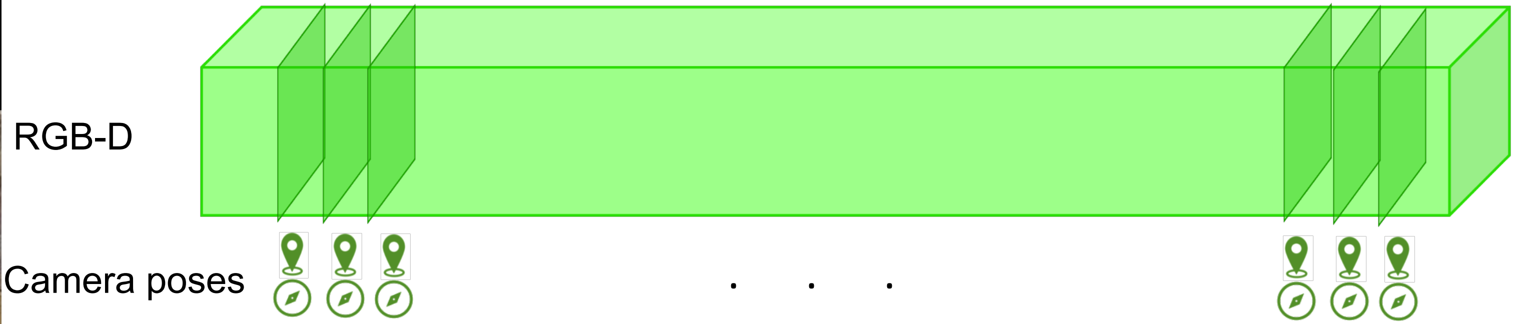
Environment Predictive Coding for Visual Navigation

Anonymous ICLR 2022 submission

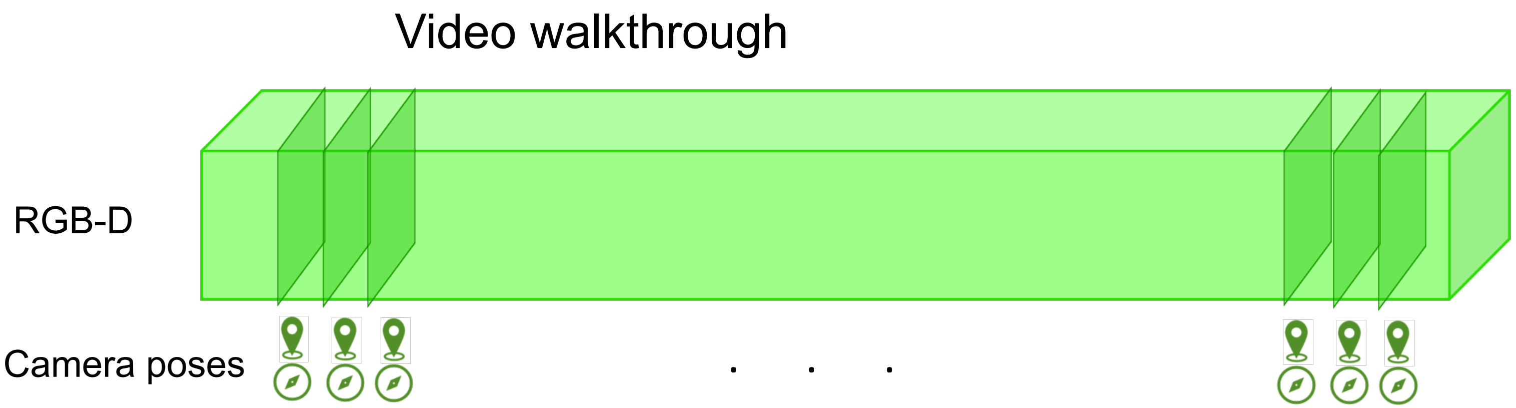
Environment Predictive Coding (EPC)

