

A. Open-Vocab Instance-ImageNav dataset

For creating the Open-Vocab Instance ImageNav (OVIIN) dataset, we build upon the dataset generation pipeline presented in [24], while increasing the diversity of object categories and instances. We start by using object categories and instances from the Open-Vocabulary ObjectNav dataset [41], instead of using just the 6 canonical ObjectNav categories [14]. Refer Tab. 3 for a full list of categories. For each object instance, we then generate image goals by capturing images from a set of candidate viewpoints around the object. Next, we filter out invalid image goals based on the object’s visibility – through frame and object coverage heuristics. We use the same parameters for sampling and thresholds same as the Instance-ImageNav (IIN) dataset [24]. Fig. 11 shows more examples of the diverse set of image goals from the resultant OVIIN dataset.

B. LanguageNav Dataset

As outlined in Sec. 4, we sample goal instances for the LanguageNav dataset from the OVON dataset [41]. We start by generating viewpoints for these goal instances – for effectively procuring language description annotations for them. This involves sampling candidate viewpoints within a radius $r \in [1.0, 1.5, 2.0]$ for every 10° sections around the centroid of the object. Next, we collect 512×512 resolution images (using HelloRobot’s Stretch embodiment) from these viewpoints and filter out images from which the target object is not sufficiently visible – by computing the object’s frame coverage. Next, we feed the viewpoint image with the highest frame coverage into the BLIP-2 model [47] to obtain a detailed description of the target object, using the following prompt:

Question: describe the <category>?
 Answer:.

Next, we leverage the semantics of nearby objects to generate meaningful captions. We do so by using ground truth semantics and depth information from the simulator – by selecting objects at least $4.5m$ away (using average depth) and with total frame coverage exceeding 0.5%. For each object, we consolidate its bounding box coordinates, area coverage, object category, semantic index, and a boolean for whether the object is the target object or not.

Finally, using the BLIP-2 generated caption and the bounding box metadata from the previous step, we create a prompt for ChatGPT, using the template shown in Tab. 4. We use the *GPT-3.5-Turbo* model and store the generated response as the language goal description for the target object. We present additional examples from the LanguageNav dataset in Fig. 11.

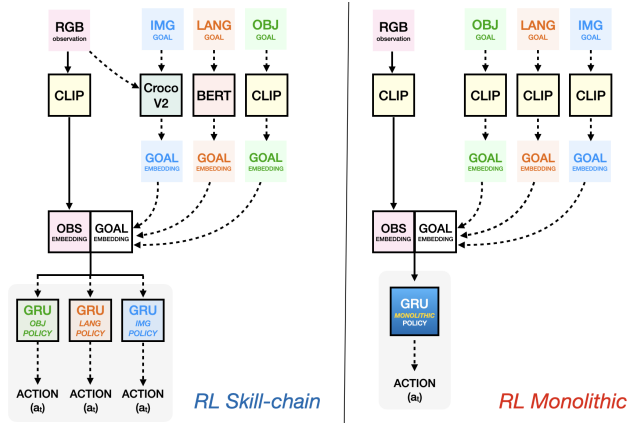


Figure 8. **Model architecture for End-to-End RL Baselines:** We evaluate two types of end-to-end RL policies: **skill chain** (left) and a **monolithic policy** architecture (right). Depending on the current goal specification (object category, language, or image), both methods first concatenate the goal +and current visual observation embeddings. The Skill Chain baseline then uses a policy pre-trained on the task corresponding to the current goal specification (e.g. ObjectNav policy for an object category goal). Once the current goal is reached, the hidden state is dropped and the agent switches to using a different policy depending on the subsequent goal. The monolithic policy, on the other hand, uses the same policy for all three types of goal specifications to predict the next action (while maintaining a hidden state memory).

C. Baselines

We present architecture diagrams summarizing our end-to-end RL (skill chain and monolithic) and Modular GOAT baselines in Fig. 8 and Fig. 10.

