ViNT: A Large-Scale, Multi-Task Visual Navigation Backbone with Cross-Robot Generalization

Anonymous Author(s)
Affiliation
Address
email

Abstract: We present ViNT: a foundation model for visual robot navigation problems. Using a very general navigation pre-training objective, we train a large Transformer model to act as a generalist policy on a wide array of existing datasets containing hundreds of hours of cumulative robot navigation data and including a diverse set of environments and robot embodiments. ViNT demonstrates strong zero-shot transfer capabilities to new environments and even unseen robots. It also exhibits *positive transfer*, outperforming even specialist models trained on narrower datasets. We equip the model with a novel diffusion-based exploration mechanism, and demonstrate that it is capable of navigation and coverage tasks in large unseen environments. When equipped with a long-range goal-directed heuristic, ViNT can be used to achieve kilometer-scale navigation on any robot. ViNT's zero-shot capabilities can be further enhanced by fine-tuning the model for increased performance on specific downstream tasks, robots, or environments with only a small amount of data. Fine-tuning can also be used to adapt the model to new methods of task specification other than the visual goals used in the pre-training objective, greatly expanding the model's capability with minimal additional effort. These properties establish ViNT as an effective foundation model for mobile robotics.

Keywords: visual navigation, multi-task learning, planning

1 Introduction

Machine learning methods have achieved broad success in natural language processing [1], visual perception [2–4], and other domains [5, 6] by leveraging Internet-scale data to effectively generalize. Generalization typically requires learning from diverse datasets. These large-scale "foundation" models can be adapted to new tasks by zero-shot transfer, prompt-tuning, or fine-tuning on target data [7–10]. Although this paradigm has been successful in many domains, it is difficult to apply in robotics due to the sheer diversity of environments, platforms, and applications. In this paper we ask the question: *what is required of a broadly capable foundation model for mobile robotics?*

We specifically consider the problem of vision-based navigation. A general pre-trained robot navigation model should enable a wide range of navigation tasks, readily allow fine-tuning to downstream tasks, and generalize to a broad array of tasks, environments, and robotic platforms. Such a model should provide a broadly capable navigation policy on top of which applications to specific domains can be constructed, giving a base level of generalization and basic capabilities to new robotic platforms either in zero shot or with minimal fine-tuning.

To this end, we propose **Vi**sual **N**avigation **T**ransformer, or ViNT: an embodiment-agnostic foundation model for visual navigation with strong zero-shot generalization. We train ViNT to reach goals specified by camera images, providing a very general pre-training objective that can be applied to almost any mobile robot dataset. We propose a novel exploration algorithm for the visual navigation paradigm using a diffusion model to propose short-horizon goals, and demonstrate that it enables ViNT to navigate in novel environments. ViNT can be applied zero-shot to control new robots, explore previously unseen environments, perform indoor mapping, and navigate kilometer-scale outdoor environments without interventions. Furthermore, we show that the base ViNT model

Submitted to the 7th Conference on Robot Learning (CoRL 2023). Do not distribute.

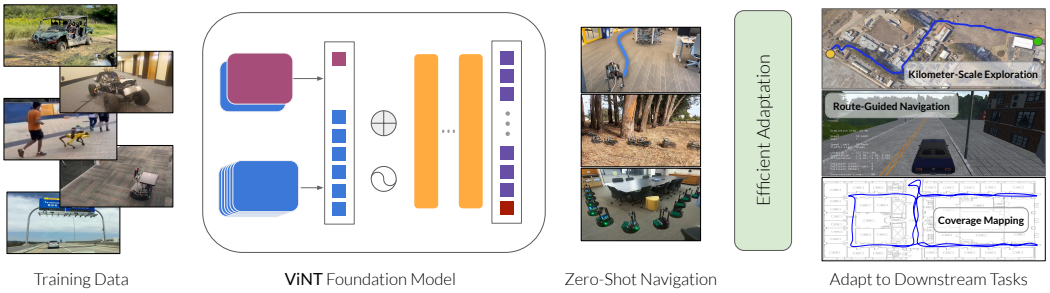

Figure 1: Overview of the ViNT foundation model. ViNT generalizes zero-shot across environments and robot embodiments, and can be directly applied to tasks including exploration and short-horizon navigation. ViNT can be fine-tuned with a small amount of data to further improve performance in out-of-distribution target environments or to expand its capabilities to new tasks.

can be fine-tuned on a small amount of data to achieve high performance with new task specification modalities – such as GPS waypoints or high-level routing commands – allowing ViNT to serve as a foundation for a wide variety of navigation applications.

We hope that this represents a step towards general-purpose visual navigation models that can be deployed on a wide range of robots, and on a wide range of tasks, and serve as a foundation for diverse mobile robotic applications. Model weights for ViNT as well as training and deployment code are available on our project page.

