VENTURA: Adapting Image Diffusion Models for Unified Task Conditioned Navigation

Anonymous Author(s)

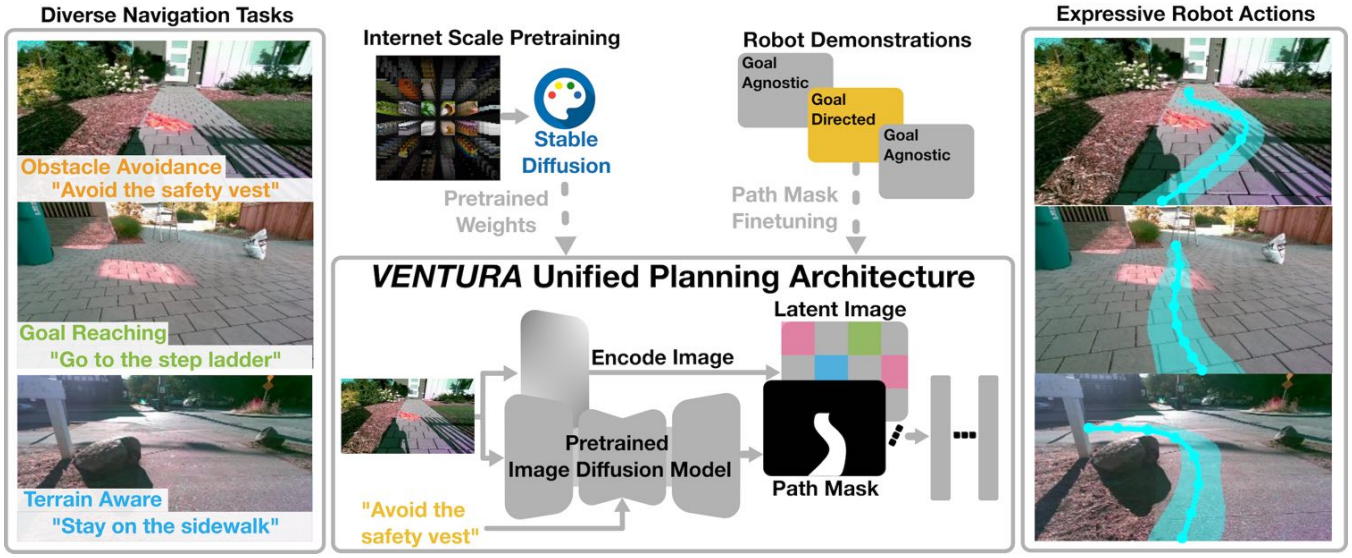

Fig. 1: Given an image and a language instruction (e.g. *avoid the safety vest*), VENTURA uses a fine-tuned image diffusion model to render a path mask in the image space. The path mask is then passed to a lightweight policy network to produce executable robot actions. By training on a mix of goal-agnostic and goal-directed demonstrations, VENTURA grounds diverse language instructions in safe and precise robot motions.

***Abstract*— Robots must adapt to diverse human instructions and operate safely in unstructured, open-world environments. Recent Vision–Language models (VLMs) offer strong priors for grounding language and perception, but remain difficult to steer for navigation due to differences in action spaces and pretraining objectives that hamper transferability to robotics tasks. Towards addressing this, we introduce VENTURA, a vision–language navigation system that finetunes internet-pretrained image diffusion models for path planning. Instead of directly predicting low-level actions, VENTURA generates a path mask (i.e. a visual plan) in image space that captures fine-grained, context-aware navigation behaviors. A lightweight behavior-cloning policy grounds these visual plans into executable trajectories, yielding an interface that follows natural language instructions to generate diverse robot behaviors. To scale training, we supervise on path masks derived from self-supervised tracking models paired with VLM-augmented captions, avoiding manual pixel-level annotation or highly engineered data collection setups. In extensive real-world evaluations, VENTURA outperforms state-of-the-art foundation model baselines on object reaching, obstacle avoidance, and terrain preference tasks, improving success rates by 33% and reducing collisions by 54% across both seen and unseen scenarios. Notably, we find that VENTURA generalizes to unseen combinations of distinct tasks, revealing emergent compositional capabilities. Videos, code, and additional materials: https://venturapath.github.io.**

