FGPrompt: Fine-grained Goal Prompting for Image-goal Navigation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Learning to navigate to an image-specified goal is an important but challenging task for autonomous systems like household robots. The agent is required to well understand and reason the location of the navigation goal from a picture shot in the goal position. Existing methods try to solve this problem by learning a navigation policy, which captures semantic features of the goal image and observation image independently and lastly fuses them for predicting a sequence of navigation actions. However, these methods suffer from two major limitations. 1) They may miss detailed information in the goal image, and thus fail to reason the goal location. 2) More critically, it is hard to focus on the goal-relevant regions in the observation image, because they attempt to understand observation without goal conditioning. In this paper, we aim to overcome these limitations by designing a Fine-grained Goal Prompting (FGPrompt) method for image-goal navigation. In particular, we leverage fine-grained and high-resolution feature maps in the goal image as prompts to perform conditioned embedding, which preserves detailed information in the goal image and guides the observation encoder to pay attention to goal-relevant regions. Compared with existing methods on the image-goal navigation benchmark, our method brings significant performance improvement on 3 benchmark datasets (*i.e.,* Gibson, MP3D, and HM3D). Especially on Gibson, we surpass the state-of-the-art success rate by 8% with only 1/50 model size.

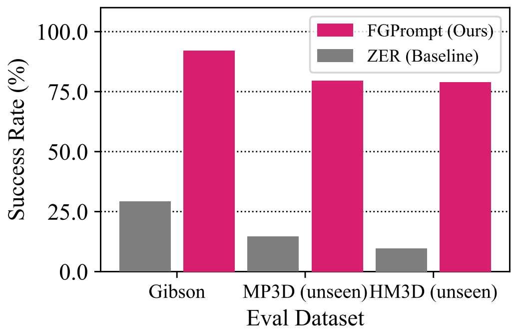
1 Introduction

We focus on the image-goal navigation (ImageNav) task [41] that requires an agent to navigate to an image-specified goal position and face the same orientation as where the photo is taken. In this task, the agent needs to explore the environment and try to find the objects with their surroundings that best match the ones specified in the goal image. As an image is a clearer description than language, it shows a wide range of application prospects on household robots [19] or self-driving vehicles, serving as a navigation goal or intermediate landmark.

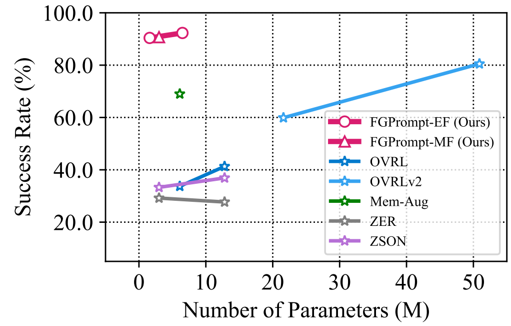
Despite its wide applications, this task is still very challenging for the embodied agent due to the following two aspects. First, compared to object-goal navigation which assigns goal descriptions with specific semantic categories, it requires the agent to perceive the visual observation as well as the goal image and make a comprehensive understanding of the scene in order to identify goal-relevant objects. Second, objects share similar semantic meanings within one environment, making it challenging to accurately find out the desired object instance.

Previous methods [25, 7, 15, 8, 30, 6, 2] seek to solve this task by decomposing the navigation system into several modules in isolation. In general, they tend to adopt efficient exploration skills to build a map as the understanding of the scene, incrementally update the map and localize the agent's position at each time step, and further predict a waypoint to navigate to. However, these map-based methods

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.



(a) Success rate comparison with *baseline* (ZER [42]) on three different datasets. Our method performs efficiently and robustly in both seen (*i.e.*, Gibson) and unseen (*i.e.*, MP3D and HM3D) environments.



(b) Comparison with SOTA both on success rate and the number of parameters. the FGPrompt-EF, an early fusion variant of our method, achieved 90.4% success rate with only 1/50 model size compared to SOTA.

Figure 1: Main results of our proposed FGPrompt on the image navigation task.

require depth maps or the agent's GPS position to build the occupancy map or topological map. The latest methods [11, 23, 42, 22, 37, 36] instead try to learn a navigation policy in an end-to-end manner using reinforcement learning. These methods set up two different encoders to obtain semantic embeddings from goal and observation images independently. Subsequently, a recurrent model takes these embeddings as input to predict a possible action sequence. However, they suffer from two major limitations: 1) As the details in the goal image are gradually overlooked as it goes through deeper network layers, it is harder to find useful cues for reasoning and finding the goal location. 2) Existing methods leave the goal image apart from the observation when performing encoding, it is hard for the agent to focus on the goal-relevant regions in the observation since there is no goal prompting to guide the agent to understand the observation. In this paper, we focus on addressing these limitations to improve navigation performance.

When people try to find a place captured in an image, they must look for the contextual cues presented with objects, shapes, colors, and textures in both the goal images and current visual observation. Spatial reasoning based on this information plays a critical role in understanding the scene, as people always compare and identify similarities, in order to consider the relative position of various elements and gain insights into the current position and the target location. Motivated by this fact, instead of considering only semantic features of goal and observation images, we propose a novel fine-grained goal prompting (**FGPrompt**) architecture to learn observation embeddings conditioned on the fine-grained and high-resolution features of the goal image.