C.1. RL Skill Chaining

For our RL skill chain baseline, we train individual policies for LANGUAGENAV, OBJECTNAV and INSTANCE IMAGE-NAV tasks using VER [61]. Each policy takes in RGB observation size 224×224 , which we obtain by resizing and center cropping the original 360×640 image input. The agent also has access to GPS+Compass sensors, which provides location and orientation relative to the start of episode. We embed the pose information to a 32-dimensional vector and concatenate with RGB observation embedding to form current state embedding. Next, as shown in Figure 8, we compute an N-dimensional goal embedding vector using a modality specific goal encoder (separate for each baseline), and concatenate it with state embedding to form an observation embedding o_t . Finally, we feed the observation embedding with previous action into a 2-layer 512-d GRU at each timestep to prediction a distribution over actions a_t . **ObjectNav Policy.** For OBJECTNAV, we encode the object goal category to a target embedding g_t (1024-d vector) using a frozen CLIP pretrained transformer based sentence

air conditioner, amplifier, antique clock, antique telephone, aquarium, arcade game, archway, artwork, baby changing station, backrest, backslash, bag, balcony railing, balustrade, banner, bar, bar cabinet, barbecue, basket, bath towel, bathrobe, bathroom accessory, bathroom cabinet, bathroom shelf, bathroom towel, bathtub, bathtub platform, bed, bed comforter, bed curtain, bench, binder, blanket, board, board game, book cabinet, book rack, bottle, box, brochure, bucket, bulletin board, bunk bed, cabinet, cabinet clutter, cabinet table, candle, candle holder, canvas, cardboard, cardboard box, cart, case, casket, chair, chimney, cleaning clutter, clock, closet, closet shelf, closet shelving, cloth, clothes, clothes bag, clothes dryer, clothing stand, clutter, coat, container, cooker, copier machine, cosmetics, couch, counter, countertop, cradle, crate, crib, curtain valence, cutting board, decorative plate, decorative quilt, desk, desk cabinet, desk lamp, dinner table, dish cabinet, dish rack, dishrag, display cabinet, display table, double armchair, drawer, drawer sink table, dressing table, duct, easy chair, electrical controller, elephant sculpture, elevator, exercise ladder, exhibition panel, fence, figure, fire extinguisher, firewood, fish tank, flag, folding stand, folding table, foosball game table, foot spa, fuse box, gas furnace, gate, globe, grass, grill, guitar frame, gym equipment, hammock, handbag, handle, hat, headboard, heater, high shelf, hook, hose, hunting trophy, hutch, icebox, iron board, jacket, jacuzzi, jar, jewelry box, kitchen appliance, kitchen cabinet lower, kitchen countertop items, kitchen extractor, kitchen lower cabinet, kitchen shelf, kitchen sink, kitchen table, kitchen top, ladder, lamp stand, lampshade, laptop, laundry, laundry basket, light switch, locker, lounge chair, magazine, mantle, massage bed, medical lamp, mixer, monitor, motorcycle, newspaper, note, office chair, office table, ornament, oven, oven and stove, painting frame, pantry, paper, paper storage, paper towel, patio chair, photo stand, pile of magazines, pillar, plate, plush toy, pool, pool table, pot, pouffe, power breaker box, purse, rack, radio, range hood, record player, relief, robe, rocking chair, round chair, sauna oven, scarf, schedule, screen, seat, sewing machine, shade, sheet, shelf cubby, shelving, shower, shower bar, shower cabin, shower rail, shower soap shelf, shower stall, shower tap, shower-bath cabinet, sideboard, sign, sink, sink cabinet, sink table, sled, sleeping bag, sofa chair, sofa seat, sofa set, soft chair, solarium, spa bench, spice rack, stack of papers, stack of stuff, stage, staircase trim, stand, stereo set, stone support structure, stool, storage, storage cabinet, storage shelving, storage space, stove, stovetop, sunbed, support beam, swivel chair, table, table stand, tank, telephone, telescope, tent, tile, toilet, tool, towel, toy, trampoline, trashcan, tray, treadmill, tv, umbrella, urinal, vacuum cleaner, violin case, wardrobe, washbasin, washbasin counter, washer-dryer, washing machine, water dispenser, water fountain, water tank, whiteboard, window, window shade, window shutter, wine cabinet, wood, workstation, worktop, wreath