2 Related Work

Learning from large, diverse robotic datasets has been studied for various robotic applications where data sharing across *similar* robots provides a larger training set for more generalizable models [11–13]. However, for applications in mobile robotics, with varying dynamics and camera configurations (e.g., focal length, field of view, and extrinsics), current approaches tend to rely on learning either from small real-world datasets, which are only representative of a single robotic platform, or from simulation, with paired robot and environment models to transfer learned policies [14–16]. Instead, our paper follows the paradigm of learning navigation behavior from data collected across multiple different real-world robotic systems [17, 18], while focusing on learning a foundation model that can be adapted for a variety of downstream tasks in zero shot or with small amounts of data.

Our goal is to train an effective visual navigation policy that can solve a range of downstream tasks, such as navigating to GPS goals [19], goal images [20], skill-conditioned driving [21]. Following a large body of research in visual navigation, we use a combination of topological graphs for maintaining a spatial representation of the environment and learned policies for low-level control [22–25], and use learned heuristics to guide the robot in novel environments [15, 26]. But unlike these works, our goal is to train a single generalist model rather than specialist solutions to each of these problems, showing how a single high-capacity model can be adapted for diverse tasks.

The closest related works to ViNT are I2O, GNM, and RT-1 [15, 18, 27], which study broad generalization across environments and embodiments for robots deployed in real-world settings. While RT-1 demonstrates impressive performance in following diverse instructions, our focus is on adapting a single model across *many* robots to solve *different tasks*, by fine-tuning with small amounts of data. I2O and related efforts [15, 16] show impressive transfer from simulation to real-world environments, but we emphasize that our aim is orthogonal to the specific choice of algorithm: we focus on learning a capable navigation policy that can be efficiently adapted to solve different downstream tasks. [18] demonstrate policy learning from heterogeneous RGB datasets, but focuses on the singular task of reaching image goals. Instead, ViNT learns a single generalist policy that can be adapted to a variety of downstream tasks with small amounts of data.

3 The ViNT Model

Our model is trained for image-goal navigation, providing general navigational capabilities that can then either be utilized directly, or serve as a pre-trained foundation for downstream fine-tuning with other task specifications. In the image-goal navigation task, the robot is tasked with navigating to

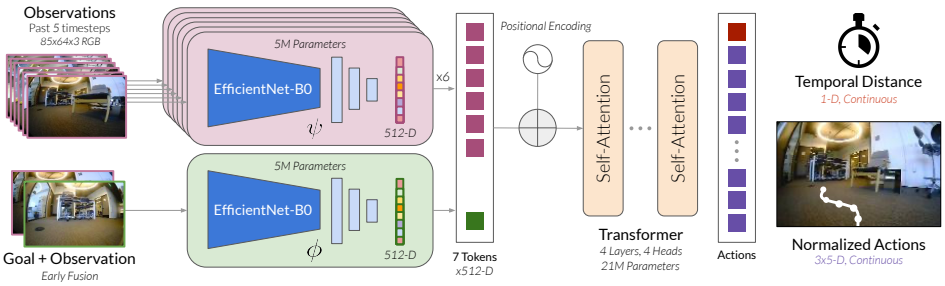

Figure 2: ViNT Model Architecture. ViNT uses two EfficientNet encoders ψ, ϕ to generate input tokens to a Transformer decoder. The resulting sequence is concatenated and passed through a fully-connected network to predict (temporal) distance to the goal as well as a sequence of $H = 5$ future actions.

a subgoal specified by an image observation s (i.e., the robot's observation at the goal). Unlike alternative mechanisms for goal specification such as PointGoal [28], GPS navigation, or semantic objectives [29], a model can be trained for image-goal navigation with minimal assumptions, utilizing any data that contains videos and actions, without requirements on ground-truth localization, semantic labels, or other metadata. This makes it practical to train on a large and diverse dataset sourced from many different robots, facilitating broad generalization.

ViNT takes as input current and past visual observations $o_{t-P:t}$ and a subgoal image s, and predicts (i) the number of time steps needed to reach the subgoal (the *dynamical* distance), and (ii) a sequence with length H of future actions leading towards the subgoal. Our 31M-parameter model, ViNT, is built on the Transformer architecture [30] and is optimized for: (i) fast and efficient inference on resource-constrained robots, and (ii) the ability to prompt and fine-tune for downstream tasks. We initialize all networks from scratch and train them end-to-end with the training objective in Eqn. 1. The model architecture is summarized in Figure 2, and described in detail in Appendix A.

Tokenization: The ViNT architecture (Fig. 2) first tokenizes its inputs into an embedding of size $d_{\text{model}} = 512$. ViNT independently tokenizes current and $P = 5$ past visual observations by encoding them with an EfficientNet-B0 [31] model, which takes $85 \times 64 \times 3$ images as input and outputs a flattened feature vector $\psi(o_i)$ from the final convolutional layer [27].

Goal fusion: We found that naïvely extracting features from the goal image $\phi(s)$ using an EfficientNet encoder ϕ led to poor performance, often ignoring the goal entirely (see Appendix A). We hypothesize that effective features for image-based goal-reaching tasks are often *relative*, encoding the *difference* between the current observation and the goal rather than an absolute representation of the goal itself. Hence, we use a separate *goal fusion* encoder $\phi(o_t, s)$ to jointly encode the current and goal observations. We stack the two images along their channel dimensions, pass them through a second EfficientNet-B0 encoder, and flatten to obtain the goal token.