I. INTRODUCTION

Mobile robots deployed in diverse, unstructured environments have untapped potential in domains such as construction inspection [1], urban maintenance [2], and last-mile delivery [3]. In these settings, robots must adapt their behavior to changing human preferences and environmental contexts. For example, at a construction site, a robot should avoid areas marked by caution tape, but it may enter if instructed by a worker to perform an inspection. In residential neighborhoods, robots should generally avoid disturbing private lawns, but may cut across to take out the trash when directed by a homeowner. Because situations often change quickly and unpredictably, robots must be able to adapt their behaviors rapidly based on diverse human instructions.

Language is a natural interface for conveying human intent, making it a flexible tool for adaptive autonomy in the open world. Recently, Vision-Language-Action (VLA) models output actions that enable robots to follow language instructions [4, 5, 6, 7, 8]. By leveraging internet-scale data, VLAs have shown promising open-set image and language understanding capabilities [6, 9]. However, existing VLA navigation systems struggle to ground language instructions in *precise* robot motions. For example, methods [8, 10, 11, 12] based on CLIP-style [13] embedding typically use language only to locate the target (e.g. *go to the red chair*) due to their limited language-conditioned planning capabilities. Systems that use transformer-based VLMs can plan at a coarse level using pre-defined image markers [7] or discrete actions [14]. However, consider the instruction "*Keep a safe*

distance from kids": a robot must not only understand the meanings of "safe" and "kids", but also generate precise motions to satisfy the intent. As of today, it remains an open question how to most effectively leverage the open-world knowledge in foundation models and ground it in precise navigation plans.

Driven by this question, we propose a new architecture VENTURA, that leverages pretrained image diffusion models [15] for planning. Like denoising an image from language description, VENTURA denoises a path mask (e.g. a "visual plan", see Fig. 1) to represent the robot's intended path across the scene. By formulating planning as an image generation problem, VENTURA leverages the rich visual-linguistic priors and strong image generation ability of diffusion models to render realistic and instruction-aligned visual plans. A lightweight Behavior-Cloning (BC) policy is sufficient to convert the visual plans into executable waypoints, thereby enabling VENTURA to ground diverse language commands in precise actions.

Notably, VENTURA uses a visual tracking approach to automatically construct the ground truth path masks. In this way, we obtain pixel-precise and natural-looking path masks that natively handle occlusions. It does not need robot odometry, an assumption required by current approaches [11, 16, 17]. Akin to the classifier-free guidance training in image-diffusion models, we train VENTURA with a mix of 8.5 hours unlabeled robot videos and 1.5 hours of language-trajectory data. This makes VENTURA potentially scalable to millions of internet videos. Combined, the high-quality groundtruths and label-efficient training scheme bolster VENTURA's generalization capabilities and precision.

We evaluate VENTURA in challenging outdoor environments, finding that VENTURA outperforms SOTA VLA navigation systems on a variety of common navigation tasks ranging from terrain-aware navigation, obstacle avoidance, and object-centric goal reaching. Our contributions are as follows: 1) A simple finetuning protocol that adapts image diffusion models for multi-task path planning, 2) A scalable label generation pipeline for image space planning from unstructured robot demonstrations, and 3) An open-source language-captioned navigation dataset to support future research towards VLA models for navigation.

II. RELATED WORK

Learning-based Navigation. Driven by demands for models that understand diverse goal instructions and intricate affordances, recent works have shifted towards learning-based methods for robot navigation. These approaches range from general-purpose, single-task, and multi-task navigation models. General-purpose models [3, 17, 18] learn task-agnostic policies that reason about various environmental factors for producing safe paths. Single-task models [19, 20, 21, 10] learn specialized costs or actions to achieve a single objective, but must be retrained for each new task. Multi-task models [7, 22] seek to unify these works, learning a policy capable of following multiple tasks and constraints. Achieving this requires models that generalize across a combinatorial set of tasks and environments, and is typically achieved by leveraging large pre-trained foundation models [23, 24]. While these models offer internet-scale priors that make this problem tractable, adapting them for robotics tasks requires overcoming novel challenges discussed in the next section.

