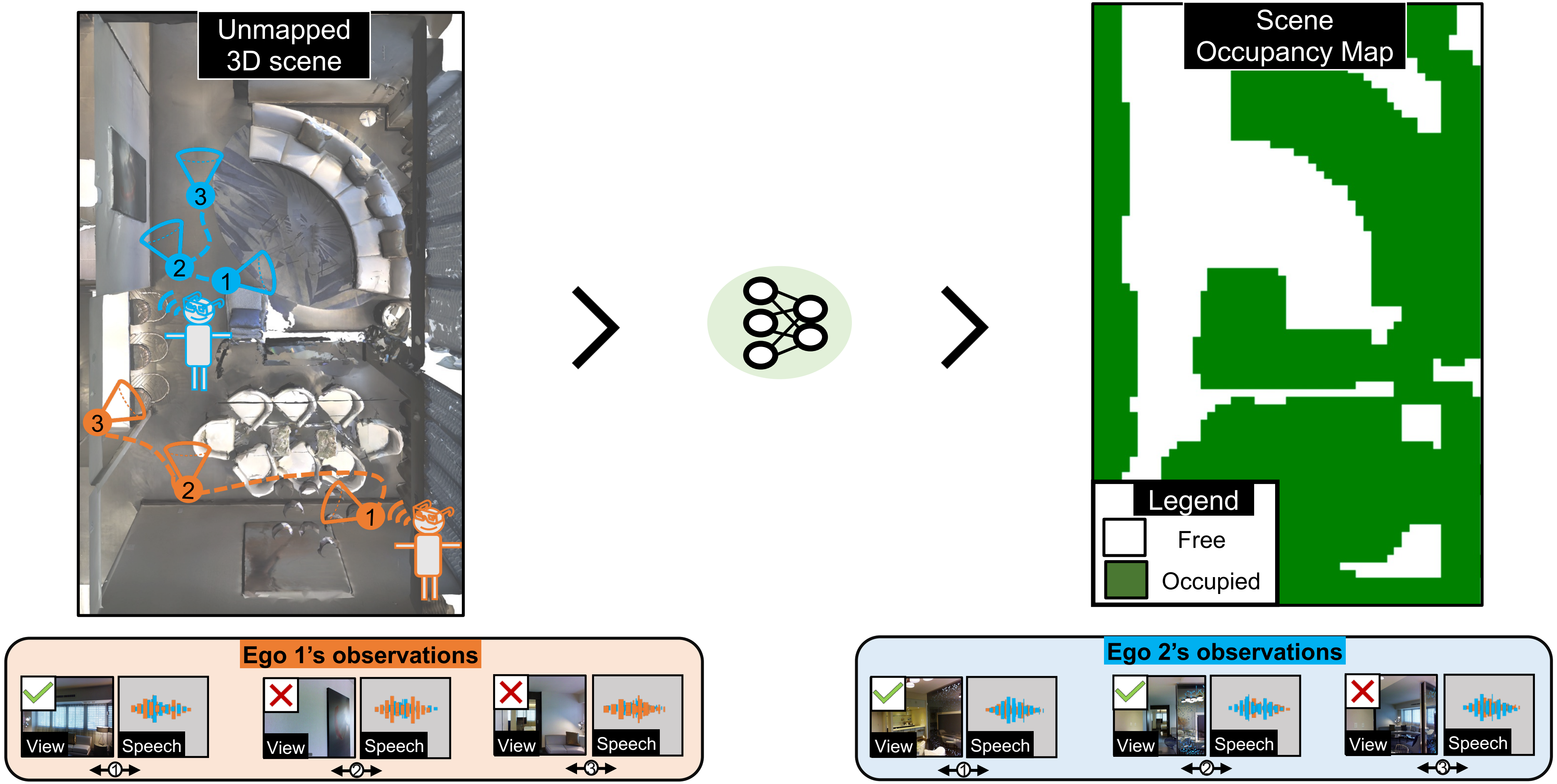


Each **egocentric audio-visual** observation contains

- the **90° field-of-view image** of the scene
- **multi-channel audio** comprising speech from **self** and the **other ego** *

