Appendix

A. Details on Data, Model, and Training

City Walking Videos. We source our training video data mainly from the city walking[1] and driving[2] playlists on YouTube. The full sourced videos have a total length of 2522 hours. We use 2000 hours of them for training. These videos cover different weather and lighting conditions. Figure I shows a detailed distribution of each condition.

The lower part of Fig. I illustrates the proportion of each critical scenario in our offline expert data based on our definitions. We observe that the union of critical scenarios accounts for less than half of the dataset. However, these scenarios contribute most to the success rate in real-world experiments. This highlights the need for future work to enhance model performance in these critical areas.

Hyperparameters for Model and Training. For model and training hyperparameters, we largely follow previous work [1] and adapt some parameters to our case, as shown in Tab. II. Note that DINOv2 [2] uses ViT [3] so it can adapt to any input resolution as long as it is divisible by the patch size. Therefore, we center-crop the 360×640 city walking videos to 350×630, and the 400×400 teleoperation video to 392×392 to keep the aspect ratio and as much visual content as possible.

B. More Quantitative Results

Full Ablation Study. In Tab. I, we provide an extended ablation study, including all the critical scenarios. We can observe that the addition of orientation loss and feature hallucination loss does not result in significant performance improvements. This lack of noticeable enhancement can be attributed to several factors, including the limited size of our training data (1000 hours) and the constrained nature of our test dataset, which is prone to substantial noise in the evaluation results. Consequently, we consider errors beyond the decimal point to be negligible.

Another interesting observation is the decline in performance within the Turn scenario following fine-tuning. We attribute this performance drop to the disproportionate representation of Turn scenarios in our fine-tuning data (8%) compared to the original video data (32%), leading to insufficient training examples for effectively handling turns.

VLM Performance. In Tab. III, we present the performance of the VLM (GPT-4o [4]) on our urban navigation tasks. Our results indicate that GPT-4o struggles to generate reasonable navigation actions off-the-shelf via prompting. However, it performs reasonably in predicting the arrival status, likely because this sub-task is inherently more straightforward given the input of past and target locations.

[1]https://www.youtube.com/@WALKS_and_the_CITY/playlists
[2]https://www.youtube.com/@jutah/playlists

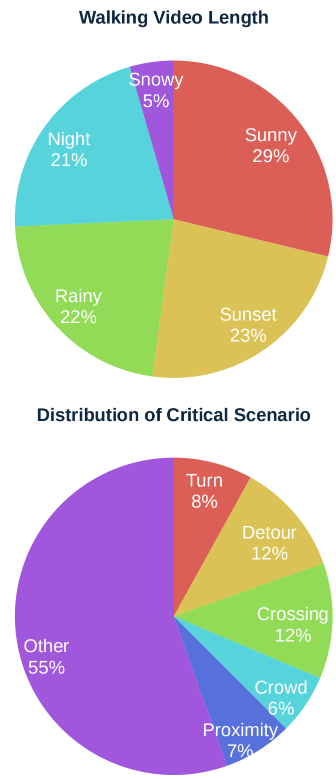

Figure I. **Data distribution**. *Top*: The distribution of different weather and lightning conditions in our video training data. *Bottom*: The distribution of different critical scenarios in our collected data.

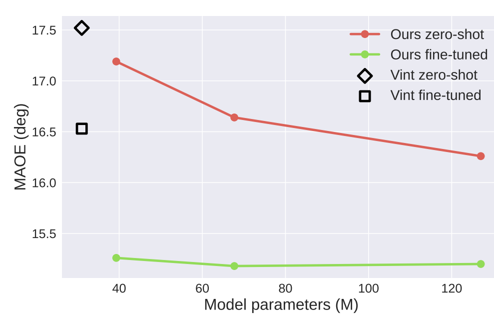

Figure II. **Performance and Model Szie**. We show model performance with respect to the size of the model, measured by the number of parameters in the model.

Impact of Model Size. We also run experiments to discover the impact of model size on navigation performance. This is done by modifying the number of layers in the transformer model. From Fig. II, we can observe the clear trend that a larger model with more parameters leads to better performance, especially in the zero-shot case. Note that all models in the figure are trained with 2000 hours of video data and we can see a trend of saturation with even larger models. This aligns with the scaling law observed in previous

Table I. **Full Ablation Study**. Here we provide a extended ablation study in supplementary **??**. The result is evaluated for all scenarios.