Adapting Pre-trained Vision-Language Models. With the emergence of vision–language models (VLMs) trained on internet-scale datasets, a number of works have explored adapting them for navigation. These methods typically rely on prompting VLMs for tasks that resemble their pre-training objectives, such as annotating images [9], selecting between in-context examples [7, 12], or performing visual question answering (VQA) [14] to ground language instructions to robot actions. Other efforts fine-tune VLMs into vision–language–action (VLA) models to directly produce robot actions [25, 6], promoting more precise control. However, due to the significant differences between the original pre-training tasks and output space, these approaches struggle to follow semantically diverse task instructions and generate myopic local plans to reach long-horizon goals.

Learning from Robot Foundation Models and Internet Data. A complementary line of work directly distills affordance priors and actions from robot foundation models using large collections of internet data. These approaches [11, 19] condition on natural language or preference instructions to regress actions generated by an oracle navigation policy. While effective for following simple commands (e.g. *go to object x*), these methods struggle to accommodate multiple tasks or generalize beyond the training data. Moreover, methods that directly predict robot actions require supervision from robot odometry, limiting their scalability in domains where accurate odometry data is difficult to obtain, such as internet videos.

Relation to Prior Work. Transformer-based VLMs, while understanding high-level visual semantics, struggle with fine-grained spatial reasoning and planning [26]. In comparison, image diffusion models can generate high-fidelity images that align with language descriptions precisely. By formulating the navigation planning problem as an image generation problem, diffusion models can be more effective *visual planners*.

In terms of goal-conditioning, our work is most similar to LeLAN [11], which conditions on object-goal language instructions. While LeLAN is limited to single-task conditioning, VENTURA generalizes to diverse language instructions, enabling a multi-task policy that better aligns with open-world navigation demands. Methodologically, our approach also differs from prior efforts that employ diffusion models for navigation. Image-goal-conditioned policies [22] use diffusion models to generate intermediate image subgoals for exploration, while diffusion policies [27, 17] directly synthesize robot actions. By contrast, VENTURA leverages internet-scale priors from Stable Diffusion [15] to plan full trajectories directly in image space before grounding them into the robot's action space, providing a structured and interpretable representation that supports diverse, language-