Video walkthrough



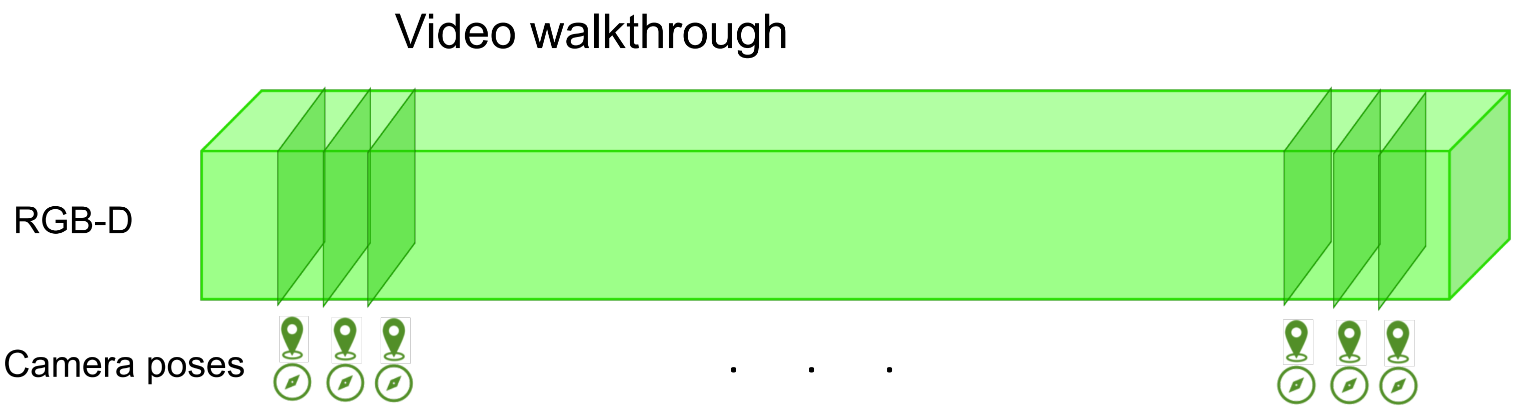
We are given *video walkthroughs* collected by another agent navigating in various indoor environments.

Environment Predictive Coding (EPC)



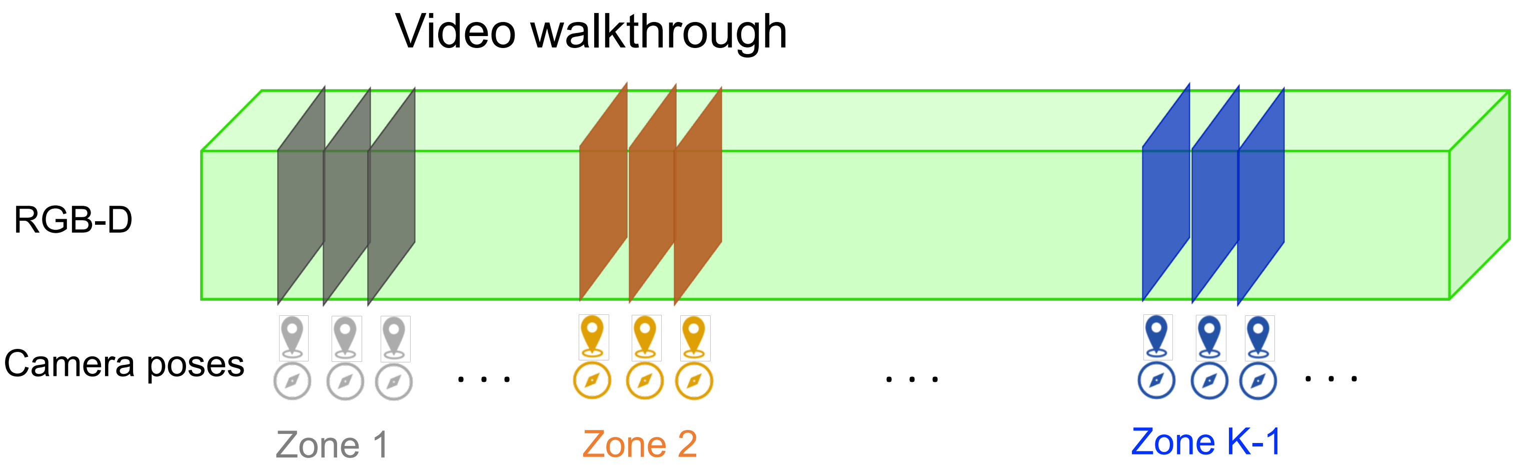
The *observed portions* of the environment are shown in red.