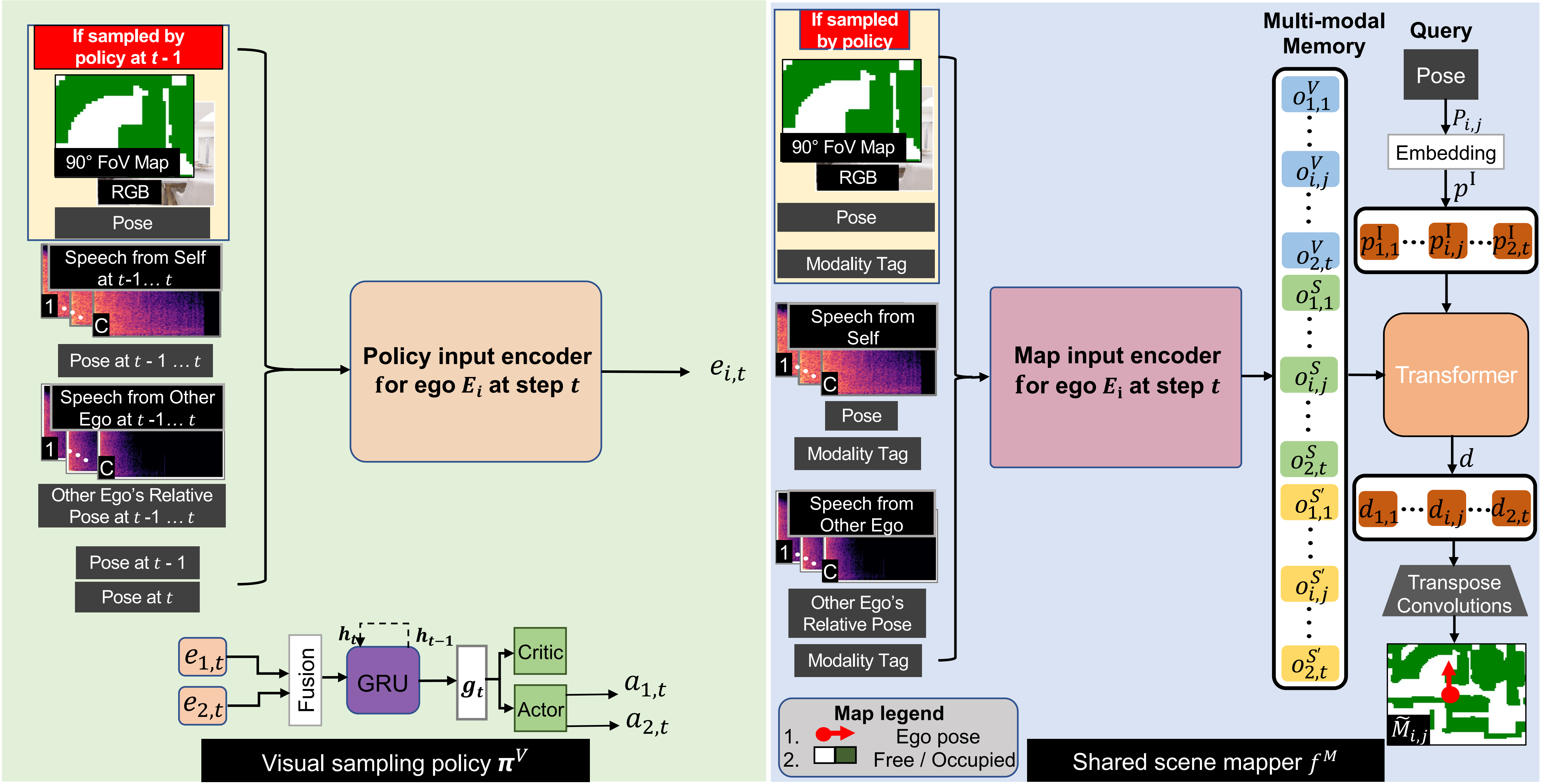
* The audio **could be mixed** with **ambient environment sounds**