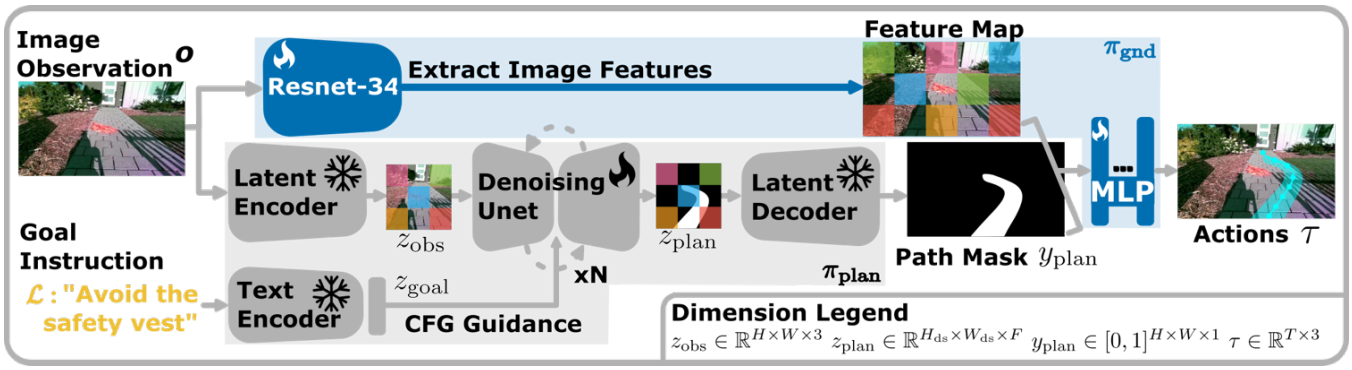

Fig. 2: VENTURA Architecture Overview. Our model is composed from an image diffusion planner π_{plan} and grounding policy π_{gnd}. π_{plan} generates a "visual plan" in the form of a path mask, guiding π_{gnd} to generate a sequence of xyz waypoints that satisfy the goal instruction $\mathcal{L}$.

conditioned tasks.

III. OVERVIEW

Our approach addresses the task-conditioned path planning problem, where the robot receives local camera observations $o \in \mathcal{O}$ and a language-specified task instruction $\mathcal{L}$, and must plan a path Γ to accomplish the task [28]. Similar to prior work on multi-task robot learning [29], we identify two tasks as distinct if they differ in what they optimize for and their task-specific constraints. For instance, object goal navigation requires precise maneuvering to an object whereas abiding by terrain preferences requires maneuvering on the most desirable terrain available.

To accomplish the task, the robot is given a policy $\pi : (o, \mathcal{L}) \rightarrow \mathcal{A}$ mapping the current observation and task tuple to a sequence of primitive actions $a \in \mathcal{A}$, where $a \in \mathbb{R}^3$ represents a sequence of xyz Cartesian waypoints. We assume the robot continuously replans given new observations, tracing a path Γ^π. The robot's objective is to generate a path Γ^π that lies within the set of acceptable paths given by the expert policy such that $\Gamma^\pi \in \{\Gamma | \Gamma \sim \pi^E\}$.

IV. APPROACH

We posit that adapting an image diffusion model to generate image-space plans is a highly effective way to transfer internet-scale semantic knowledge into navigation policies. To make this scalable, we exploit advances in off-the-shelf point tracking [30] to automatically extract plan labels from uncalibrated egocentric video, enabling supervision from diverse, unstructured data. Building on these ideas, VENTURA introduces two main components: a diffusion-based planner π_{plan} (see Fig. 2) that performs task-conditioned path planning in image space, and an auto-labeling pipeline (see Fig. 3) that provides the supervision needed to train π_{plan}. In the remainder of this section, we describe the VENTURA architecture, training objective, and scalable auto-labeling pipeline.

A. VENTURA *Architecture*

Our architecture is composed of two components, a language-conditioned image diffusion policy π_{plan} that generates path masks, and a grounding policy π_{gnd} that grounds these visual plans to trajectory waypoints (see Fig. 2).

We initialize π_{plan} from a pre-trained text-to-image latent diffusion model (Stable Diffusion v2 [15]) and freeze the variational autoencoder (VAE) and text encoder for the duration of training. We unfreeze the latent diffusion U-Net so that π_{plan} can adapt its denoising process for path mask generation. Our planner π_{plan} encodes the image observation o and natural language goal instruction $\mathcal{L}$ using the pre-trained image and text encoders to obtain a latent image z_{rgb} and goal z_{goal}. We sample an image $\hat{z}_{\text{mask}}$ from Standard noise and stack $\hat{z}_{\text{mask}}$ and z_{rgb} along the channel dimension. Our latent diffusion U-Net conditions on a stacked feature map consisting of the latent image and goal z_{goal}, and learns to denoise a latent path mask z_{plan}. Finally, we decode z_{plan} using the frozen VAE decoder. Since the pre-trained VAE decodes three-channel images, we average along the channel dimension to obtain a scalar likelihood map for the final image space plan y_{plan}.

VENTURA implements the grounding policy π_{gnd} using a ResNet-34 [31] to encode the current observation z_{obs} and stacks z_{obs} with y_{plan} along the channel dimension to construct the context vector c. We pass c to a Spatial Convolution [32] layer before using a Multilayer Perceptron (MLP) to predict a sequence of xyz waypoint targets.