Environment Predictive Coding (EPC)

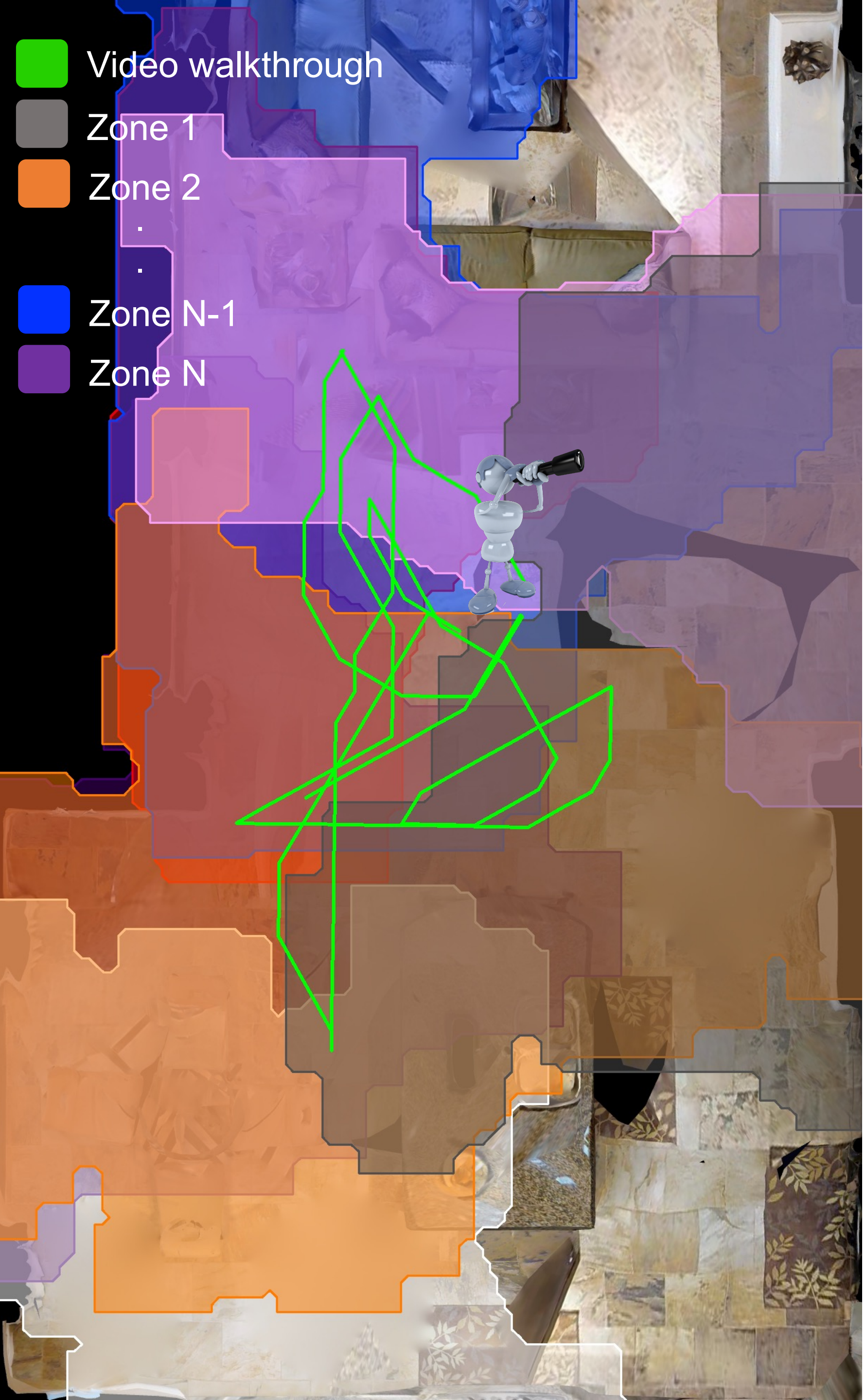


- We propose the *masked-zone prediction* task for self-supervised learning.
- The goal is to learn environment-level representations of *egocentric observation sequences.*