(a) Training Split

bag, balustrade, basket, bath towel, bathrobe, bathroom cabinet, bathroom towel, bathtub, bathtub platform, bed, bench, blanket, board, box, cabinet, cabinet table, cardboard box, case, chair, closet, closet shelf, cloth, clothes, clothes dryer, container, cooker, couch, countertop, crib, desk, display cabinet, drawer, fire extinguisher, flag, handbag, heater, high shelf, iron board, kitchen appliance, kitchen countertop items, kitchen lower cabinet, kitchen shelf, mantle, monitor, note, office chair, oven, oven and stove, pillar, plush toy, rack, shower, shower cabin, shower rail, shower soap shelf, sideboard, sink, sink cabinet, sofa chair, sofa seat, sofa set, spa bench, stool, storage shelving, stove, support beam, table, toilet, towel, toy, tv, wardrobe, washbasin, washer-dryer, washing machine, window, window shade, window shutter, worktop

(b) Val Seen Split

appliance, armchair, backpack, bar chair, bath cabinet, bath sink, bathroom counter, bed sheet, bed table, bedside lamp, bicycle, bookshelf, chest drawer, chest of drawers, clothes rack, coffee table, computer, computer chair, computer desk, curtain, curtain rod, cushion, desk chair, dining chair, dining table, electric box, exercise equipment, file cabinet, fireplace, folding chair, furnace, ironing board, kitchen cabinet, kitchen counter, kitchen island, lamp, lamp table, railing, refrigerator cabinet, shelf, shower curtain, shower tub, stairs railing, table lamp, tablecloth, throw blanket, tv stand, water heater, window curtain, window frame

(c) Val Seen Synonyms Split

bedframe, blinds, boiler, book, bowl of fruit, calendar, carpet, christmas tree, clothes hanger rod, coffee machine, decorative plant, dishwasher, dresser, exercise bike, flower vase, flowerpot, food, footrest, footstool, freezer, glass, guitar, handrail, hanger, hanging clothes, island, microwave, mirror, nightstand, ottoman, parapet, photo, photo mount, piano, picture, pillow, plant, printer, radiator, refrigerator, rug, shower dial, shower glass, speaker, stair, staircase handrail, statue, vase, window glass

(d) Val Unseen Split

Table 3. Full list of object categories for Open-Vocab ObjectNav dataset [41].

encoder [25]. We train the policy using RL with a navigation reward to minimize the distance-to-target. We train the policy till convergence (300 million steps in our case)

using 4xA40 GPUs with 32 environments on each GPU. We present results of the model checkpoint with the best average performance across all 3 evaluation splits of OVON HM3D

Generate an informative and natural instruction to a robot agent based on the given information(a,b):

a. Region Semantics: <bbox_metadata>

b. Target Object Description: <obj_category>:<BLIP-2 Caption>

Based on the region semantics dictionary which contains information about 2d bounding boxes given in the form of (xmin,ymin,xmax,ymax) in a view of the target object where (0,0) is the top left corner of the frame and a description of the appearance of target object write an language instruction describing the location of the target object, <obj_category>, spatially relative to other objects as references.

There are some rules:

Don't use any absolute values of the numbers, only use relative directions. Do not show bounding box coordinates in the output. Think of giving this as an instruction to a robot agent based on the given details. Add a prefix "Instruction: Find the .." or "Instruction: Go to .." to the generated instruction.

Table 4. ChatGPT prompt used for LanguageNav dataset generation.

Task	VAL SEEN		VAL UNSEEN EASY		VAL UNSEEN HARD	
	Success (↑)	SPL (↑)	Success (↑)	SPL (↑)	Success (↑)	SPL (↑)
1) OVON	32.5	16.3	28.60	14.0	15.70	7.0
2) LanguageNav	17.1	7.8	16.9	8.1	13.5	6.1
3) OVIIN	48.2	22.3	45.8	21.6	45.3	22.3

Table 5. Performance of individual skills used for RL Skill Chain baseline on Open-Vocabulary ObjectNav [41], LanguageNav, and Open-Vocabulary Instance-ImageNav (OVIIN) val splits. We explain each of these policies in Appendix C.1.

dataset.

LanguageNav Policy. For LANGUAGENAV, we encode the language goal to a target embedding g_t (768-d vector) using frozen BERT base uncased [60] sentence encoder. Specifically, we use the output of the [CLS] token as language goal embedding g_t . Similar to the ObjectNav policy, we train the policy till convergence (300 million steps in our case) using 4xA40 GPUs with 32 environments on each GPU and choose the checkpoint with the best average performance across all 3 evaluation splits.