B. VENTURA *Objective*

The planner π_{plan} and grounding policy π_{gnd} are trained in two stages with the following loss function:

$$\mathcal{L}_{\text{VENTURA}} = \mathcal{L}_{\text{plan}} + \mathcal{L}_{\text{gnd}}. \tag{1}$$

For diffusion training, we approximate the conditional distribution $p(y_{\text{plan}} | o, \mathcal{L})$, where $\mathcal{L}$ is the language task instruction. In the *forward* process, we start from $y_{\text{plan},0} := y_{\text{plan}}$ and gradually add Gaussian noise at levels $t \in \{1, ..., T\}$ to obtain noisy samples $y_{\text{plan},t}$:

$$y_{\text{plan},t} = \sqrt{\alpha_t} y_{\text{plan},0} + \sqrt{1 - \alpha_t}\epsilon, \tag{2}$$

where $\epsilon \sim \mathcal{N}(0, I)$, $\alpha_t = \prod_s^t = 1 - \beta_s$, and $\{\beta_1, ..., \beta_T\}$ is the process variance schedule. We train π_{plan} to predict

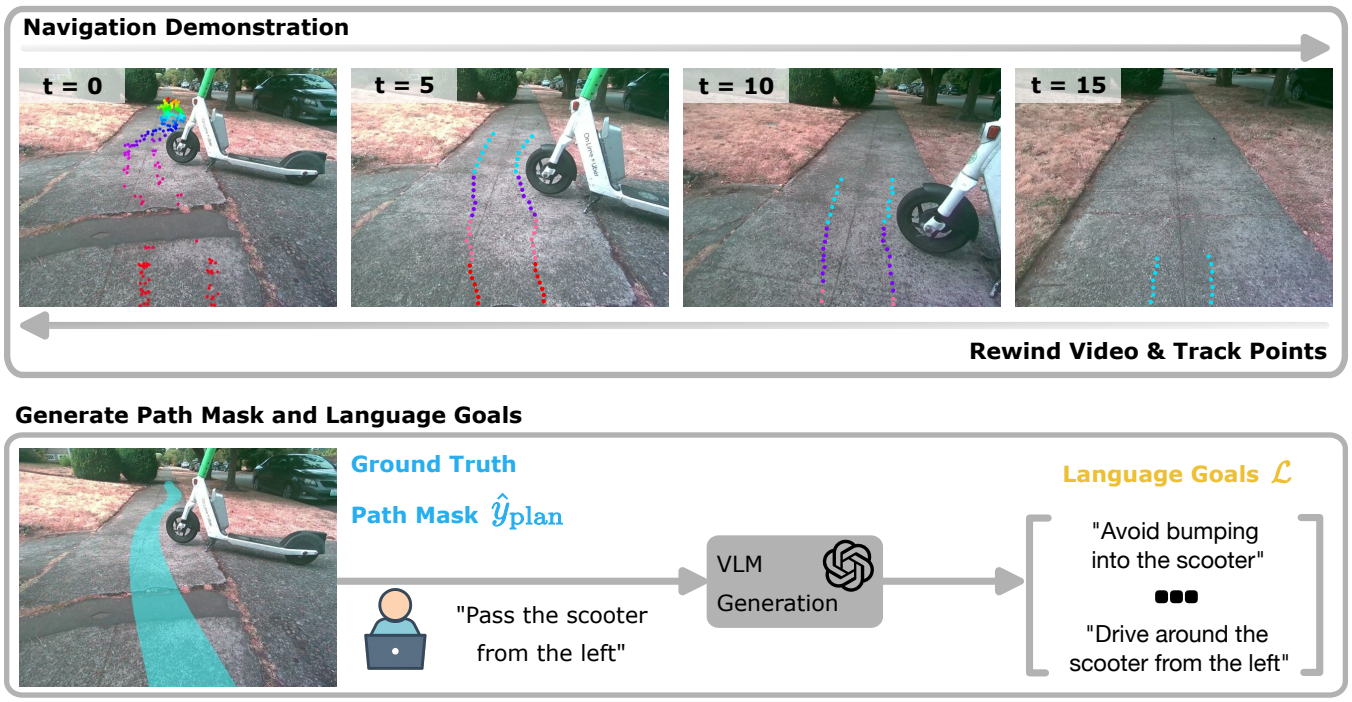

Fig. 3: VENTURA Data Generation Pipeline. We generate ground truth path masks by using an off-the-shelf tracker [30] to track points traversed by the robot in reversed videos. Each dots' color corresponds to the first time that it is tracked from. We provide seed captions to a VLM to generate diverse language goals that explain the path.

the image plan by gradually removing noise in the *reverse process*.

At train time, we sample a data point $(o, y_{\text{plan}}, \mathcal{L})$ and inject ϵ noise from a random timestep t to obtain the noise estimate $\hat{\epsilon} = \epsilon(y_{\text{plan, t}}, o, \mathcal{L}, t)$ and minimize the standard diffusion objective [33]:

$$\mathcal{L}_{\text{plan}} = \mathbb{E}_{y_{\text{plan},0},\epsilon \sim \mathcal{N}(0,I), t \sim \mathcal{U}(T)} ||\epsilon - \hat{\epsilon}||_2^2. \quad (3)$$

At inference time, we iteratively apply π_{plan} starting from noise $y_{\text{plan},T}$ to reconstruct the true image space plan $y_{\text{plan},0}$.

After training π_{plan} to convergence, we freeze π_{plan} and train the grounding policy π_{gnd} to minimize the mean squared error (MSE) loss between the predicted and ground truth actions a using the predicted image plan y_{plan}:

$$\mathcal{L}_{\text{gnd}} = MSE(\pi_{\text{gnd}}(c) - a), \quad (4)$$

where $c = \{y_{\text{plan}}, z_{\text{obs}}\}$.

C. Autolabeling Pipeline

Previously, we described the VENTURA objective, which assumes access to a ground truth path mask $\hat{y}_{\text{plan}}$ and robot actions a. In this subsection, we expand on how to automatically extract these ground truth masks.