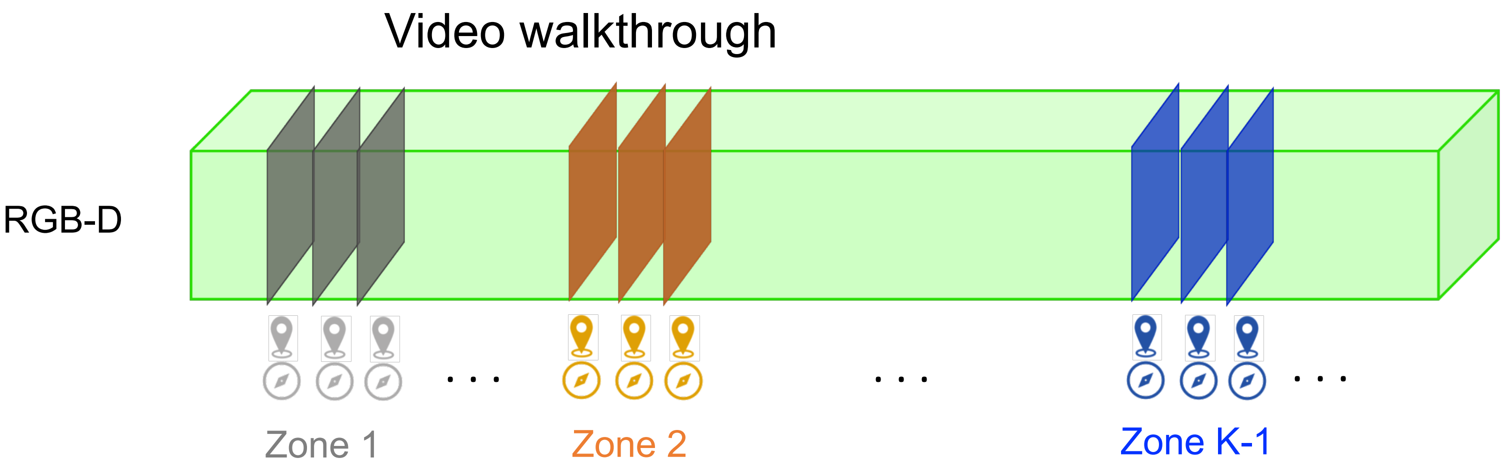
Masked-zone prediction - Step 1: zone creation



- First, we segment the walkthrough into K disjoint frame sets.
- Each frame set is called a *zone*.
- Each zone contains a temporally contiguous set of N frames in the video.