Specifically, we implement the goal prompting scheme as a fusion process between the goal and observation images and design a mid fusion (FGPrompt-MF) mechanism. This mechanism leverages fine-grained and high-resolution feature maps in the intermediate goal network layers as the prompts. These feature maps are proven to contain informative object details [16, 40]. Hereafter, conditioned on these feature maps, we utilize FiLM [26] layers to learn a transform function to adjust the observation activations to focus on goal-relevant objects. In addition, we also design an early fusion (FGPrompt-EF) mechanism by concatenating the goal and observation images at the pixel level. We then use a neural network to jointly model the concatenated image and implicitly fuse their information. Experimental results on the ImageNav benchmark show our proposed method significantly outperforms state-of-the-art methods, especially in both generalization ability to unseen environments and efficiency, as shown in Figure 1.

To sum up, our contributions are as follows: 1) We propose a novel fine-grained goal prompting method for the image-goal navigation task, from which the agent learns to understand visual observations conditioned on the fine-grained information from the goal image, and thus pay more attention to goal-relevant objects to reason the target location. 2) We explore different mechanisms to perform fine-grained goal prompting and find that both the mid fusion (FGPrompt-MF) and early fusion (FGPrompt-EF) mechanisms draw significant improvements compared to the late fusion baseline. 3) With FGPrompt, our agent robustly understands the scene and finds objects relevant to the goal image. On ImageNav, our method improves the navigation success rate by 10.3% and 14.4% under default and panoramic settings, respectively.

2 Related Work

Modular methods. Modular methods leverage strictly defined modules that are handcrafted [30, 19] or learnable [7, 15, 14, 8, 30, 6, 2] to address the image-goal navigation task step by step. Classical modular methods typically combine the exploration [38] component, simultaneous localization and mapping (SLAM [12, 35]) component, and path planning component to achieve the navigation goal. In order to localize the agent in an unknown environment, some approaches build an explicit metric map of the environment [7, 15], while others propose to obtain an implicit latent map [14] like a topological map [8, 30] or simply adopt object detectors without mapping [28]. Chaplot *et al.* [6] and Avraham *et al.* [2] train supervised deep models to tackle the sub-tasks, which require a lot of annotated data. Although off-the-shelf modules can be used with zero fine-tuning [19], they still heavily rely on pose and depth sensors, which greatly limits their applicability in the real world.

RL-based navigation. Another pipeline for ImageNav is to directly learn from interactions with the environment using reinforcement learning (RL). RL-based navigation tends to learn an end-to-end reward-driven policy that maps observation to action [37, 36, 42, 22, 23] and shows great potential in this task. However, these methods still face the challenge of the sparse reward mechanism and poor generalization performance. To address these issues, previous works propose different methods to encourage the agent to explore more efficiently. Yu *et al.* [11] combines RL policy and visual representation learning model in a min-max game way to incentivize the agent to explore its environment. Al-Halah *et al.* [42] proposes a zero-shot transfer learning approach with a novel reward for its semantic search policy. Similarly, Majumdar *et al.* [22] uses a pre-trained CLIP to enhance image embedding. To tackle the long-horizon planning problem, an external memory module has been proposed by [23, 13, 3, 30, 20, 18] that learns a topological graph [13, 3, 30, 20, 18] or attention map [23] online. Self-supervised learning paradigm has also been explored by Yadav *et al.* [37, 36] to endow the navigation model with better representation ability. Different from existing approaches, we proposed a goal-prompted observation understanding method that learns to focus on goal-relevant objects through fine-grained goal prompts.

Goal-conditioned learning. Existing RL-based navigation methods can be interpreted as learning a goal-conditioned policy, since they only perform fusion on the latent goal embedding and observation embedding. Only semantic-level information can be exchanged during fusing. Some embodied robot planning methods [4, 33, 17, 39] learn a goal-conditioned observation encoder by injecting the goal embedding to it. Stone *et al.* [33] and Brohan *et al.* [4] only consider the language as the goal description, while Jang *et al.* [17] and Yu *et al.* [39] try to fuse the goal image with the intermediate feature maps of observation encoder using an affine transformation proposed by FiLM [26]. However, they still focus on the latent embedding of goal images and neglect the fine-grained information in high-resolution activation maps. In this paper, we propose to make use of the intermediate activations in the goal encoder as informative guidance to condition the learning of the observation encoder.

3 Image Goal Navigation using Fine-Grained Goal Prompting

3.1 Task definition

Image-goal navigation (ImageNav) requires an agent to navigate to a goal position that matches where the goal image v_g was shot. Specifically, the agent starts at a random location p_0 and only receives a goal image v_g from the environment. At each time step t, the agent receives an egocentric RGB image v_t captured by a RGB sensor fixed on its body, and executes an action a_t conditioned on v_t and v_g. In RL-based methods, the action a_t is selected based on the learned policy. After performing the action a_t, the agent will be assigned a reward r_t that encourages the agent to reach the goal position as soon as possible. A more detailed definition of our setup can be found in Section 4.