Transformer: The $P + 2$ observation and goal tokens are combined with a positional encoding and fed into a Transformer backbone. We use a decoder-only transformer with $n_L = 4$ multi-headed attention blocks, each with $n_H = 4$ heads and $d_{\text{FF}} = 2048$ hidden units.

Training objective: During training, we first sample a minibatch of trajectories τ from the dataset $\mathcal{D}$. We then choose P consecutive observations to form the temporal context $o_{t:t-P}$ and randomly select a future observation o_{t+d}, with d sampled uniformly from $[l_{\text{min}}, l_{\text{max}}]$, to serve as the subgoal [32]. The corresponding H future actions $a_{t:t+H}$ and the distance d are used as labels and trained with a maximum likelihood objective:

$$\mathcal{L}_{\text{ViNT}} = \mathbb{E}_\tau \mathbb{E}_t \mathbb{E}_d \left[\log p(a_{t:t+H} | f(\psi(o)_{t:t-P}, \phi(o_t, s)) + \lambda \log p(d | f(\psi(o)_{t:t-P}, \phi(o_t, s)) \right], \quad (1)$$

where ϕ, ψ, f are as defined above, and λ balances the two losses.

Embodiment-agnostic action space: To effectively train a single model across robots of varying sizes, speeds, and dynamics, we follow Shah et al. [18] and choose an embodiment-agnostic action space for ViNT. To abstract away low-level control, ViNT uses relative waypoints as its action space; to account for the large variance in speeds and sizes of the robots, we normalize these waypoints by scaling them according to the robot's top speed.

Training data: We train ViNT using a large-scale dataset of heterogeneous navigation trajectories from a diverse set of environments and robotic platforms with varying dynamics, camera parameters, and behaviors. The training dataset contains over 100 hours of real-world trajectories sourced

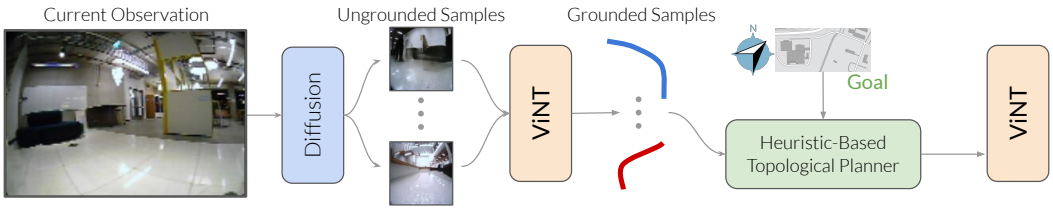

Figure 3: Long-horizon navigation in unseen environments with ViNT. We use physical search with a topological graph-based planner to explore the environment. An image-to-image diffusion model proposes diverse exploration targets which are spatially grounded using ViNT, and scored using a goal-directed heuristic h. The global planner then selects the best subgoal towards the goal, and drives towards it using ViNT.

entirely from existing datasets, spanning 8 distinct robotic platforms with varying speeds and dynamics. For more details about the dataset, see Appendix C.

Deployment: ViNT can run on any robot equipped with an onboard camera and a low-level velocity tracking controller. Given a subgoal image s at time t, we run the model at 4Hz, and use a PD controller to track the predicted waypoints $a_{t:t+H}$ in a receding-horizon fashion.

4 Long-Horizon Navigation with ViNT

While the goal-conditioned policy learned by ViNT captures a general understanding of navigational affordances and obstacles, it has limited applicability on its own. Many practical tasks are either not defined by goal images, or require a much longer horizon than what ViNT supports. We apply ViNT to several downstream applications by combining it with an episodic memory formed by a topological graph, which can provide short-horizon subgoals for reaching faraway locations. In previously unseen environments, we can further augment this graph-based planner with a mechanism for proposing novel subgoals to ViNT, which can then be used to reach the goal. To do this, we train an *image diffusion model* that generates diverse future subgoal candidates conditioned on the current observation, and score them with a goal-directed heuristic to identify the best subgoal that makes progress towards the goal using a process akin to *physical search* [26]. Past observations in the environment and unexplored frontiers are stored as nodes in a topological graph, with their connectivity signifying paths taken in the environment. During exploration, we build this topological graph on-the-fly; during future deployments, this graph may be used for discovering short-cuts to arbitrary goals in the environment.

4.1 Subgoal Generation with Diffusion

To generate diverse subgoal proposals $\mathcal{S}$ that are reachable from the current state of the robot o_t, we train a conditional generative model $g(s_i|o_t)$ on a large amount of training data. Specifically, we use an image-to-image diffusion model [33, 34], which is particularly well-suited for generating diverse samples from multimodal distributions over high-dimensional spaces. We train a 318M parameter diffusion model using randomly-sampled future observations from trajectories in the ViNT dataset, and sample K subgoal candidates $\mathcal{S} = \{s_1, \ldots, s_K\}$ from the model at inference time.