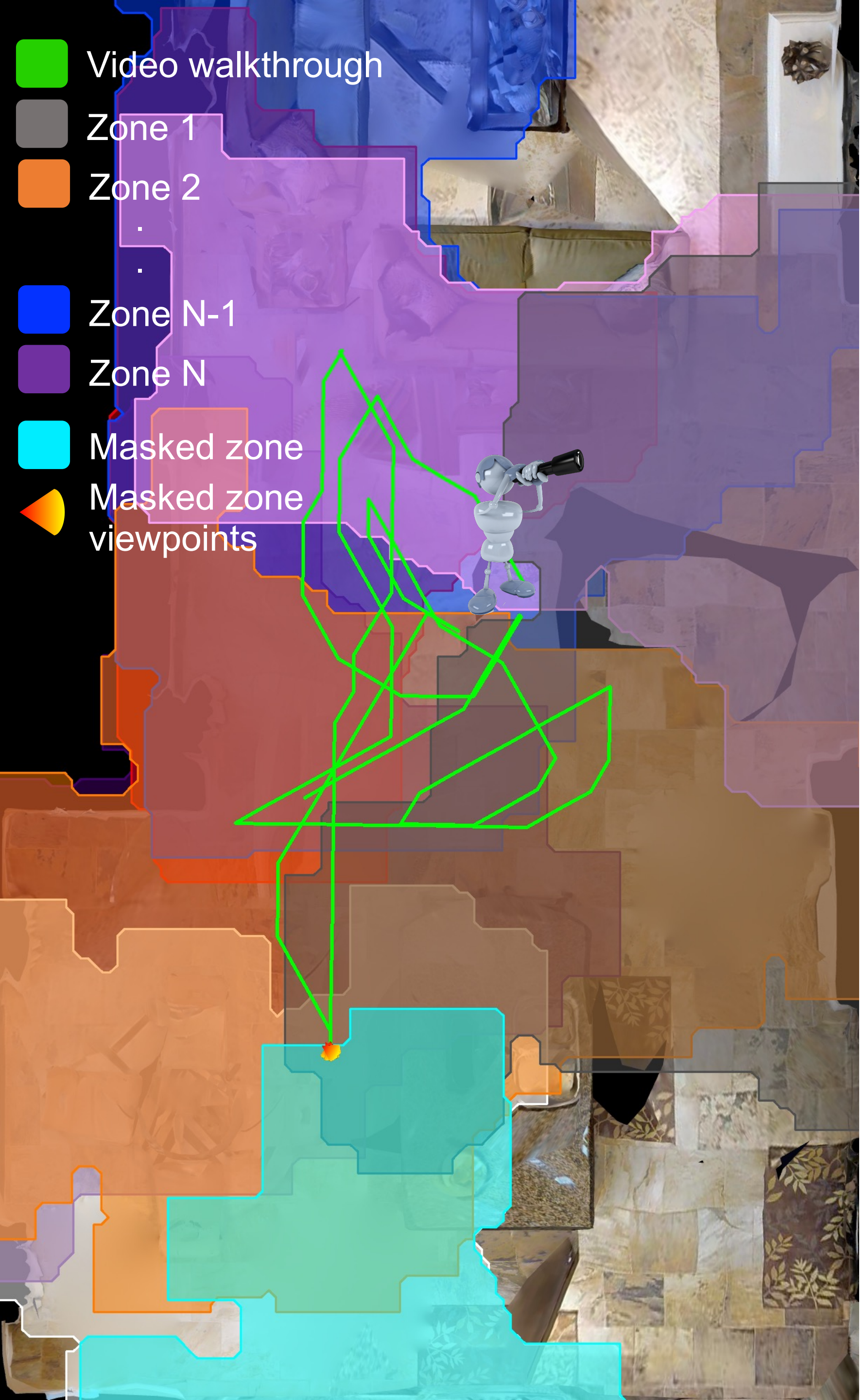


Masked-zone prediction - Step 1: zone creation



- The structure of the zones is shown on the top-down view to the left.
- These zones typically capture partially overlapping regions in 3D.

Masked-zone prediction - Step 2: zone masking