Existing RL-based methods tackle the ImageNav problem by learning an observation encoder and a goal encoder separately, and then fusing their output embeddings together. As shown in Figure 2 (a), this fusion module is commonly equipped on most of the baseline methods. However, those embeddings preserve little detailed information, *e.g.*, shape, texture, and spatial relationship, to promote finding and comparing objects relevant to the goal image [40, 16]. To tackle this challenge, we propose to leverage fine-grained information from lower-level goal image features as prompts to promote the agent's ability to focus more on goal-relevant objects.

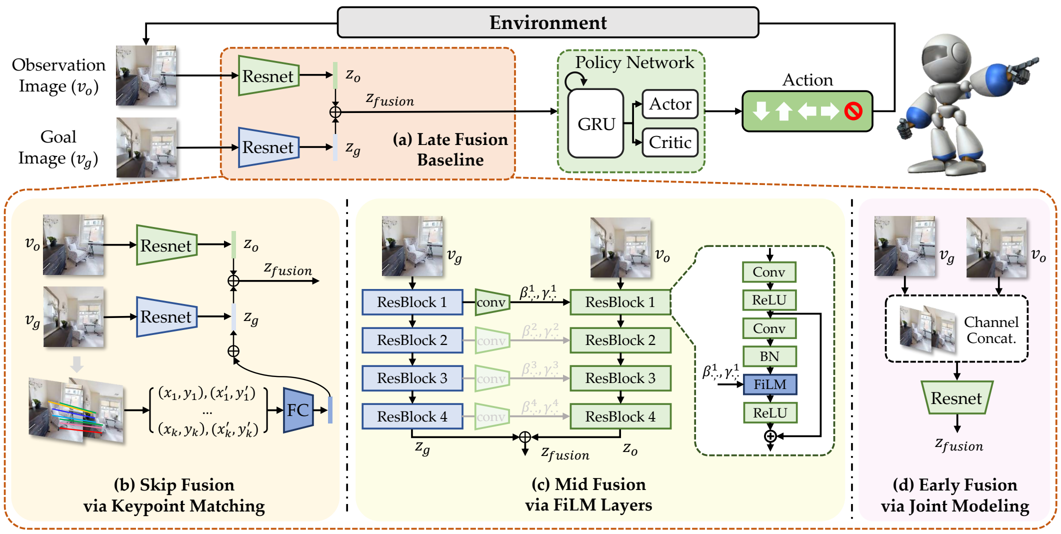


Figure 2: **Illustration of baseline fusion (a) and our goal prompting (b, c, d) for image-goal navigation**. All these methods take observation and goal images as input and output fused features.

3.2 Fine-grained Goal Prompting

We design and explore three different fine-grained goal prompting methods that vary from fusion mechanism, namely **Skip Fusion**, **Mid Fusion**, and **Early Fusion**. For the first Skip Fusion mechanism, we investigate injecting fine-grained goal prompting utilizing a handcrafted keypoint matching module. After that, we replace the handcrafted matching module with learnable affine transform layers to enable active prompt learning and propose the Mid Fusion mechanism. Finally, we simplify the above mechanisms by introducing a joint modeling framework to perform implicit fusion, w.r.t. Early Fusion. Details of our proposed methods are as follows.

Skip Fusion via Keypoint Matching. We first attempt to equip the baseline late fusion model with the ability to benefit from fine-grained information in the goal image, we attach an additional low-level fusion module using handcrafted keypoint matching methods [21, 29], as an improvement of the Late Fusion baseline. We name this mechanism Skip Fusion as it fuses the goal image and observation image in the both early and later stage but skip the others, as shown in Figure 2 (b).

Keypoint matching, which aims to discover representative keypoints in an image and then describe and match them with the most similar ones in another image. As these points are detected based on the low-level statistic [21, 9] of image pixels, we leverage them to play a role as low-level fusion. This scheme is handcrafted as it is not learnable during training. To enable batch inference, we leverage a deep learning-based keypoint detecting [10] and a matching [29] method to obtain matched keypoint between the goal image and the observation image. Hereafter, we select top-k matched points according to their matching score to compose a variable z_k and concatenate them together with z_g and z_o as the fusion result:

$$z_{fusion} = z_g \oplus z_o \oplus \text{FC}(z_k) \quad (1)$$

where $z_k = (x_1, y_1, x'_1, y'_1, ..., x_k, y_k, x'_k, y'_k)$ is a flattened vector of k keypoints. The default value is set to -1 if the number of matched keypoints is fewer than k.

Mid Fusion via FiLM Layers. The handcrafted keypoint matching module may not work in a situation *where the observation does not shoot the same objects with the goal image*. A feasible solution is replacing the handcrafted low-level fusion module with a learnable fusion scheme. Previous literature [17, 39] inputs the goal embedding into the ResNet visual backbone via FiLM [26] layers, which adapt a learnable affine transformation conditioned on the input embedding to the intermediate activation maps in each residual blocks. Through these layers, we can easily connect the intermediate layers in both the goal encoder and the observation encoder to perform mid fusion.