However, these subgoal generations are not *spatially grounded*: they do not include an actionable relationship to o_t. We ground these candidates by using ViNT to compute temporal distances $d(s_i, o_t)$ and action rollouts $a(s_i, o_t)$, yielding a set of grounded subgoals as in Fig. 5. While the samples generated by the diffusion model do not necessarily match any true observation (see Figure 3), they preserve sufficient relative features from o_t to be plausible, and we find that ViNT generalizes well to generated subgoals. Please see Appendix B.2 for more details about the diffusion model architecture, training procedure, and samples from the model.

4.2 High-Level Planning and Physical Search

We incorporate these subgoal candidates into a planning framework for goal-directed exploration in novel environments, where the user provides a high-level goal G, which may be abstract (e.g., map the entire floor), or arbitrarily far away. We largely follow prior work [26], but swap out their

learned models with ViNT and the diffusion model. We summarize the system here, and provide a more complete discussion in Appendix B.3.

We construct a topological graph $\mathcal{M}$ online to act as episodic memory, with each node as an individual subgoal observation and with edges representing paths between two subgoals, added when the path is taken by the robot. We frame goal-directed exploration as a *physical search* problem, where the robot incrementally builds $\mathcal{M}$ while searching for the goal. To guide search towards the goal, the robot uses a goal-directed heuristic $h(o_t, s_i, G, \mathcal{M}, C)$ to *score* subgoal candidates according to their likelihood of reaching the goal, given additional context C — for example, a floor plan or satellite imagery [15, 26]. This heuristic may be geometric (e.g. Euclidean distance) or learned.

During deployment in a new environment, the robot uses the diffusion model to generate subgoal candidates $\mathcal{S}$ from o_t, spatially grounds them using ViNT, and scores them using the goal-directed heuristic $h(.)$. The robot then selects the best subgoal s^* according to this heuristic using an A*-like planner, adds it to $\mathcal{M}$, and drives towards it using ViNT (Figure 3). During subsequent deployments in the same environment, the robot can use $\mathcal{M}$ to discover shortcuts to arbitrary goals in the environment. Please see Appendix B.3 for more details about the planner and heuristics.

5 ViNT: A Foundation Model For Downstream Tasks

In addition to direct use, ViNT learns strong navigational priors that can be adapted to many downstream tasks by fine-tuning part or all of the model in novel environments, with new modalities of data, or with different objectives.

Full model fine-tuning: While ViNT demonstrates strong zero-shot generalization to new environments and robots, we can further improve on-task performance by fine-tuning the entire model with the same objective but using on-task data. ViNT can master new environments and embodiments with as little as 1.5 hours of data, transferring the capabilities of the original model to a new setting without retraining from scratch.

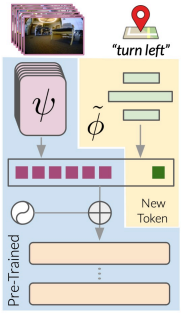

Figure 4: Adapting ViNT to different modalities.

Adapting to new modalities: While specifying image goals gives us a general pre-training objective, ViNT can easily be *adapted* to other common forms of goal-specification by learning a "soft prompt" mapping from the desired goal modality to the ViNT goal token [10]. We build on the Transformer architecture's ability to attend to multimodal inputs projected into a shared token space [35, 36]. Given a subgoal σ in a new modality (such as 2D coordinates or high-level routing directions [21]), we train a small neural network $\tilde{\phi}$ that maps the subgoal to this shared token space as shown in Figure 4, and replace $\phi(o_t, s)$. We fine-tune ViNT using a small amount of on-task data $\mathcal{D}_F$ using the modified objective:

$$\mathcal{L}_{\text{adapt}} = \mathbb{E}_{\tau \in \mathcal{D}_F} \mathbb{E}_t \mathbb{E}_d \left[\log p(a_{t:t+H} | f(\psi(o)_{t:t-P}, \tilde{\phi}(\sigma)) + \lambda \log p(d | f(\psi(o)_{t:t-P}, \tilde{\phi}(\sigma))) \right]. \quad (2)$$

This allows adaptation to new tasks with minimal data, while still leveraging the performance and generalization of ViNT. Appendix B.4 includes additional details.

6 Real-world Evaluation

We deployed our ViNT foundation model on five distinct robotic platforms, including a drone, a quadruped, and two other *novel* robots with no corresponding training data. We designed our experiments to answer the following questions:

- **Q1.** Can ViNT efficiently explore previously unseen environments and incorporate *hints*?
- **Q2.** Does ViNT generalize to *novel* robots, environments, and obstacles?
- **Q3.** Can ViNT be fine-tuned to improve performance in out-of-distribution environments?
- **Q4.** Can the ViNT policy be adapted to handle new task specification and modalities?

Please see Appendix D for more details on platforms used in the training data and in evaluation.