Training Components										
Ori. Loss	Feature Hall.	Fine-tuning	**Mean**	**Turn**	**Crossing**	**Detour**	**Proximity**	**Crowd**	**Other**	**All**
			17.03	**27.09**	16.25	16.72	16.99	13.28	11.88	13.16
✓			17.00	27.14	16.40	16.43	16.74	13.19	12.12	13.32
✓	✓		17.02	27.17	15.92	16.51	17.19	13.23	12.10	13.32
		✓	15.23	28.94	**13.90**	13.14	14.39	**11.19**	**9.91**	**11.12**
✓		✓	15.21	28.69	14.05	**13.12**	**14.17**	**11.19**	10.01	11.18
✓	✓	✓	**15.16**	27.36	14.05	13.20	14.44	11.59	10.31	11.41

Table II. Hyperparameters for training the CityWalker model.

Hyperparameter	Value
CityWalker Model	
Total # Parameters	214M
Trainable # Parameters	127M
Image Encoder	DINOv2 [2]
Backbone Arch.	ViT-B/14
City Walking Input Res.	350×630
Teleop Input Resolution	392×392
Token Dimension	768
Attn. Hidden Dim.	768
# Attention Layers	16
# Attention Heads	8
Input Context	5
Prediction Horizon	5
Input Cord. Repr.	Polar Cord.
Fourier Feat. Freq	6
Training	
# Epochs	10
Batch Size	32
Learning Rate	2×10^{-4}
Optimizer	AdamW [5]
LR Schedule	Cosine
Compute Resources	$2 \times$ H100
Training Walltime	30 hours
Fine-tuning LR	5×10^{-5}
L1 Loss Weight φ_{l1}	1.0
Ori. Loss Weight φ_{ori}	5.0
Arr. Loss Weight φ_{arr}	1.0
Feat. Loss Weight φ_{feat}	0.1

works [2, 6–8] that a larger model should be accompanied with larger data to produce better results.

Image Backbones. In Tab. IV, we show that our model performance is not sensitive to the choice of image backbones. This makes embodied depolyment very efficient. While our model with DiNOv2 backbone only has 1.7 fps inference speed on a RTX 3060 laptop, this can be boosted to 20 fps by switching to EfficientNetB0 backbone without sacrificing model performance.

Table III. VLM Results on Offline Data.

Scenario	**GPT-4o [4]**			**Ours**		
	↓AOE(5)	↓MAOE	↑Arrival	↓AOE(5)	↓MAOE	↑Arrival
Mean	72.22°	87.39°	69.38%	7.97°	15.23°	81.85%
Turn	68.61°	88.02°	68.66%	19.67°	26.63°	68.91%
Cros.	65.33°	81.12°	66.52%	5.43°	14.07°	75.03%
Detour	76.86°	90.76°	68.81%	8.71°	13.94°	78.54%
Prox.	75.65°	95.74°	66.33%	5.54°	14.32°	90.64%
Crowd	75.85°	84.88°	75.47%	4.77°	12.01°	87.50%
Other	71.03°	83.85°	70.49%	3.67°	10.40°	90.19%
All	71.51°	85.03°	70.04%	4.63°	11.53°	87.84%

Table IV. **Comparison of backbones and architecture.** All models are *pretrained* with 2000 hours of video and fine-tuned with expert data. Both metrics are taking the category mean. *Pretrained from ACO [9].

Metric	**EfficientNetB0**	**ResNet50**	**DiNOv2**	**ResNet34***	**ViNT****
MAOE (↓)	15.33°	15.16°	15.23°	**15.13** °	15.26 °
L2 (↓)	1.11 m	1.15 m	1.12 m	1.09 m	**1.08 m**

C. More Qualitative Results

In Fig. III, we provide more qualitative resting results on the offline data. We divide the results into three categories. Success: predicted action aligns well with ground truth action. Large error: predicted action does not align with ground truth but may still lead to successful navigation. Fail: predicted action may lead to failed navigation. The most significant observation is that large errors in offline data do not necessarily lead to failure in navigation, due to the multimodality characteristic of policy learning. For example, in the fifth row, although the ground truth action takes a detour to the right of the traffic drum, the predicted action that goes straight from the left of the drum should also lead to successful navigation.