Masked zone

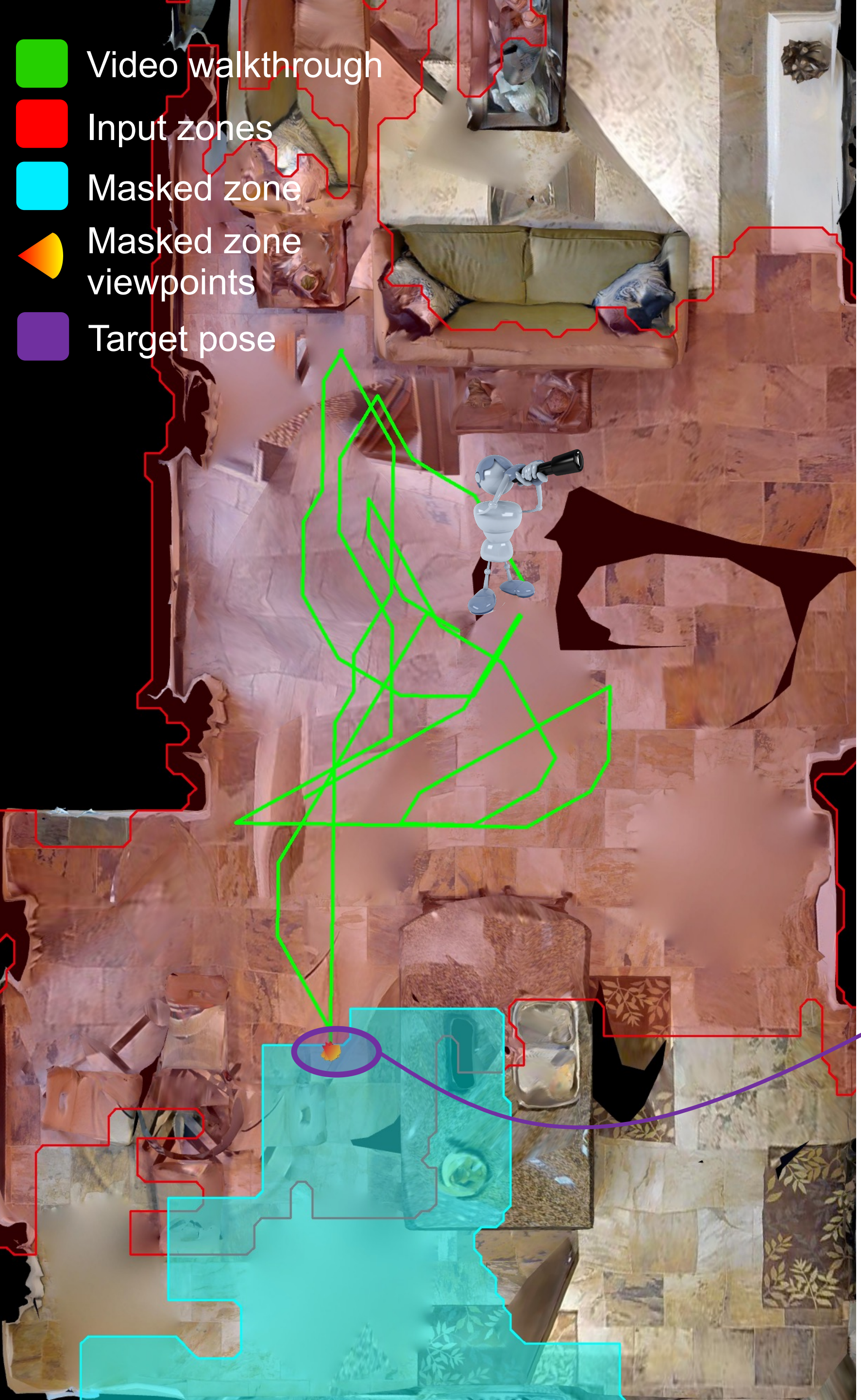
- Next, we mask out one or more zones from the left.
- The viewpoints belonging to a masked zone are shown on the left.
- Some images sampled from the masked zone are shown above.
- This zone contains a *part of a kitchen.*