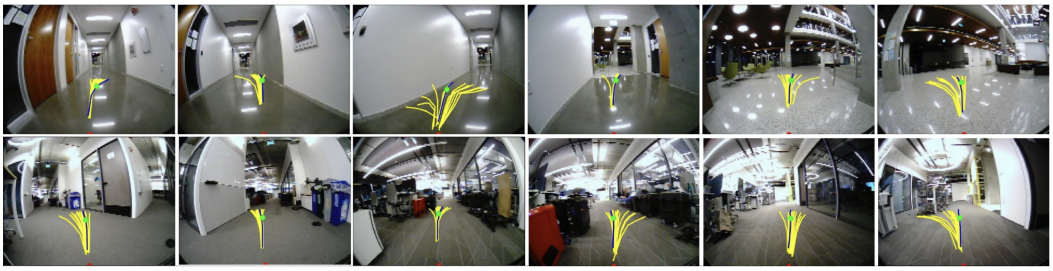

Figure 5: Visualizing rollouts of two different robots — Vizbot (top) and LoCoBot (bottom) — in challenging indoor environments. Future action samples $a_{t:t+H}$ obtained by *spatially grounding* the diffusion samples.

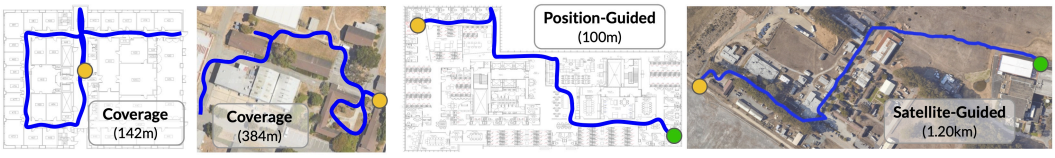

Figure 6: ViNT accomplishes long-horizon navigation with a variety of objectives in indoor and outdoor environments. Goal-reaching behavior can be achieved with a goal-directed heuristic (optionally guided by satellite imagery), while removing this heuristic allows for undirected exploration to maximally cover a workspace.

6.1 Core Results

Towards understanding **Q1**, we deploy our full graph-based navigation pipeline (Section 4.2) in a variety of challenging indoor and outdoor environments. We evaluate the performance of ViNT on two challenging tasks in *previously unseen environments*: (i) coverage exploration, where the objective is to visit as many novel locations as possible, and (ii) guided exploration, where the objective is to reach a goal location using contextual information such as GPS coordinates or a satellite image. We compare ViNT to a variety of baselines, including end-to-end policies trained with imitation or RL [15, 37], and a prior graph-based approach using a VIB for exploration [38].

Figure 6 shows example trajectories taken by ViNT in challenging indoor and outdoor environments for the different tasks. For the base task of coverage (Table 2), we find that while end-to-end baselines can avoid collisions with their surroundings, they cannot optimize for coverage at a global level, and often get stuck in small areas of the environment. Graph-based methods avoid this by explicitly reasoning about coverage with the search objective, and we find that ViNT paired with a diffusion model generates more diverse subgoal predictions than the VIB baseline, leading to better exploration. We hypothesize that the diffusion model can better capture the multimodal distribution of future subgoals, as compared to a VIB with a unimodal prior, and hence perform consistently in both indoor and outdoor environments.

This observation extends to the long-range navigation task (Table 3), where we find that the diverse subgoal predictions from the diffusion model allow ViNT to make better use of context in selecting goals. ViNT effectively solves the long-horizon navigation problem in all outdoor test cases, while the VIB-based sampler only succeeds 24% of the time. Our approach is also significantly more effective at avoiding collisions in crowded indoor spaces and requires fewer interventions. Please see Appendix E.1 for details of the experimental setup, visualizations, and videos from the experiments.

Model	LoCoBot	Go 1	Vizbot	Jackal
In-Domain	40	12	40	184
GNM [18]	60	8	20	**427**
ViNT	**120**	**45**	**110**	**438**

Table 1: The same ViNT policy can control many robots, driving hundreds of meters without collisions and beating lower-capacity models (GNM) and models trained on a single dataset in the target domain.

6.2 Zero-Shot Generalization: a Single Policy to Drive Any Robot

Towards answering **Q2**, we deploy the *same* pre-trained ViNT policy on four distinct robotic platforms *without* any fine-tuning for the task of coverage exploration, as described above. Table 1 summarize our evaluation for the task of coverage exploration in a variety of indoor and outdoor en-

Method	Indoor		Outdoor	
	Survival Time	Distance Covered	Survival Time	Distance Covered
End-to-End BC	7:42	47.0	8:11	217.0
End-to-End GCG [37]	—	—	11.41	357.0
VIB [38]	3:23	24.3	1:04	58.3
Diffusion (Ours)	**9:40**	**91.0**	**20:00+**	**1007+**

Table 2: ViNT with Diffusion can consistently outperform all baselines and explore large areas without interventions. Indoor experiments time out at 10:00, and outdoor at 20:00 minutes. GCG is not run indoors due to lack of suitable labeled data.

Method	Indoor		Outdoor	
	Progress %	Distance	Progress %	Distance
ViNT + VIB [26]	60%	56m	24%	107
ViNT + Diffusion	**90%**	**91m**	**100%**	**1170**

Table 3: Evaluating goal-directed exploration on large-scale goal-reaching problems.

vironments, reporting the total distance traversed before collision. Most notably, ViNT successfully generalizes zero-shot to control a Go 1 quadruped, which does not appear in the training set.