Instance ImageNav Policy. To encode the instance image goal INSTANCE IMAGENAV we use frozen CroCo-v2 [34] image encoder, pre-trained on the Cross-view completion task on a dataset of 3D scanned scene images from Habitat [9] and real world images from ARKitScenes [64], MegaDepth [65], 3DStreetView [66] and IndoorVL [67] datasets. Similar to a recent work [68], we also use adapter layers [69] with CroCo-v2 image encoder during training. We use the publicly released pretrained CroCo-v2 ViT-Base Small-Decoder model from [34]. Due to resource constraints, we resize and center crop the input image to 112×112 during training. We train the policy till convergence (200 million steps in our case) using 4xA40 GPUs with 16 environments on each GPU. Similar to the ObjectNav and LanguageNav policies, we choose the checkpoint with the best average performance across all 3 evaluation splits of INSTANCE IMAGENAV dataset.

The individual task performances are presented in Tab. 5.

We see that the InstanceImageNav (IIN) baseline performs the best across all the three tasks – indicating the efficacy of the cross-view consistent goal embeddings using the CroCo-V2 encoder [34]. On the other hand, the LanguageNav task baseline struggles the most due to the inadequacy of CLIP features in capturing instance-specific information about objects.

C.2. RL Monolithic Policy

Fig. 8 shows the architecture of the monolithic RL policy *i.e.* a single end-to-end policy using multimodal goal encoder and implicit memory trained on the GOAT task. The policy takes in RGB observation size 224×224 , obtained by resizing and center-cropping the original input. We encode the image ($i_t = \text{CNN}(I_t)$) using a frozen CLIP [25] ResNet50 [59] encoder. The agent also has access to GPS+Compass sensors, which provides location and orientation relative to the start of episode. The GPS+Compass inputs, $P_t = (\Delta x, \Delta y, \Delta z)$, and $R_t = (\Delta \theta)$, are passed through fully-connected layers $p_t = \text{FC}(P_t), r_t = \text{FC}(R_t)$ to embed them to 32-d vectors. Next, we compute a 1024-d goal embedding vector g_t^k using frozen CLIP image or sentence encoder based on the subtask s_k goal modality (object, image, or language). All these input features are concatenated to form an observation embedding, and fed into a 2-layer, 512-d GRU at every timestep to predict a distribution over actions a_t - formally, given current observations $o_t = [i_t, p_t, r_t, g_t^k], (h_t, a_t) = \text{GRU}(o_t, h_{t-1})$. To leverage memory from the agent's past experiences in the scene, we carry forward hidden state of the policy from the last subtask $h_T^{(s_{t-1})}$ as initial hidden state for a new subtask $h_0^{(s_t)}$ in a single GOAT episode. We train the policy till convergence using VER [61] (for 500 million steps in our case) using 4xA40 GPUs with 32 environments on each GPU. We choose the checkpoint which has the best average performance across all 3 evaluation splits of the GOAT HM3D dataset.