Task Description: Efficient Scene Mapping from Multi-Ego Conversations



At each step of the conversation episode, we aim to learn a model that **decides** for each ego whether to **sample** the **current visual frame** or **skip** it, given the audio stream and sampled visual frames from the past , and **predicts** the **scene map** in a **cost-effective** manner

Our approach



Novel policy reward: $r(t) = \Delta Q(t) - \eta \, * \, \rho(t)$

Mapper loss: $\mathcal{L}^M(t) = \dfrac{1}{2 \times t} \sum_{i=1\dots2} \sum_{j=1\dots t} \mathrm{BCE}(\tilde{M}_{i,j}, M_{i,j})$

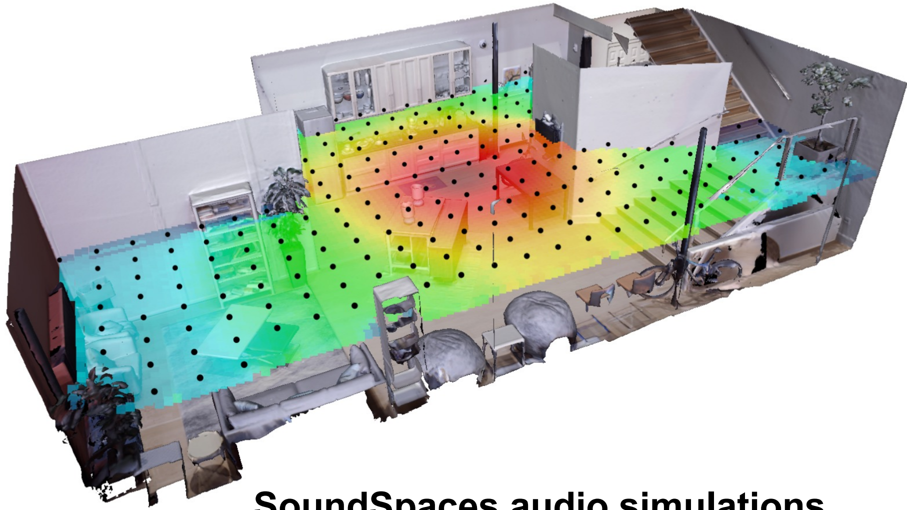
Experimental setup

Evaluation settings:

- **Passive mapping:** the mapper has access to all visual samples
- **Active mapping:** the model actively choses visual frames given a fixed budget

3D scenes and spatial audio:

- **State-of-the-art audio-visual simulator:**
  - Matterport3D [1] scenes
  - SoundSpaces [2] acoustics simulator
- **Real-world data:**
  - Mock-up apartment for both visual and impulse response data capturing