We compare ViNT trained across all the combined datasets and robots to the best single-robot baseline, a model trained using data only from the target environment, as well as the GNM model [18] trained on all datasets. We observe that policies trained across robot embodiments can not only match, but also *outperform*, single-robot models across all the embodiments we studied. We also find that the larger capacity of ViNT leads to improved generalization compared to GNM on out-of-distribution robots such as the Go 1 which do not appear in the dataset. Crucially, we find that ViNT demonstrates *positive transfer* for in-domain robots (Vizbot), greatly outperforming a model trained on only the target environment, while smaller models such as GNM fail to do so. This indicates that the model *generalizes* between multiple tasks to improve performance, a key property of foundation models.

6.3 Broader Generalization via Fine-Tuning

To answer **Q3**, we consider the problem of fine-tuning ViNT in the low data regime. We study fine-tuning for the task of autonomous driving in the CARLA simulator for two reasons: (i) CARLA has low-fidelity graphics and is perceptually very different from the real-world data used to train ViNT (Fig. 7), and (ii) the task of on-road driving requires very specific semantic behavior, i.e., driving in a lane and making smooth turns, that is not present in the training data. We show that ViNT can be fine-tuned on a small amount of data to achieve strong performance on this task by effectively leveraging the navigational prior encoded in the model.

Figure 7: Third-person view of the CARLA test environment (top), and a bird's eye view showing the high-level routing commands for the routing task.

We compare the pre-trained ViNT backbone to several alternatives, including visual representations trained with supervised learning [2], unsupervised objectives [7, 39, 40], and an embodiment-agnostic navigation policy [18]. We use the same fine-tuning data and procedure for all models (see Section 5); please see Appendix E.3 for more details.

Table 4 summarizes our findings. While pre-trained visual representations significantly improve task performance over a policy trained entirely from scratch, we observe that the learned policies suffer from frequent collisions and poor performance. GNM [18] outperforms these baselines due to its strong navigation prior, but the lower-capacity model is unable to generalize fully to the task. ViNT, on the other hand, is able to achieve strong performance, achieving substantially higher success rate than the next best baseline. Sweeping over fine-tuning dataset size (Table 4-right) shows that ViNT achieves strong performance with as little as 20 minutes of fine-tuning data, demonstrating its ability to generalize to new environments with very little data.

Method	Images		Positions	Routing
	Success	In Lane	Success	Success
Scratch	0.45	0.74	0.79	0.43
ImageNet	0.22	0.71	0.59	0.45
SimCLR [7]	0.21	0.63	0.70	0.64
VC-1 [40]	0.19	0.65	0.49	0.38
GNM [18]	0.49	0.66	0.45	0.49
ViNT	0.82	0.82	0.89	0.72

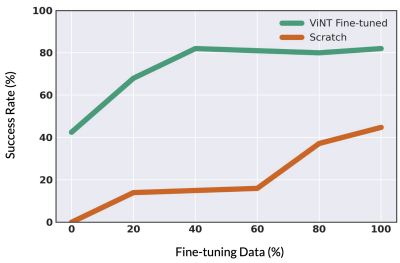

Table 4: Left: fine-tuning and task adaptation results across multiple baselines. **Right:** fine-tuning results of ViNT compared to training from-scratch across different data scales.

6.4 Adapting ViNT to Downstream Tasks

To evaluate **Q4**, we investigate whether ViNT can serve as a foundation model for a broader range of downstream tasks by considering goal modalities beyond subgoal images (see Section 6.4). We consider the same CARLA driving task but with two different high-level planners: (i) a position-based planner that commands a sequence of GPS waypoints, and (ii) a routing planner with similar functionality to Google Maps that commands high-level navigation directions (left/right/straight) to the policy [21]. We compare the pre-trained navigational priors learned by ViNT to the baselines discussed earlier, corresponding to pre-trained visual representations and policies, each adapted to the downstream task using the same on-task data (see Appendix E.3 for more details).

Table 4 summarizes our results for the two tasks. We again find that the general pre-trained visual representations such as ImageNet or VC-1 are not sufficient to extract navigational affordances for challenging downstream tasks, suggesting that effective generalization requires more than general visual representations [40, 41]. We also find that unlike fine-tuning, GNM [18] struggles with the adaptation task, suggesting that the architecture and increased capacity of ViNT are essential for broad generalization and adaptation. ViNT achieves strong performance on both tasks, demonstrating its ability to serve as a foundation model for downstream tasks.

7 Discussion and Future Work

We presented ViNT, a navigational foundation model that is trained for a generic image-goal navigation task on diverse data from many different robots, and can then support a variety of different navigational functionalities. ViNT can be deployed for long-horizon navigation in combination with a topological graph planning method, it can explore new environments with goals proposed by a diffusion model, and it can be fine-tuned to new domains, such as autonomous driving, and new task specification methods, such as GPS coordinates or even turn-by-turn instructions. Our results show that ViNT can successfully generalize across robots and environments, outperforms prior navigational models, and can be efficiently fine-tuned to new domains and tasks.