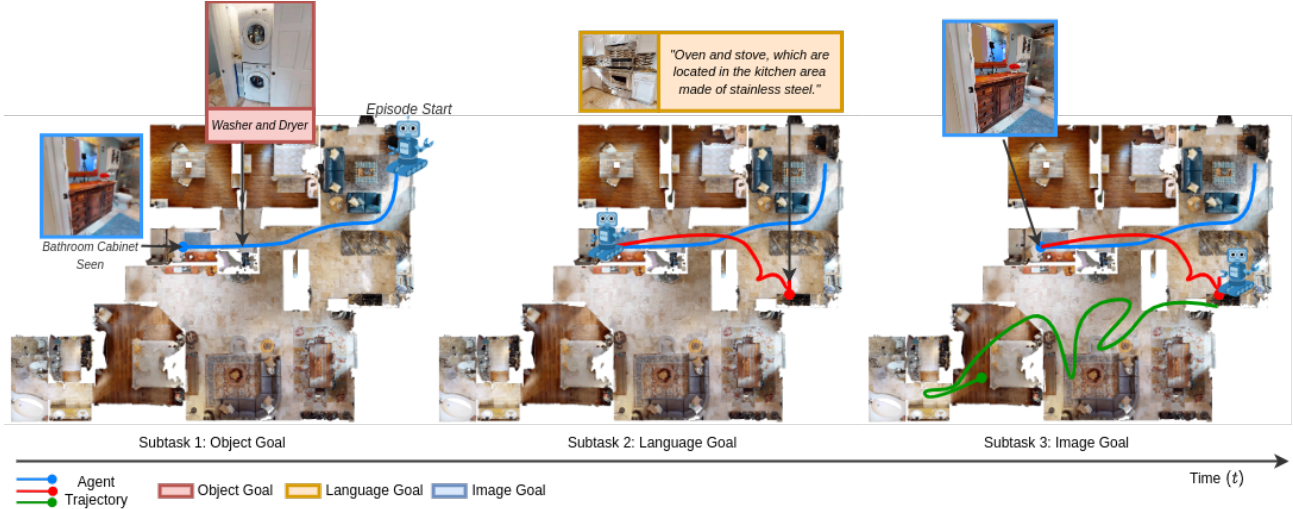


Figure 9. Qualitative example of the monolithic RL baseline agent not remembering objects or regions already seen in the environment. The agent starts by navigating to a washer dryer and sees the bathroom cabinet on its way. However, when tasked with navigating to the same bathroom cabinet in the third sub-task of this GOAT episode, the agent does not go back to the seen region of the house, but instead keeps exploring new regions.

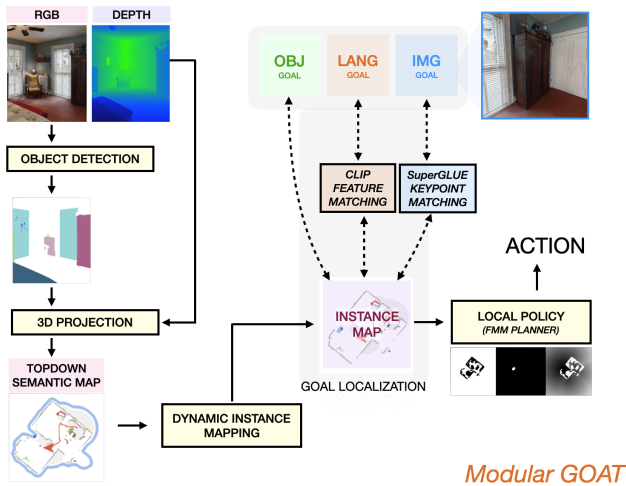


Figure 10. **Modular GOAT Baseline Architecture:** This baseline [42] maintains a semantic and instance-specific topdown map of the environment using a perception module combined with ground truth pose and depth information. This is then used to localize object, language, and image goals – by matching CLIP features or image keypoints. The agent explores the scene using frontier-exploration until a match is found. The goal is then passed to a local policy which predicts low-level actions to reach the goal.

C.3. Modular GOAT

We provide a visual overview of the Modular GOAT baseline, as proposed in [42], in Fig. 10. We direct the reader to prior work [42] for more information about the method.

D. Qualitative Analysis of Monolithic Agent

In this section, we discuss the inability of the monolithic RL policy in capturing past experiences in its implicit hidden state memory. Through Fig. 9, we share a qualitative example exhibiting this behaviour. The agent first looks for a washer dryer in the house (specified by object category). While navigating to the washer dryer, note that the agent also sees the bathroom cabinet shown in the figure. Ideally, we would expect the agent to keep track of this seen object when asked to navigate to it in the future. For the second sub-task, the agent successfully navigates to an oven and stove specified by language. However, for the third sub-task, when the agent it is tasked with navigating to the bathroom cabinet it saw before, the agent wanders around the house and does not make any attempt to visit the seen part of the house again. This highlights the lack of effectiveness of the GRU hidden state in the current implementation – in keeping track of seen objects and regions of the house.