Vision-based navigation models scale favorably with dataset size and diversity, motivating the need for flexible auto-labeling methods that work reliably across a variety of data collection setups. While it is possible to compute the robot's 3D position with a calibrated and synchronized hardware setup, this approach is unreliable for long trajectories as small state estimation errors can cause points in lethal regions to be considered traversable. We address these limitations by adopting an approach described in LRN [34] that uses Co-Tracker [30] to track masks in the image that correspond to the robot's future positions. Our approach generates masks that accurately represent the robot's path in the image without relying on accurate calibrations, human labels, or complex hardware setups.

To compute these masks, we play the video in reverse and drop a set of "breadcrumb tracks" on the pixels beneath the robot at the bottom of the image. We track these points across the entire video sequence, adding new points every 0.25 seconds. For each frame, we use the visibility value predicted by Co-Tracker to determine the set of visible points and construct a binary segmentation mask that best fits these points.

V. IMPLEMENTATION DETAILS

In this section, we describe the experimental setup and model-specific details to ensure fair evaluation. All baselines are evaluated on a wheeled quadrupedal robot (Unitree GO2-W) using monocular RGB observations from an Intel Realsense RGB-D camera (depth not used). Each method predicts a sequence of 8 cartesian waypoints spaced 0.4m apart that are tracked using the same model predictive controller (MPC).

To train VENTURA, we collect a dataset of 10 hours of navigation demonstrations, which we describe in detail in Sec. VI. We train the image planner π_{plan} for 50 epochs with a learning rate of 3e-4 and a batch size of 512, adopting the same training settings as Marigold [35] for the remaining hyperparameters. Additionally, we train with a classifier-free

Fig. 4: Ground truth language and path mask labels from the VENTURA dataset. We co-train VENTURA on a collection of navigation demonstrations with and without language captions. To bolster generalization to novel language prompts, we augment human-labeled captions using a pre-trained VLM to automatically generate diverse caption variations.

guidance weight of 0.05 to learn a task-conditioned and task-agnostic planner. We train the grounding policy π_{gnd} for another 50 epochs using the same settings as prior language-conditioned behavior cloning work [11].