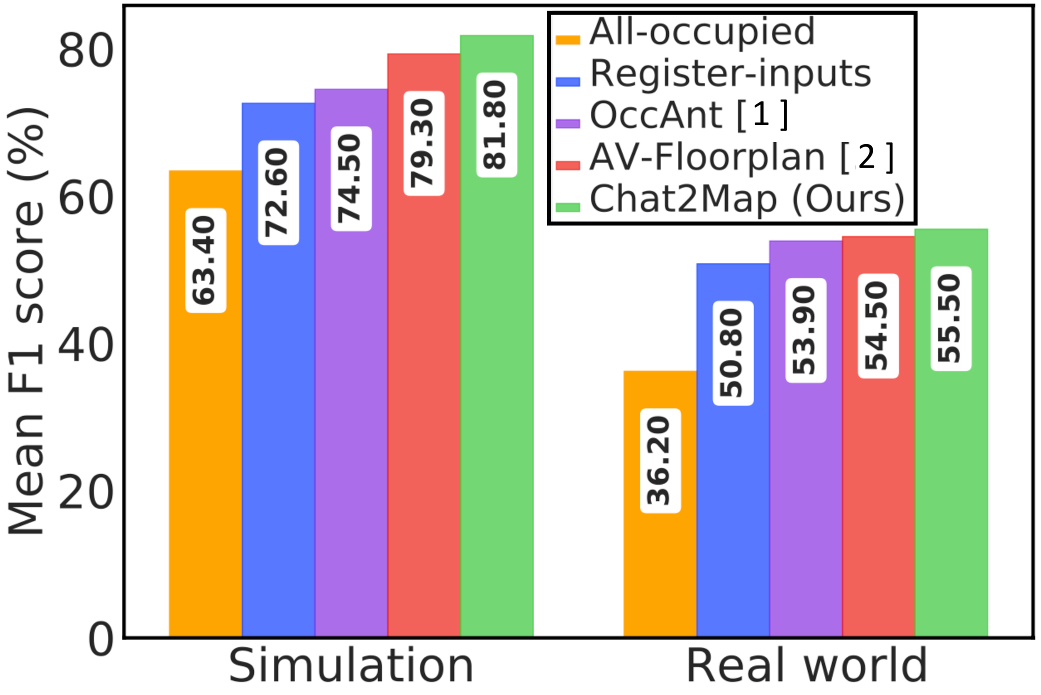
SoundSpaces audio simulations

Evaluation metrics:

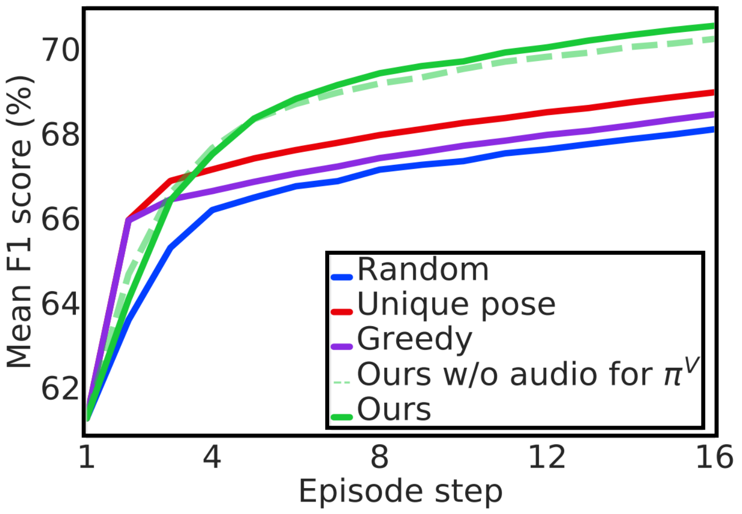
- Mean F1 score
- Mean Intersection over Union (IoU)

[1] Matterport3D. *Chang et al., 3DV 2017.*
[2] SoundSpaces. *Chen et al., ECCV 2020.*