Different from the existing approaches that leverage abstract language embedding as a global condition for all layers, we propose to use the hierarchical representations from the intermediate goal encoder layers. This allows us to make good use of the fine-grained information in high-resolution feature maps. Specifically, we perform FiLM affine transformation on the resnet blocks of the observation encoder, where the affine factors $\beta^i_{\cdot,\cdot}, \gamma^i_{\cdot,\cdot}$ in block i are conditioned on the shaped activation map z^i_g from the correspondent block of the goal encoder. This process can be formulated as:

$$\gamma^i_c = f_c(z_g) \quad \beta^i_c = h_c(z_g) \tag{2}$$

$$\hat{z}^i_o = \gamma^i_c z^i_o + \beta^i_c \tag{3}$$

where $\hat{z}^i_o$ denotes a transformed activation map in block i and c denotes the c^{th} feature of the feature map. The function f and h learn to map the condition variable into the affine factors. In practice, we implement them as 1×1 convolutions to maintain the same resolution between the input and target activation map. Section 4.2 further investigates the choices of the mapping function and the number of FiLM layers. The output from the conditioned observation encoder f_o can then be viewed as the fused feature z_{fusion}, as shown in Figure 2 (c). The fused feature can be written as:

$$z_{fusion} = f_o(v_o|v_g) \tag{4}$$

Early Fusion via Joint Modeling. As discussed above, the mid fusion mechanism casts the intermediate activation map of goal observation v_g as a fine-grained prompt for the observation encoder f_o, however, it requires separate encoding, introducing multi-stage projection and transformation with additional parameters and computation. One possible solution to simplify this mechanism is directly fusing those two images very early and then jointly modeling them using the same encoder. In particular, we concatenate the goal image with the observation image on the RGB channel dimension, resulting in an input tensor shaped $128 \times 128 \times 6$. This concatenated tensor is then fed into a ResNet encoder with a stem convolution layer that takes the 6-channel image as input. Detailed ablation on the early fusion operation can be found in Section 4.2. In this case, the fusion mechanism can be written as:

$$z_{fusion} = f_o(v_o \oplus v_g) \tag{5}$$

3.3 Navigation Policy

Based on the fused embedding z_{fusion} of the goal image and observation image, we train a navigation policy π using reinforcement learning (RL):

$$s_t = \pi(z_{fusion} \oplus a_{t-1}|h_{t-1}) \tag{6}$$

where s_t is the embedding of the agent's current state. h_{t-1} denotes hidden state of the recurrent layers in policy π from previous step. Following previous methods [42, 22], we adopt an actor-critic network to predict state value c_t and action a_t using s_t and train it end-to-end using PPO [32]. We utilize the ZER reward [42] to encourage the agent to not only reach the goal position but also face the goal orientation. More details can be found in Appendix.

4 Experiments

Datasets. We use the Habitat simulator [31, 34] and train our agent on the Gibson dataset with 72 training scenes and 14 testing scenes under the standard setting. We use the training episodes provided by [23] and trained our agent for 500M steps. We report results under multiple datasets to allow direct comparison to various prior works. On the Gibson dataset, we validate our agent on split A generated by [23], and split B generated by [15]. On the MP3D and HM3D, we use the test episodes collected by [42].

Agent configuration. We follow the recipe of previous trails [42, 22, 37] to initialize an agent equipped with only RGB cameras of 128×128 resolution and $90°$ FOV. When compared with methods that use a panoramic input, we initialize four RGB sensors to the front, left, right, and back directions of the agent, following [23, 37]. The agent's action space is comprised of four discrete actions, including MOVE_FORWARD, TURN_LEFT, TURN_RIGHT, STOP. The minimum units of rotation and forward movement are $30°$ and 0.25m respectively.

Method	Backbone	Pretrain	Sensor(s)	Memory	Split	SPL	SR
NTS [8]	ResNet9	N/A	RGBD+Pose	✗	A	43.0%	63.0%
Act-Neur-SLAM [6]	ResNet9	N/A	RGB+Pose	✗	A	23.0%	35.0%
SPTM [30]	ResNet9	N/A	RGB+Pose	✗	A	27.0%	51.0%
ZER [42]	ResNet9	N/A	RGB	✗	A	21.6%	29.2%
ZSON [22]	ResNet50	OSD	RGB	✗	A	28.0%	36.9%
OVRL [37]	ResNet50	OSD	RGB	✗	A	27.0%	54.2%
OVRL-V2 [36]	ViT-Base	HGSP	RGB+Pose	✗	A	58.7%	82.0%
FGPrompt-MF (Ours)	ResNet9	N/A	RGB	✗	A	62.1%	90.7%
FGPrompt-EF (Ours)	ResNet9	N/A	RGB	✗	A	66.5%	90.4%
FGPrompt-EF (Ours)	ResNet50	N/A	RGB	✗	A	**68.5%**	**92.3%**
Mem-Aug [23]	ResNet18	N/A	4 RGB	✓	A	56.0%	69.0%
VGM [20]	ResNet18	N/A	4 RGB	✓	A	64.0%	76.0%
OVRL [37]	ResNet50	OSD	4 RGB	✗	A	62.5%	79.8%
TSGM [18]	ResNet18	N/A	4 RGB	✓	A	67.2%	81.1%
FGPrompt-EF (Ours)	ResNet9	N/A	4 RGB	✗	A	**75.0%**	**94.2%**
NRNS [15]	ResNet18	N/A	RGBD	✗	B	12.4%	24.0%
FGPrompt-EF (Ours)	ResNet9	N/A	RGB	✗	B	**70.5%**	**93.0%**

Table 1: **Comparison with state-of-the-art methods on Gibson**. All methods are trained and evaluated both on the Gibson dataset.