Limitations and future work: As with many large-scale models, ViNT carries a heavier computational burden at inference-time, which can present a challenge for power-constrained platforms such as quadcopters. While our design aims to enable efficient inference, our Transformer-based model is still significantly costlier to run at deployment time than simpler feedforward convolutional networks. Additionally, although ViNT generalizes effectively across robots in our experiments, it does assume a degree of structural similarity. For example, it cannot control the altitude of a quadcopter or handle other changes in the action representation, nor accommodate new sensors such as LIDAR. Training on a variety of modalities and action spaces in future work could unlock this capability. More broadly, while ViNT illustrates the promise of a general-purpose and broadly reusable navigational foundation model, we believe that the most exciting developments for general-purpose cross-robot models are still ahead: as larger and larger multi-robot datasets are assembled, perhaps we will see even broader generalization and more flexible specification with increasingly powerful and general-purpose robotic models. We hope that ViNT represents a step in this direction.

References

[1] A. Radford and K. Narasimhan. Improving language understanding by generative pre-training, 2018. 1

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 1, 7

[3] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[4] K. Grauman et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[5] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents, 2022. 1

[6] M. Chen et al. Evaluating large language models trained on code. *arXiv*, 2021. 1

[7] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. *International Conference on Machine Learning (ICML)*, 2020. 1, 7, 8

[8] D. Shah, B. Osiński, B. Ichter, and S. Levine. Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on Robot Learning (CoRL*, 2022.

[9] X. Liu, Y. Li, C. Liang, and X. Li. The power of scale for parameter-efficient prompt tuning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[10] J. Wei et al. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations (ICLR)*, 2022. 1, 5

[11] C. Devin, A. Gupta, T. Darrell, P. Abbeel, and S. Levine. Learning modular neural network policies for multi-task and multi-robot transfer. In *2017 International Conference on Robotics and Automation (ICRA)*, pages 2169–2176, 2017. doi:10.1109/ICRA.2017.7989250. 2

[12] S. Dasari, F. Ebert, et al. Robonet: Large-scale multi-robot learning. In *Conference on Robot Learning (CoRL)*, 2020.

[13] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 5

[14] P. Anderson, A. Shrivastava, J. Truong, A. Majumdar, D. Parikh, D. Batra, and S. Lee. Sim-to-real transfer for vision-and-language navigation, 2020. 2

[15] J. Truong, A. Zitkovich, S. Chernova, D. Batra, T. Zhang, J. Tan, and W. Yu. Indoorsim-to-outdoorreal: Learning to navigate outdoors without any outdoor experience, 2023. 2, 5, 6

[16] A. Kadian, J. Truong, A. Gokaslan, A. Clegg, E. Wijmans, S. Lee, M. Savva, S. Chernova, and D. Batra. Sim2Real Predictivity: Does Evaluation in Simulation Predict Real-World Performance? *IEEE Robotics and Automation Letters*, 2020. 2

[17] A. Loquercio, A. I. Maqueda, C. R. del Blanco, and D. Scaramuzza. Dronet: Learning to fly by driving. *IEEE Robotics and Automation Letters*, 2018. 2

[18] D. Shah, A. Sridhar, A. Bhorkar, N. Hirose, and S. Levine. GNM: A General Navigation Model to Drive Any Robot. In *International Conference on Robotics and Automation (ICRA)*, 2023. URL `https://arxiv.org/abs/2210.03370`. 2, 3, 6, 7, 8

[19] A. Kadian, J. Truong, A. Gokaslan, A. Clegg, E. Wijmans, S. Lee, M. Savva, S. Chernova, and D. Batra. Sim2real predictivity: Does evaluation in simulation predict real-world performance? *IEEE Robotics and Automation Letters*, 5(4):6670–6677, oct 2020. doi:10.1109/lra.2020.3013848. URL `https://doi.org/10.1109%2Flra.2020.3013848`. 2

[20] 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 *International Conference on Robotics and Automation (ICRA)*, 2017. 2

[21] F. Codevilla, M. Müller, A. López, V. Koltun, and A. Dosovitskiy. End-to-End Driving Via Conditional Imitation Learning. In *International Conference on Robotics and Automation (ICRA)*, 2018. 2, 5, 8, 6

[22] N. Savinov, A. Dosovitskiy, and V. Koltun. Semi-Parametric Topological Memory for Navigation. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[23] J. Bruce, N. Sunderhauf, P. Mirowski, R. Hadsell, and M. Milford. Learning deployable navigation policies at kilometer scale from a single traversal. In A. Billard, A. Dragan, J. Peters, and J. Morimoto, editors, *2nd Conference on Robot Learning*, 2018.

[24] A. Faust, K. Oslund, O. Ramirez, A. Francis, L. Tapia, M. Fiser, and J. Davidson. Prm-rl: Long-range robotic navigation tasks by combining reinforcement learning and sampling-based planning. In *2018 International Conference on Robotics and Automation (ICRA)*, pages 5113–5120. IEEE, 2018.