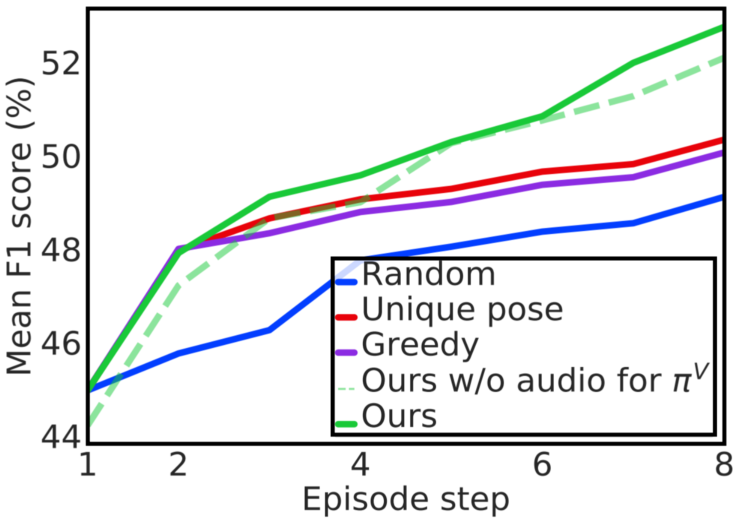
Results



Passive **mapping performance**



Simulation



Real world

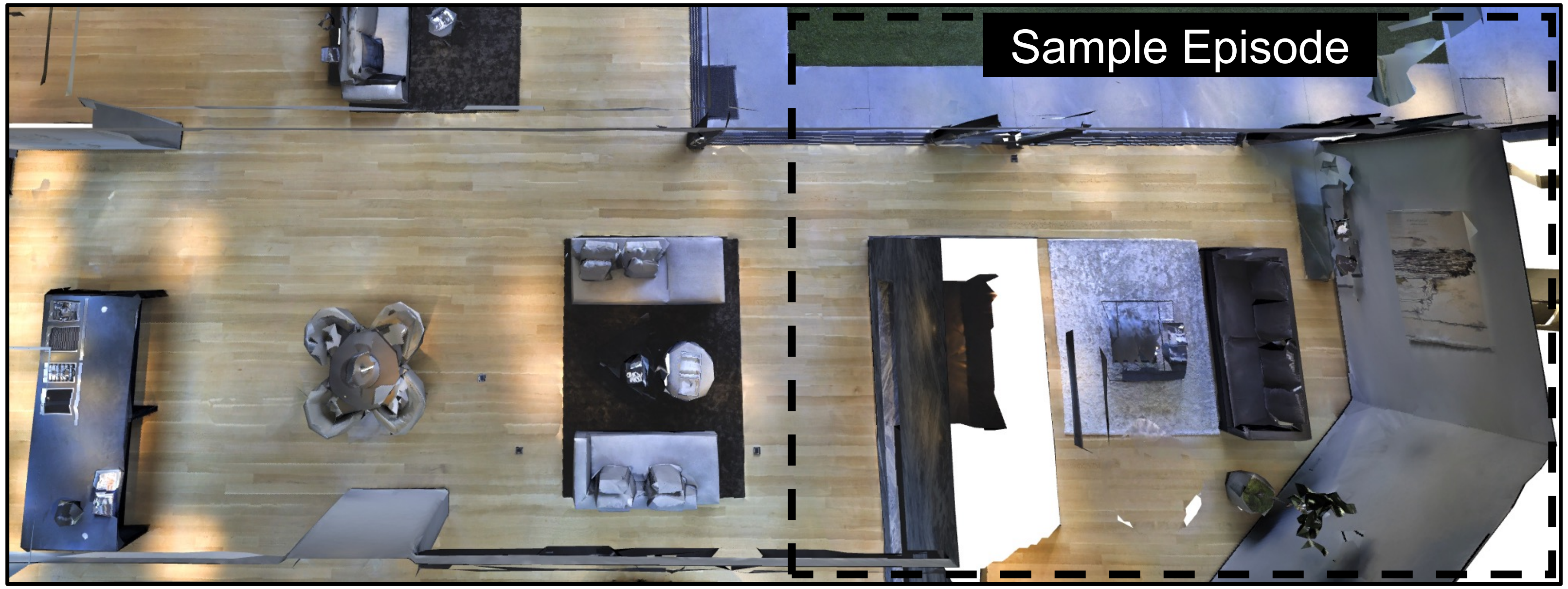
Active **mapping performance**

Our model outperforms all baselines on both **simulated** and **real-world** data for both **passive** and **active mapping**

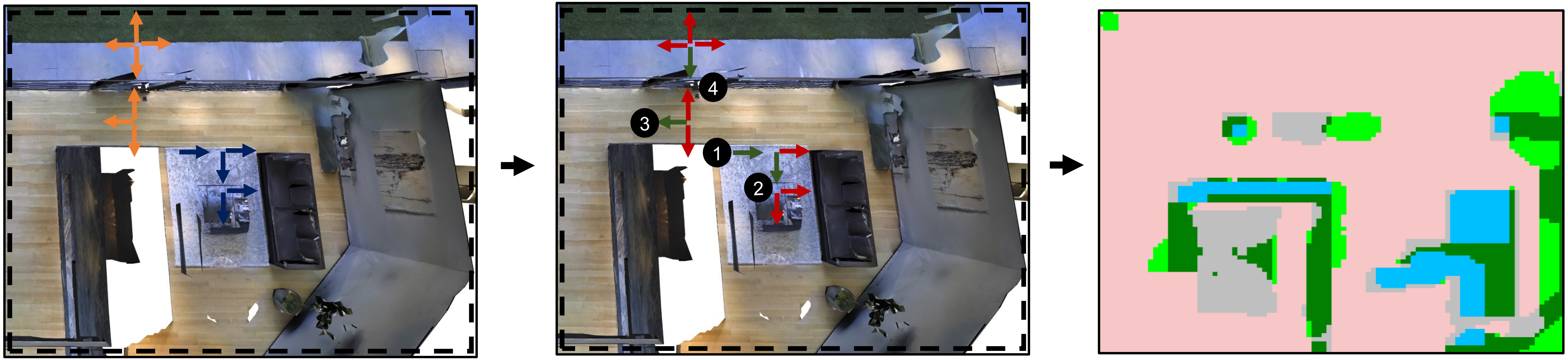
Compared to passive mapping, our active mapper *saves* as much as **74 Watts** with a budget of 2 frames while the **performance declines marginally**