Methods	Backbone	MP3D		HM3D	
		SPL	SR	SPL	SR
Mem-Aug [23]	Resnet18	3.9%	6.9%	3.5%	1.9%
NRNS [15]	Resnet18	5.2%	9.3%	4.3%	6.6%
ZER [42]	Resnet9	10.8%	14.6%	6.3%	9.6%
FGPrompt-MF (Ours)	Resnet9	**50.4%**	**77.6%**	**49.6%**	**76.1%**

Table 2: **Cross-domain evaluation on MP3D and HM3D**. The agent is trained in Gibson environments and directly transferred to new environments for evaluation.

Evaluation metrics. We report the success rate (SR) and Success weighted by Path Length (SPL) [1], which takes into account path efficiency of the navigation process. An episode is considered successful if the agent stops within 1.0m Euclidean distance from the goal location and the maximum number of steps in an episode is set to 500 as the default setting.

4.1 Comparison with State-of-the-art Methods

Evaluation on Gibson. In Table 1, we report the ImageNav results on Gibson averaged over three random seeds (the variances of all random seed results are less than 1e-4.). We compare our methods with state-of-the-art methods in two different settings, one takes only one RGB sensor as input following [42, 22, 37] and another one takes 4 RGB sensors to assemble a panoramic view following [23, 37]. For the SLAM-based methods in the first three rows, we report the number reproduced by Mezghani *et al.* [23]. We found that our proposed FGPrompt-MF and FGPrompt-EF methods take an absolute advantage compared with all previous methods. Even compared to OVRL-V2 [36], a method that utilizes a much larger visual backbone (ViT-B) pre-trained on an in-domain image dataset, we still achieved large performance gains on both SR (92.3% vs. 82.0%) and SPL (68.5% vs. 58.7%) in the absence of additional pose sensor input. This finding reveals the effectiveness and efficiency of our proposed method.

We extend our FGPrompt-EF to the panoramic view setting (4 RGB) for direct comparison with some memory-based methods [23, 20, 18] and pre-trained method [37]. We found that our FGPrompt-EF outperforms these memory-based methods by at least 13.1% in success rate and 7.8% in SPL, even without additional external memory module and pre-training phase. Besides, we also provide a comparison result on the non-mainstream testing episodes (split B) following [15]. Compared with the self-supervised method NRNS [15] that pretrained on passive videos, our FGPrompt-EF brings 58.1% improvement in success rate and 69.0% in SPL, which shows a great advantage by learning to understand the scene based on goal prompting through interacting with the environment.

Setting	SPL	SR
Later Fusion (baseline)	11.2%	13.0%
Skip Fusion via keypoint matching (FGPrompt-SF)	24.7%	41.6%
Mid Fusion via FiLM layers (FGPrompt-MF)	50.4%	77.3%
Early Fusion via joint modeling (FGPrompt-EF)	**54.7%**	**78.9%**

Table 3: **Comparison of different goal prompting methods on Gibson ImageNav task**. Fusing the fine-grained goal prompts with the observation instead of directly concatenating their semantic embeddings yield significant improvement.

Mapping Method	SPL	SR
N/A	11.2%	13.0%
Semantic Mapping	24.0%	32.0%
FG/HR Mapping	**50.4%**	**77.3%**

Table 4: **How to map activation into affine factors?** Using Fine-grained High-resolution (FG/HR) mapping performs significantly better.

Depth	SPL	SR
1	**50.4%**	77.3%
2	49.3%	**77.6%**
4	50.2%	71.4%

Table 5: **How deep should the Mid Fusion perform?** Performing Mid Fusion on the early layers works better than on all layers.

Cross-domain evaluation on out-of-domain datasets. In Table 2, we report the cross-domain evaluation results on the unseen scenes in the Matterport3D (MP3D) [5] and HM3D [27] to verify the generalization ability from seen environments to unseen environments. Following [42], we directly transfer our model trained on Gibson to these two new datasets, without any tuning. Since there exists a very large visual domain gap between these datasets, as well as more complex and larger scenes in MP3D and diverse scene types in HM3D, this setting is extremely challenging. We leverage the testing episodes released by ZER [42]. Compared with the baseline method ZER, our fine-grained and high-resolution conditioned embedding method brings 7× improvements in the success rate without any additional effort, which shows the generalization ability of our method.

4.2 Ablation Study

In Section 3.2, we introduce three different types of goal prompting methods, varying from the fusion mechanism. In this section, we first compare the effectiveness of different methods on the ImageNav task. Then we present the detailed ablation of each method to empirically discover their best implementation. For convenience and fairness, all variants in the ablation study are trained for 50M steps on the Gibson dataset.