[25] X. Meng, N. Ratliff, Y. Xiang, and D. Fox. Scaling Local Control to Large-Scale Topological Navigation. In *International Conference on Robotics and Automation (ICRA)*, 2020. 2

[26] D. Shah and S. Levine. ViKiNG: Vision-Based Kilometer-Scale Navigation with Geographic Hints. In *Robotics: Science and Systems (RSS)*, 2022. 2, 4, 5, 7, 3

[27] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, T. Jackson, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, K.-H. Lee, S. Levine, Y. Lu, U. Malla, D. Manjunath, I. Mordatch, O. Nachum, C. Parada, J. Peralta, E. Perez, K. Pertsch, J. Quiambao, K. Rao, M. Ryoo, G. Salazar, P. Sanketi, K. Sayed, J. Singh, S. Sontakke, A. Stone, C. Tan, H. Tran, V. Vanhoucke, S. Vega, Q. Vuong, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint*, 2022. 2, 3

[28] P. Anderson et al. On Evaluation of Embodied Navigation Agents, 2018. 3

[29] S. Wani, S. Patel, U. Jain, A. Chang, and M. Savva. Multion: Benchmarking semantic map memory using multi-object navigation. In *Neural Information Processing Systems (NeurIPS)*, 2020. 3

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2017. 3

[31] M. Tan and Q. V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020. 3, 1

[32] K. Hartikainen, X. Geng, T. Haarnoja, and S. Levine. Dynamical Distance Learning for Semi-Supervised and Unsupervised Skill Discovery. In *International Conference on Learning Representations (ICLR)*, 2020. 3

[33] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Neural Information Processing Systems*, 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf`. 4, 3

[34] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 4, 3

[35] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. Florence. Palm-e: An embodied multimodal language model. In *arXiv preprint arXiv:2303.03378*, 2023. 5

[36] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. Fan. Vima: General robot manipulation with multimodal prompts, 2023. 5

[37] G. Kahn, A. Villaflor, B. Ding, P. Abbeel, and S. Levine. Self-Supervised Deep RL with Generalized Computation Graphs for Robot Navigation. In *International Conference on Robotics and Automation (ICRA)*, 2018. 6, 7, 5

[38] D. Shah, B. Eysenbach, N. Rhinehart, and S. Levine. Rapid exploration for open-world navigation with latent goal models. In *Conference on Robot Learning (CoRL)*, 2021. 6, 7, 5

[39] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell. Real-world robot learning with masked visual pre-training. In *Conference on Robot Learning (CoRL)*, 2022. 7

[40] A. Majumdar, K. Yadav, S. Arnaud, Y. J. Ma, C. Chen, S. Silwal, A. Jain, V.-P. Berges, P. Abbeel, J. Malik, D. Batra, Y. Lin, O. Maksymets, A. Rajeswaran, and F. Meier. Where are we in the search for an artificial visual cortex for embodied intelligence?, 2023. 7, 8

[41] S. Parisi, A. Rajeswaran, S. Purushwalkam, and A. Gupta. The unsurprising effectiveness of pre-trained vision models for control, 2022. 8

[42] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[43] P. von Platen, S. Patil, A. Lozhkov, P. Cuenca, N. Lambert, K. Rasul, M. Davaadorj, and T. Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 3

[44] D. Kingma, T. Salimans, B. Poole, and J. Ho. Variational diffusion models. *Neural Information Processing Systems (NeurIPS)*, 2021. 3

[45] J. Ho and T. Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[46] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Neural Information Processing Systems (NeurIPS)*, 2022. 3

[47] A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding, 2019. 4

[48] N. Hirose, F. Xia, R. Martín-Martín, A. Sadeghian, and S. Savarese. Deep Visual MPC-Policy Learning for Navigation. *IEEE Robotics and Automation Letters*, 2019. 5

[49] H. Karnan et al. Socially CompliAnt Navigation Dataset (SCAND): A Large-Scale Dataset Of Demonstrations For Social Navigation. *IEEE Robotics and Automation Letters*, 2022. 5

[50] A. Shaban, X. Meng, J. Lee, B. Boots, and D. Fox. Semantic terrain classification for off-road autonomous driving. In *Conference on Robot Learning (CoRL)*, 2022. 5

[51] S. Triest et al. TartanDrive: A Large-Scale Dataset for Learning Off-Road Dynamics Models. In *International Conference on Robotics and Automation (ICRA)*, 2022. 5

[52] A. Agha et al. Nebula: Quest for robotic autonomy in challenging environments; team costar at the darpa subterranean challenge. *arXiv*, 2021. 5

[53] N. Hirose, D. Shah, A. Sridhar, and S. Levine. Sacson: Scalable autonomous data collection for social navigation. *arXiv*, 2023. 5

[54] T. Niwa, S. Taguchi, and N. Hirose. Spatio-temporal graph localization networks for image-based navigation. In *International Conference on Intelligent Robots and Systems (IROS)*, 2022. 5

[55] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 7