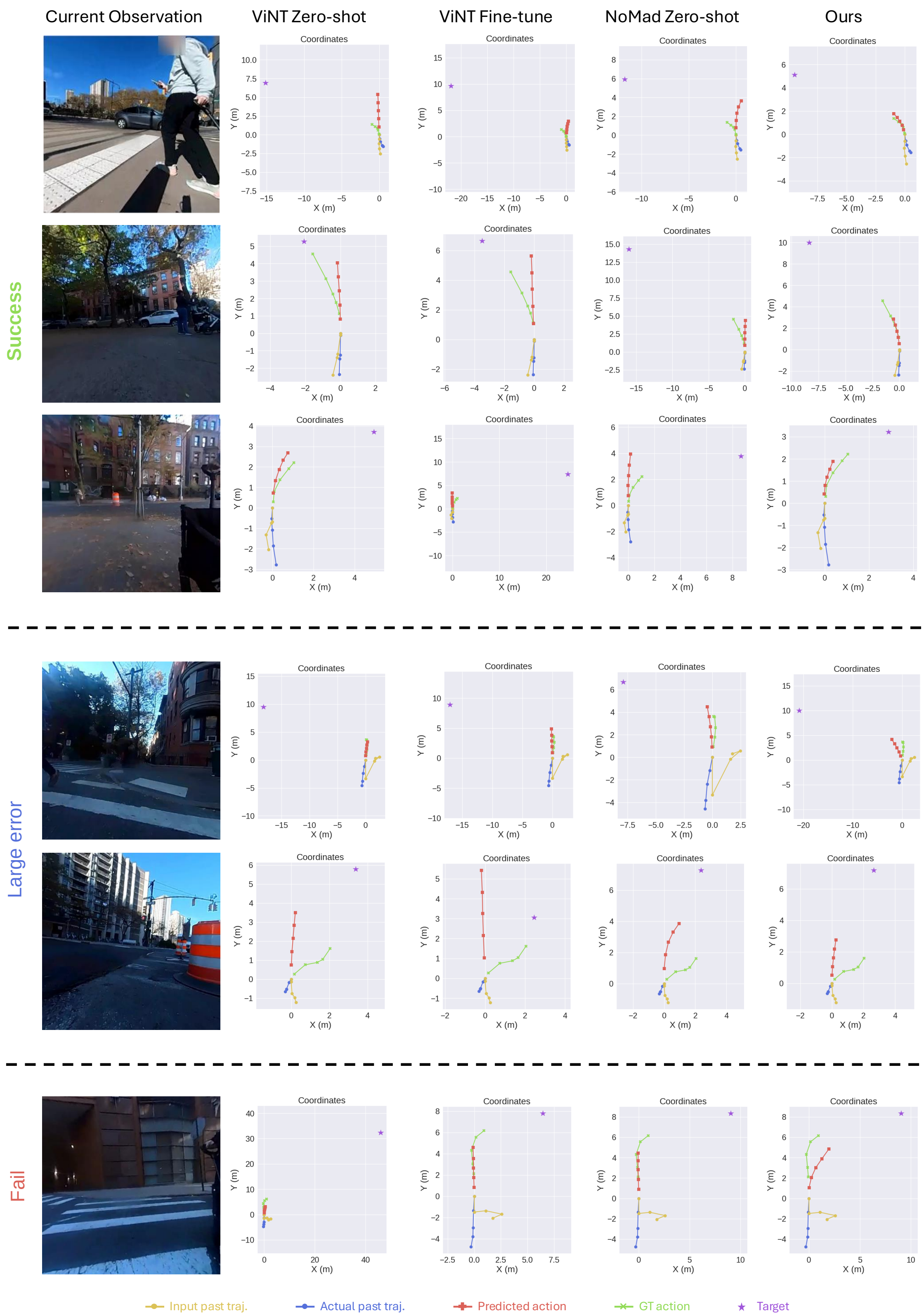

Figure III. **More Qualitative Results**. We provide more qualitative results in our offline testing. The results are categorized into success, large error, and fail. Success means the predicted action aligns with ground truth action. Large error mean prediction action does not align with ground truth but still lead to success navigation. Fails cases are those may lead to failed navigation.

References

[1] Dhruv Shah, Ajay Sridhar, Nitish Dashora, Kyle Stachowicz, Kevin Black, Noriaki Hirose, and Sergey Levine. ViNT: A foundation model for visual navigation. In *CoRL*, 2023. 1

[2] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 1, 2

[3] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[4] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2

[5] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 2

[6] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 2

[7] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023.

[8] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 2

[9] Qihang Zhang, Zhenghao Peng, and Bolei Zhou. Learning to drive by watching youtube videos: Action-conditioned contrastive policy pretraining. In *ECCV*, pages 111–128. Springer, 2022. 2