Comparing different goal prompting methods. We first compare the proposed goal prompting methods on the image-goal navigation task. As shown in Table 3, the Skip Fusion (FGPrompt-SF) variant, integrated fine-grained information by simply adding a keypoint matching-based fusion module to the baseline, performs significantly better on the ImageNav task (from 14.0% to 41.4%). This reveals that fine-grained goal prompting is important as it provides the navigation policy informative cues to compare and find goal-relevant objects. However, when the observation does not shoot the same objects with the goal image, there are no available matching keypoints to serve as low-level goal prompts, which may hinder the performance. The other two variants further exchange information in a learnable manner to tackle these problems. In detail, the Mid Fusion (FGPrompt-MF) mechanism leverages the intermediate activation maps with varied resolutions to perform goal prompting. In this case, the agent learns to understand visual input and focus on possible goal-relevant regions based on the fine-grained prompts from the goal image. As a result, this variant further increases the navigation success rate by 27.2%. Besides, as a simplified version of our proposed Mid Fusion mechanism, the Early Fusion mechanism enables an implicit fusion process through jointly modeling the goal and observation images. This scheme learns to exchange information between two input images implicitly and thus requires no expertise to design a proper fusion mechanism. In Table 3, this simple but ingenious design brings a further improvement (4.3% in SPL) compared to the Mid Fusion mechanism which is well-designed and ablated. We attribute this to its adaptive and learnable fusion fashion.

Setting	SPL	SR
3D stack	17.3%	20.5%
Edge concat	37.2%	54.8%
Channel concat	**54.7%**	**78.9%**

Table 6: **How to perform early fusion?** A naive concatenation at the channel dimension works the best.

Setting	SPL	SR
Separate modeling	11.2%	13.0%
Tied modeling	12.3%	14.6%
Joint modeling	**54.7%**	**78.9%**

Table 7: **Does joint modeling works?** Yes, it greatly boosts navigation performance compared to the baseline and another similar approach.

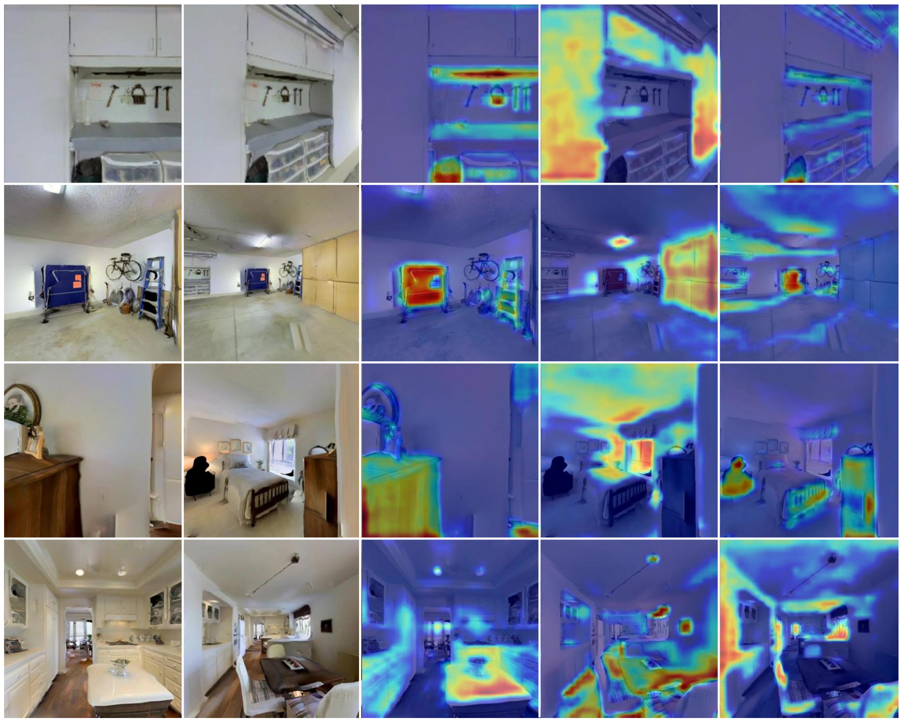


Figure 3: **EigenCAM visualization of the activation map in the fusion layer of FGPrompt-MF.** Images in different rows illustrate results in different testing episodes in Gibson. The Mid Fusion mechanism learns to focus on the objects that are relevant to the goal image.

Ablation on the Mid Fusion mechanism. We further ablate to investigate the detailed setting of our proposed Mid Fusion mechanism, which takes advantage of FiLM [26] layers to apply fusion based on fine-grained goal-conditioned affine transformation. In contrast to existing goal-conditioned methods, we point out that fine-grained information in high-resolution feature maps is a key to understanding visual observation. To verify the necessity of this information for an embodied agent, we conduct ablation studies on the FiLM layers in Table 4. We design two different mapping methods that map the activation map into the affine factors in Equation 2, namely Semantic Mapping and Fine-grained High-resolution Mapping. Specifically, for the former, we average pool the activation map in each layer within the spatial dimension, removing the fine-grained information in this layer, and then leverage two separated fully connected layers to perform mapping. For the latter method, we keep the spatial resolution of the original activation maps, hence preserving the fine-grained information. We initialize two convolution layers with 1×1 stride to learn a mapping function. Not surprisingly, only taking the coarse-grained input from the goal encoder as a condition leg a lot behind, as it lose lots of details that might serve as possible cues during the pooling,

Another important question is how deep the network layers should be considered to perform fusion. Since the perception field glows as the feature map resolution reduces in deeper layers, the information about objects and scenes in these layers could be more and more coarse-grained. We design an ablation study that integrates a different number of network layers to perform fusion. As shown in Table 5, we found that fusing the first two network layers (each layer indicates an entire Resnet block) performs well, indicating that fine-grained information in the early layers is important for goal prompting. When the fusion depth increases to 4 layers, the navigation performance slightly degrades, as considering more prompting layers increases the learning difficulty.