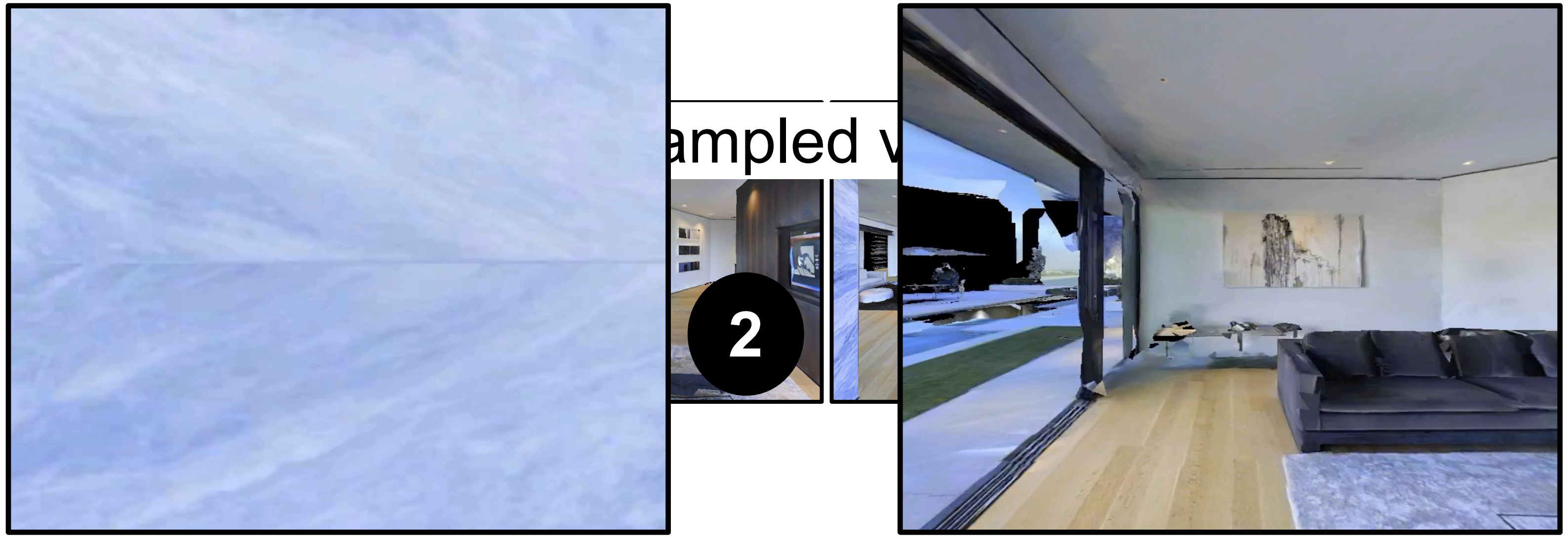
[1] **Occupancy Anticipation.** *Ramakrishnan et al., 2020.*
[2] **Audio-Visual Floorplans.** *Purushwalkam et al., 2021.*