Baselines. We evaluate VENTURA against LeLaN [11] and Convoi [7], two SOTA robot foundation model and VLM methods that predict waypoint actions given RGB observations and language commands. We pre-train LeLAN on the same GNM [16] and Youtube tour dataset used in the original work for 100 epochs before finetuning on the same dataset split used by VENTURA for another 100 epochs. We reproduce Convoi [7] as faithfully as possible since there is no open-source code release, removing the initial point cloud filtering safety layer to maintain fairness across each baseline.

VI. DATASET DETAILS

The VENTURA dataset consists of approximately 10 hours of navigation demonstrations, consisting of 8.5 hours of task-agnostic demonstrations and 1.5 hours of task-conditioned demonstrations. The task-agnostic demonstrations do not contain any unsafe actions, such as colliding into objects, and simply perform navigation to long-horizon goals. Our task-conditioned demonstrations are paired with language captions that describe behaviors such as going to objects, following spatial directions, following different terrain preferences, and avoiding objects. A small subset of these language captions and path masks are shown in Fig. 4. To generate corresponding language captions, a human labeler provides a short description to explain the observed navigation behavior. Then, we prompt gpt4o-mini [36] with a short system prompt, annotated image, and human-generated caption to automatically generate diverse, semantically identical captions for training.

Model	**Obs. Avoidance ↑**		**Obj. Goal ↑**		**Ter. Aware ↑**	
	Seen	Uns.	Seen	Uns.	Seen	Uns.
VENTURA	**13/15**	**4/5**	**9/10**	**7/10**	**6/6**	**5/6**
VENTURA-P	10/15	1/5	5/10	4/10	4/6	2/6
LeLaN [11]	9/15	1/5	8/10	3/10	3/6	2/6
Convoi [7]	8/15	3/5	7/10	**7/10**	4/6	3/6

TABLE I: **Multi-task planning evaluations.** VENTURA consistently outperforms baselines across representative robot navigation tasks in seen and unseen environments. We define the success rate criteria for each task in Sec. VII. Bolded numbers indicate the best performing method(s) for each category. We use the following abbreviations: Obs. - Obstacle, Obj. - Object, Ter. - Terrain, Uns. - Unseen, VENTURA-P - our approach without internet pre-training.

VII. EVALUATION

We evaluate VENTURA in 2 seen and 2 unseen outdoor environments and answer the following questions to understand the importance of our contributions and overall performance on multi-task and task-agnostic navigation.

- (Q_1) Does VENTURA improve success rate on diverse navigation tasks compared to SOTA approaches that leverage pre-trained foundation models?
- (Q_2) Is VENTURA able to use semantic knowledge from pre-trained foundation models to improve generalization performance?
- (Q_3) Does VENTURA improve the success rate on tasks that require long-range planning?

To investigate the preceding questions, we conduct more than 150 obstacle avoidance, object goal navigation, and preference-aware terrain navigation experiments against LeLaN [11] and Convoi [7]. Our test environments feature