Ablation on the Early Fusion mechanism. Firstly, we conduct an ablation study to find out how to perform early fusion on the goal image and observation image. To achieve a unified model for both two input images, there exists a naive approach to merge them at the pixel level. In particular, we try to concatenate these two images on the different dimensions, as shown in Table 6, where concatenation on the channel dimension performs better than other choices. We conjecture that aligning and modeling the goal and observation images enables spatial reasoning, which endows the agent with a better ability to understand and deduce the relevant regions in visual observation to explore. We also investigate stacking the two images at an additional axis and performing 3D convolution to embed them together. Interestingly, results in Table 6 show that this variant failed to learn an effective fusion process, although it aligns both images in the spatial dimension.

Secondly, in order to determine the effectiveness of our proposed joint modeling scheme that takes both the goal and observation image as input, we compare it with a similar approach that shares the same parameters between the goal encoder and observation encoder following [23], namely Tied Modeling. In Table 7 we directly compare them with a baseline that learns a goal encoder and an observation encoder separately. We observe that the Tied Modeling variant performs worse similar to the Separate Modeling baseline. Though using shared parameters to encode both goal and observation images, this architecture does not enable goal-prompted learning to focus on the goal-relevant regions and thus failed to effectively reason the goal position.

4.3 Analysis and Qualitative Visualizations

How does the fine-grained goal prompting work? We visualize the activation maps using Eigen-CAM [24] before and after the fusion layers of our mid fusion goal prompting method (FGPrompt-MF) to find out how it works in the image navigation task. Illustrations are presented in Figure 3. Prompted with the fine-grained and high-resolution activation map from the goal image, the agent is able to find out the relevant objects in the current observation and pay more attention to them, as shown in the activation visualization in the last column. Interestingly, we found that even though the agent is far away from the goal position, the mid fusion mechanism still guided the observation encoder to focus on relevant objects (see the *wooden cabinet* in the third row) or explore some candidate regions that may contain the target objects (see the *kitchen bar* in the last row). We also provide visualization and analysis of the other two goal prompting methods in Appendix.

Performance versus model size. To discuss the feasibility of application on real-world robot systems with resource-limited devices (*e.g.*, household robots), we investigate and compare the model size of our models with previous ones. We report the agent's number of parameters, as well as the ImageNav success rate on Gibson, and visualize them on the same coordinate system. As shown in Figure 1b, our FGPrompt-EF model outperforms existing methods by a large margin with a much smaller model size, indicating its promising ability on applying to real-world robot systems.

5 Discussion

Limitation and future work Although our proposed FGPrompt achieved great improvements on different ImageNav datasets, we still need a comprehensive study to find out if this method is applicable to real-world robots. In the future, we will investigate how to deploy this visual navigation methodology to a real-world robot system, to perform sim-to-real transformation.

Conclusion In this paper, we propose a novel fine-grained and high-resolution conditioned embedding method for visual navigation. In particular, we design a Mid Fusion architecture via FiLM Layers conditioning (FGPrompt-MF), which leverages the high-resolution activation maps from the goal encoder to perform an affine transformation on the observation encoder. Furthermore, we rethink it and condense it into an Early Fusion mechanism via joint modeling (FGPrompt-EF), with implicit learning of the fusion process. Experimental results on the Image-goal Navigation task show our method has excellent performance, concise architecture design, and strong generalization ability to unseen environments.

References

[1] P. Anderson, A. X. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva, and A. R. Zamir. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018. 6

[2] G. Avraham, Y. Zuo, T. Dharmasiri, and T. Drummond. Empnet: Neural localisation and mapping using embedded memory points. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 3

[3] E. Beeching, J. Dibangoye, O. Simonin, and C. Wolf. Learning to plan with uncertain topological maps. In *The European Conference on Computer Vision (ECCV)*, pages 473–490, 2020. 3

[4] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. 3

[5] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 7

[6] D. S. Chaplot, D. Gandhi, S. Gupta, A. Gupta, and R. Salakhutdinov. Learning to explore using active neural SLAM. In *International Conference on Learning Representations (ICLR)*, 2020. 1, 3, 6

[7] D. S. Chaplot, E. Parisotto, and R. Salakhutdinov. Active neural localization. In *International Conference on Learning Representations (ICLR)*, 2018. 1, 3

[8] D. S. Chaplot, R. Salakhutdinov, A. Gupta, and S. Gupta. Neural topological SLAM for visual navigation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 3, 6

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005. 4

[10] D. DeTone, T. Malisiewicz, and A. Rabinovich. Superpoint: Self-supervised interest point detection and description. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 4

[11] Y. Du, C. Gan, and P. Isola. Curious representation learning for embodied intelligence. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2, 3

[12] H. F. Durrant-Whyte and T. Bailey. Simultaneous localization and mapping: part I. *IEEE Robotics and Automation Magazine (RAM)*, 2006. 3

[13] K. Fang, A. Toshev, L. Fei-Fei, and S. Savarese. Scene memory transformer for embodied agents in long-horizon tasks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 538–547, 2019. 3