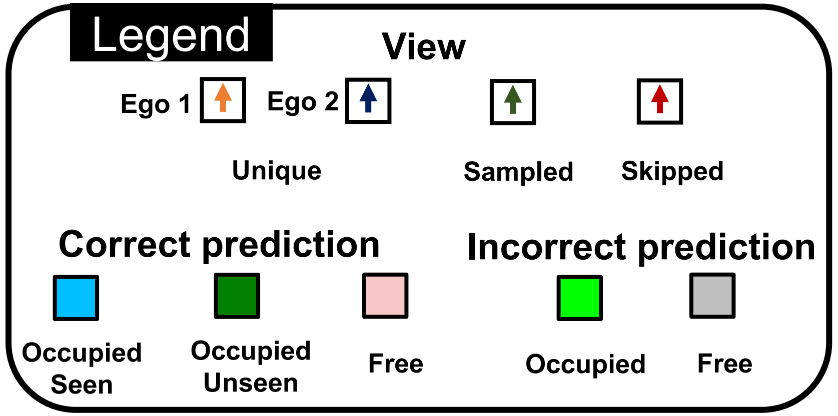
Sample map predictions in a **simulated *unseen*** environment



Sample map predictions in a **simulated *unseen*** environment



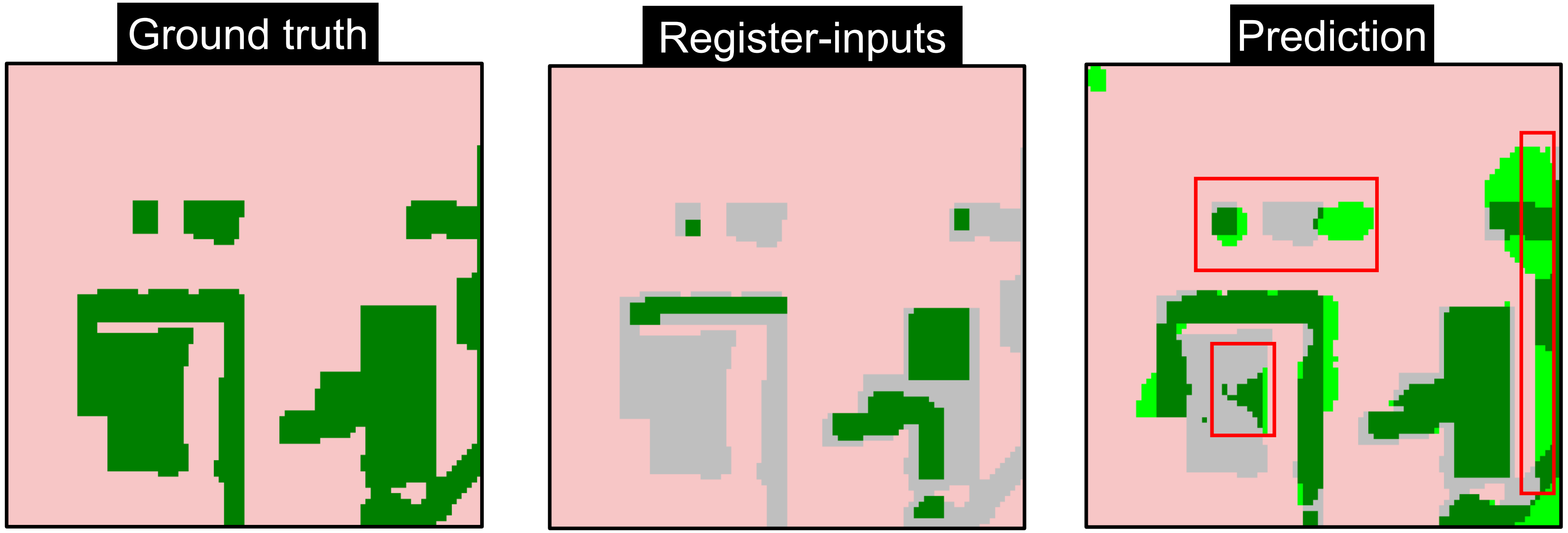


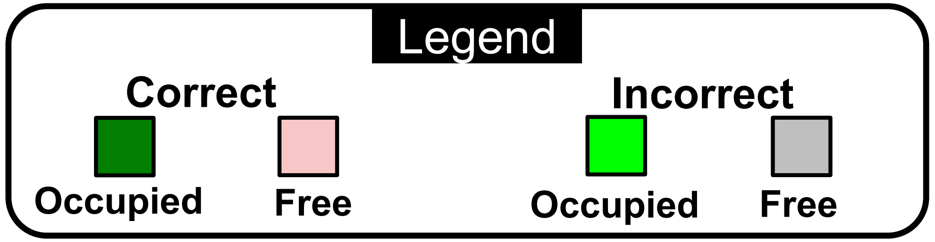


Ego 1 trajectory with self-speech

Ego 2 trajectory with self-speech

Sample map predictions in a **simulated *unseen*** environment