Masked-zone prediction - Step 2: zone masking

Masked zone

- The sensor readings from the *remaining zones* serve as inputs to the prediction model.

Masked-zone prediction - Step 3: training task



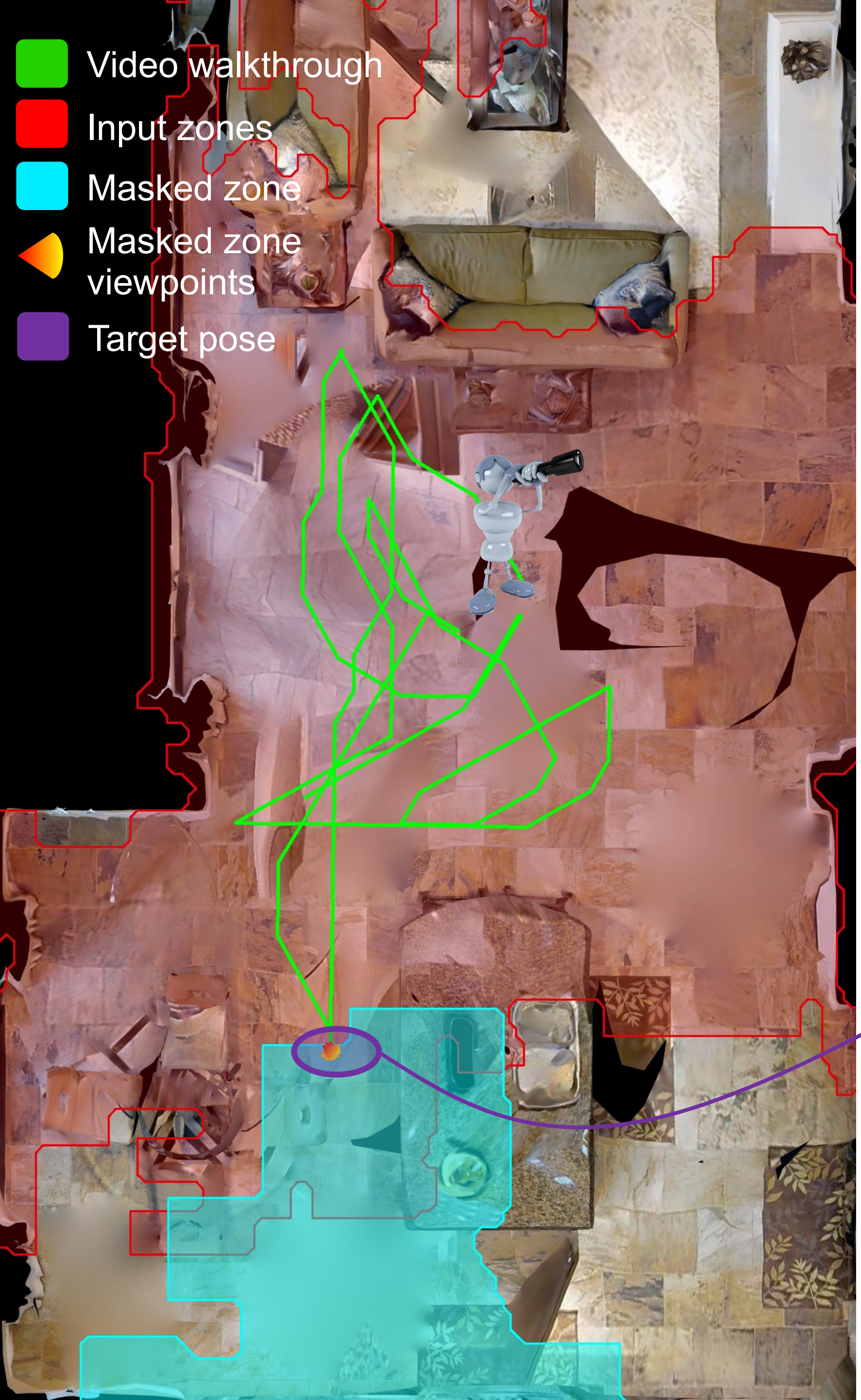
Masked zone

$\rightarrow f_i^u$

Prediction

$\hat{f}_i^u$

- Inputs: frames + camera poses from input zones
- Query: *mean* target pose from masked zone
- Output: predicted zone feature $\hat{f}_i^u$



Masked-zone prediction - Step 3: training task

Masked zone

f_i^u

Prediction

$\hat{f}_i^u$

- Inputs: frames + camera poses from input zones
- Query: a target pose from masked zone
- Output: predicted visual feature $\hat{f}_i^u$
- Contrastive loss function:

Positive zone feature

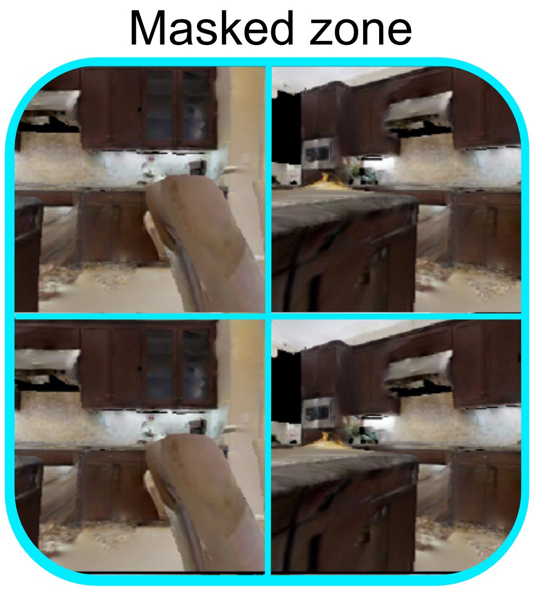
$$L_i = -\log \frac{\mathrm{sim}(\hat{f}_i^u, f_i^u)}{\mathrm{sim}(\hat{f}_i^u, f_i^u) + \sum_{i \neq j} \mathrm{sim}(\hat{f}_i^u, f_j) + \sum \mathrm{sim}(\hat{f}_i^u, f_k')}$$

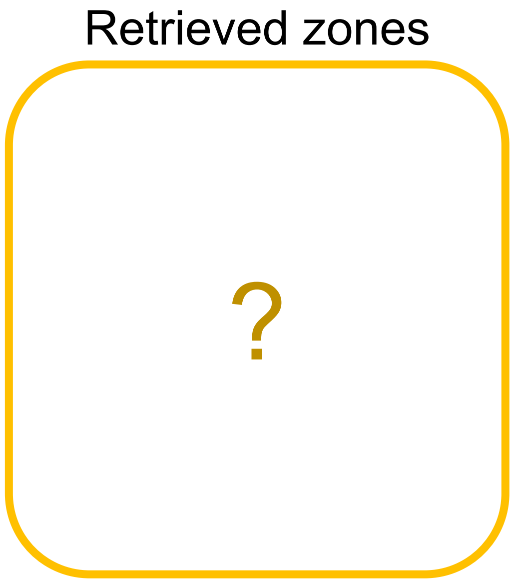
Negative zones from same video

Negative zones from other videos

Next, we visualize the predictions made by our model on the masked-zone prediction task.