[14] S. Gupta, J. Davidson, S. Levine, R. Sukthankar, and J. Malik. Cognitive mapping and planning for visual navigation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[15] M. Hahn, D. S. Chaplot, S. Tulsiani, M. Mukadam, J. M. Rehg, and A. Gupta. No rl, no simulation: Learning to navigate without navigating. In *Neural Information Processing Systems (NeurIPS)*, 2021. 1, 3, 5, 6

[16] M. A. Islam, M. Kowal, P. Esser, S. Jia, B. Ommer, K. G. Derpanis, and N. Bruce. Shape or texture: Understanding discriminative features in cnns. In *ICLR*, 2022. 2, 3

[17] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn. BC-Z: zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning (CoRL)*, 2021. 3, 4

[18] N. Kim, O. Kwon, H. Yoo, Y. Choi, J. Park, and S. Oh. Topological semantic graph memory for image-goal navigation. In *Conference on Robot Learning (CoRL)*, pages 393–402. PMLR, 2023. 3, 6

[19] J. Krantz, T. Gervet, K. Yadav, A. Wang, C. Paxton, R. Mottaghi, D. Batra, J. Malik, S. Lee, and D. S. Chaplot. Navigating to objects specified by images. *arXiv preprint arXiv:2304.01192*, 2023. 1, 3

[20] O. Kwon, N. Kim, Y. Choi, H. Yoo, J. Park, and S. Oh. Visual graph memory with unsupervised representation for visual navigation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 15890–15899, 2021. 3, 6

[21] D. G. Lowe. Object recognition from local scale-invariant features. In *IEEE International Conference on Computer Vision (ICCV)*, 1999. 4

[22] A. Majumdar, G. Aggarwal, B. Devnani, J. Hoffman, and D. Batra. Zson: Zero-shot object-goal navigation using multimodal goal embeddings. In *Neural Information Processing Systems (NeurIPS)*, 2022. 2, 3, 5, 6

[23] L. Mezghani, S. Sukhbaatar, T. Lavril, O. Maksymets, D. Batra, P. Bojanowski, and K. Alahari. Memory-augmented reinforcement learning for image-goal navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022. 2, 3, 5, 6, 9

[24] M. B. Muhammad and M. Yeasin. Eigen-cam: Class activation map using principal components. In *International Joint Conference on Neural Networks (IJCNN)*, 2020. 9

[25] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics (T-RO)*, 2015. 1

[26] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 2, 3, 4, 8

[27] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. Turner, E. Undersander, W. Galuba, A. Westbury, A. X. Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021. 7

[28] R. Ramrakhya, E. Undersander, D. Batra, and A. Das. Habitat-web: Learning embodied object-search strategies from human demonstrations at scale. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[29] P. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich. Superglue: Learning feature matching with graph neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4

[30] N. Savinov, A. Dosovitskiy, and V. Koltun. Semi-parametric topological memory for navigation. In *International Conference on Learning Representations (ICLR)*, 2018. 1, 3, 6

[31] M. Savva, J. Malik, D. Parikh, D. Batra, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, and V. Koltun. Habitat: A platform for embodied AI research. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 5

[32] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 5

[33] A. Stone, T. Xiao, Y. Lu, K. Gopalakrishnan, K.-H. Lee, Q. Vuong, P. Wohlhart, B. Zitkovich, F. Xia, C. Finn, et al. Open-world object manipulation using pre-trained vision-language models. *arXiv preprint arXiv:2303.00905*, 2023. 3

[34] A. Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. S. Chaplot, O. Maksymets, A. Gokaslan, V. Vondrus, S. Dharur, F. Meier, W. Galuba, A. X. Chang, Z. Kira, V. Koltun, J. Malik, M. Savva, and D. Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *Neural Information Processing Systems (NeurIPS)*, 2021. 5

[35] S. Thrun. Probabilistic robotics. *Commun. ACM*, 45(3):52–57, 2002. 3

[36] K. Yadav, A. Majumdar, R. Ramrakhya, N. Yokoyama, A. Baevski, Z. Kira, O. Maksymets, and D. Batra. Ovrl-v2: A simple state-of-art baseline for imagenav and objectnav. *arXiv preprint arXiv:2303.07798*, 2023. 2, 3, 6

[37] K. Yadav, R. Ramrakhya, A. Majumdar, V.-P. Berges, S. Kuhar, D. Batra, A. Baevski, and O. Maksymets. Offline visual representation learning for embodied navigation. In *International Conference on Learning Representations (ICLR)*, 2022. 2, 3, 5, 6

[38] B. Yamauchi. A frontier-based approach for autonomous exploration. In *IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA)*, 1997. 3

[39] A. Yu and R. J. Mooney. Using both demonstrations and language instructions to efficiently learn robotic tasks. *arXiv preprint arXiv:2210.04476*, 2022. 3, 4

[40] B. Zhou, D. Bau, A. Oliva, and A. Torralba. Interpreting deep visual representations via network dissection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019. 2, 3

[41] Y. Zhu, R. Mottaghi, E. Kolve, J. J. Lim, A. Gupta, L. Fei-Fei, and A. Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2017. 1

[42] S. K. R. Ziad Al-Halah and K. Grauman. Zero experience required: Plug & play modular transfer learning for semantic visual navigation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 5, 6, 7