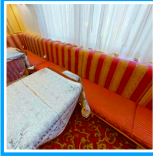
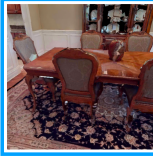
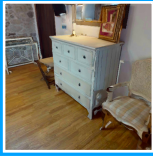
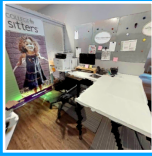

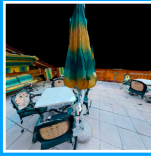


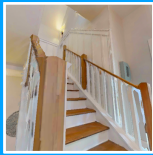



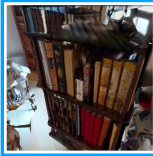

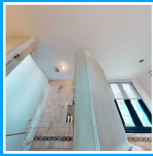
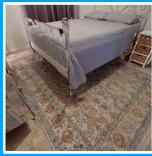


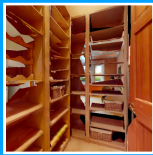
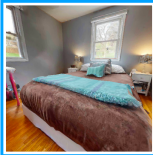




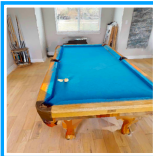

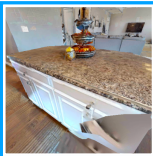
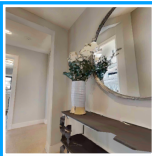


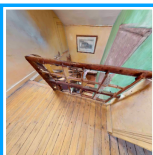

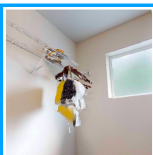

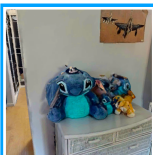

					
<p>Seat that is placed in front of a table.</p>	<p>Wing chair located near the cabinet and the kitchen cabinet lower.</p>	<p>Dresser located below the mirror in the room.</p>	<p>Desk with a computer on it. The desk is located to the left of the board and below the printer.</p>	<p>Carpet located beside the circular sofa and armchair.</p>	<p>Umbrella located to the right of the table, between the swing and the railing.</p>
					
<p>Oven and stove in the kitchen. The oven and stove are located next to the kitchen extractor and the kitchen countertop items.</p>	<p>Wooden shelf in the kitchen located near the refrigerator and table.</p>	<p>Stairs railing with wrought iron railings. The stairs railing is located below and adjacent to the stairs.</p>	<p>The ironing board, which is located near a shelf, a bag, a brush, and a box. The ironing board is a folding ironing board.</p>	<p>Shower soap shelf. The shower soap shelf is a wall-mounted soap holder. It is located near the shower and the shower tap.</p>	<p>Picture in the bedroom. It is located above the bedside cabinet and below the pillow.</p>
					
<p>Hardcover book that is located on the book rack, which is on the left side of the room.</p>	<p>Stainless steel refrigerator located on the shelf.</p>	<p>Shower glass located near the shower soap shelf and bath towel.</p>	<p>The blue and white striped rug located near the bed and bed table.</p>	<p>Footrest. It is located near the sofa, table, bookshelf, and pillow.</p>	<p>Sunbed near the window curtain.</p>
Object goal Language goal Image goal					
					
<p>Shelving by the basket on the carpet. The shelving is made of wood.</p>	<p>Blue and white blanket located on the bed, towards the top right corner, next to the pillow and the bedside cabinet.</p>	<p>Christmas tree by the fireplace on the mantle.</p>	<p>Calendar on the wall, next to the display cabinet, kitchen cabinet, refrigerator, and kitchen counter.</p>	<p>The handrail located below the painting. The handrail is made of wood.</p>	<p>Record player located near the picture and speaker.</p>
					
<p>Pool table in the room with the couch and tv.</p>	<p>Pillow located on the rack towards the middle right side.</p>	<p>Kitchen island, which is located near a bowl of fruit and a throw blanket. It is a granite island with a sink and a stovetop.</p>	<p>Flower vase located near the mirror.</p>	<p>Flowerpot located on the bottom right side of the podium.</p>	<p>Clock located above the shelf and to the right of the furnace. Observe that the clock has roman numerals.</p>
					
<p>Balustrade that is made of wrought iron. Look for it near the parapet, bench, and picture.</p>	<p>Oven next to the kitchen shelf and refrigerator.</p>	<p>Yellow handbag located near the clothes hanger rod.</p>	<p>Archway located near the painting.</p>	<p>The dresser and find the plush toy, which is a teddy bear.</p>	<p>Range hood located above the countertop and to the right of the stove. It is a stainless steel range hood.</p>

Figure 11. Additional multi-modal examples from the GOAT dataset.