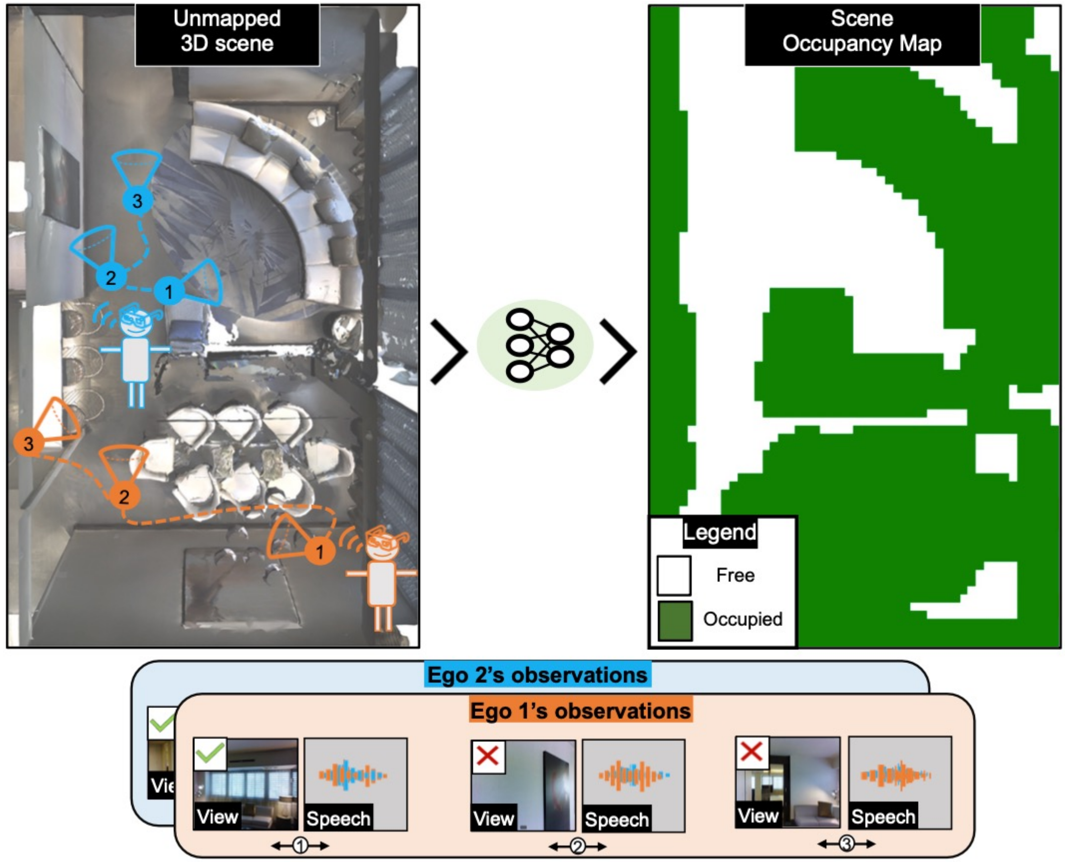
Chat2Map: Efficient Scene Mapping from Multi-Ego Conversations

Sagnik Majumder[1,2,3] Hao Jiang[2] Pierre Moulon[2] Ethan Henderson[1,2]
Paul Calamia Kristen Grauman[1,3]* Vamsi Krishna Ithapu[2]*

[1]UT Austin [2]Reality Labs Research, Meta [3]FAIR

Project page: https://vision.cs.utexas.edu/projects/chat2map/

CVPR 2023 paper tag: WED-AM-222



*Equal contribution