Masked-zone prediction - Visualization



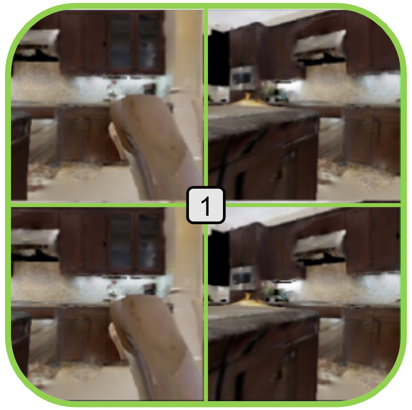


- We visualize the zone prediction $\hat{f}_i^u$ using *inter-video retrieval*.
- We compare the prediction with the ground-truth f_i^u and zones $\{f_k'\}_k$ sampled from other scenes.
- We then visualize the top-4 similar zones from $[f_i^u, f_{k_1}', f_{k_2}', \ldots]$.

Masked-zone prediction - Example

Masked zone

Retrieved zones

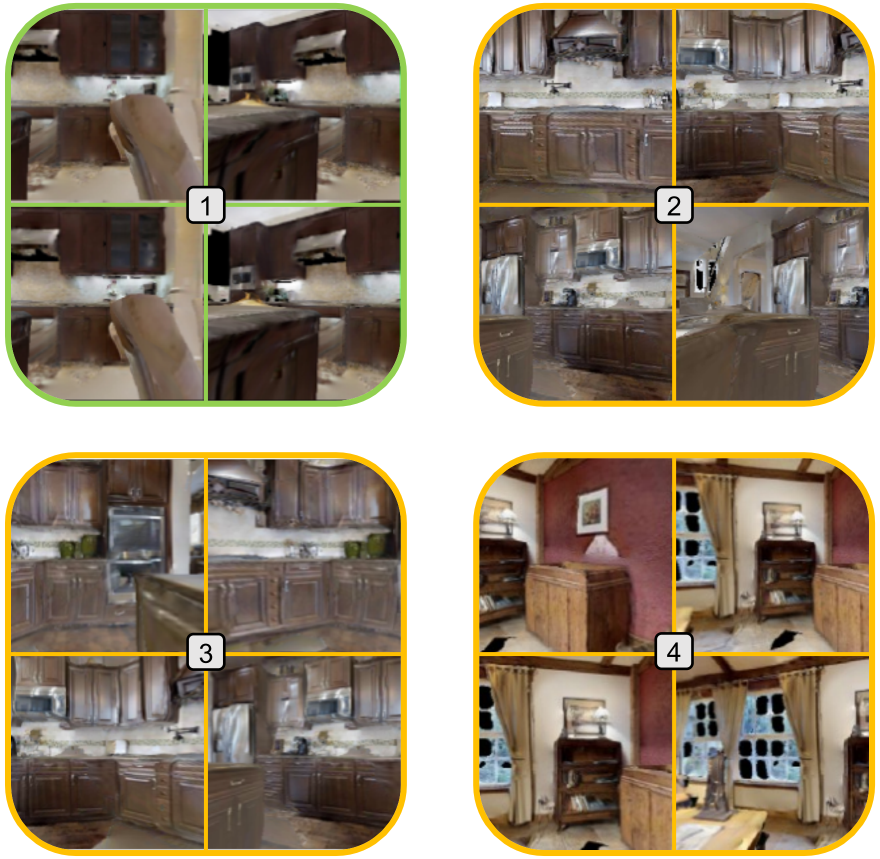


- The model accurately retrieves the ground-truth masked zone at the top.

Masked-zone prediction - Example

Masked zone

Retrieved zones



- The next two retrieved zones from other scenes also correspond to kitchens. This suggests that our learned feature representation captures general semantic concepts.