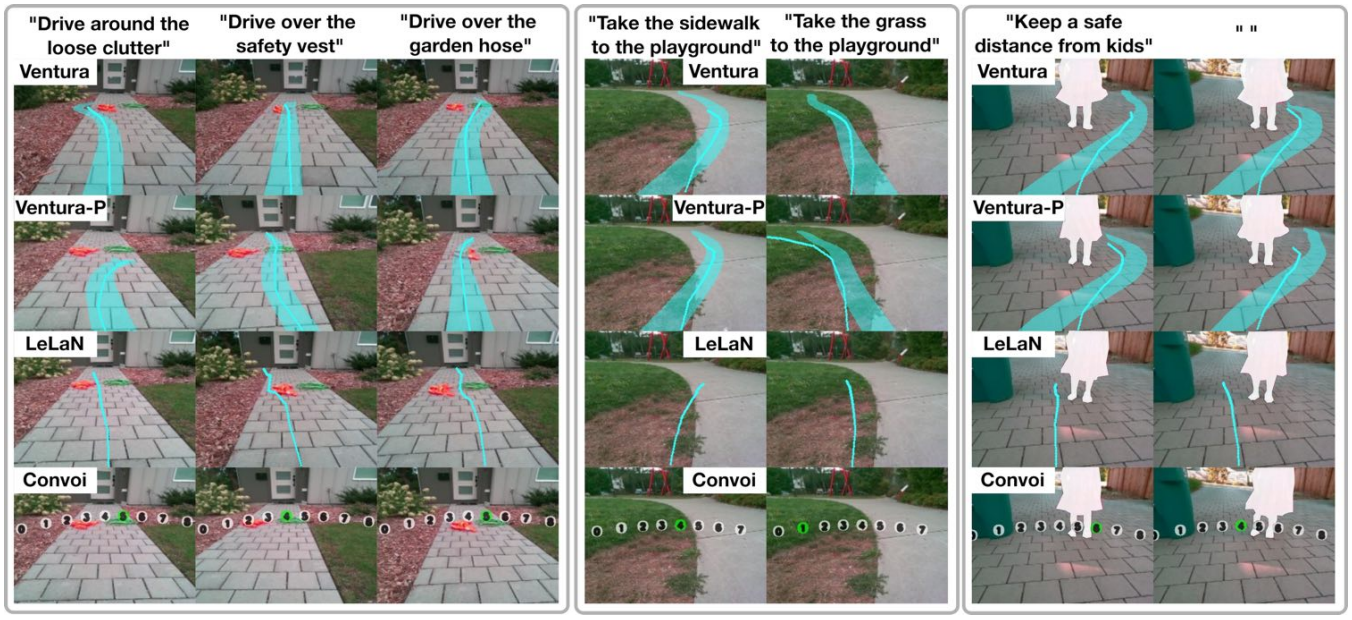

Fig. 5: **Qualitative analysis of various visual-language navigation baselines.** VENTURA consistently outperforms existing approaches in terms of task alignment and generalization to unseen entities and environments. We use the following abbreviations: VENTURA-P: Our model without initializing with internet-pretrained weights.

Model	Short ↑		Medium ↑		Long ↑	
	Seen	Uns.	Seen	Uns.	Seen	Uns.
VENTURA	**6/6**	**6/6**	**6/6**	**5/6**	**5/6**	**4/6**
VENTURA-P	3/6	1/6	3/6	2/6	3/6	1/6
LeLaN [11]	6/6	4/6	4/6	2/6	2/6	1/6
Convoi [7]	**6/6**	5/6	5/6	2/6	3/6	3/6

TABLE II: **Long range planning evaluations.** We show the success rate of reaching object goal targets across 140 trials with 10 objects. A trial is deemed successful if the robot reaches with 0.5 m of the target object. Bolded numbers indicate the highest performing method(s) per category. We use the following abbreviations: Uns. - Unseen, VENTURA-P - our approach without internet pretraining.

Model	Mean L_2 Error ↓	Hausdorff Distance ↓
VENTURA	**0.04**	**0.08**
VENTURA-P	0.06	0.09
LeLaN [11]	0.09	0.17

TABLE III: **Trajectory Error on the Test Set.** We compare the test set performance of each baseline that is trained on the VENTURA dataset. Our approach achieves the lowest average L_2 error and Hausdorff distance, indicating that our method is able to generate more precise actions that closely match the expert behavior. Bolded numbers indicate the best performing method(s) for each metric.

a diverse set of objects and terrains ranging from common entities like trash cans and sidewalks to rare entities like safety vests and playgrounds. We evaluate each method using success rate as the primary criteria, classifying trials as failures if the robot does not reach within 0.5m of the goal, collides with an obstacle, or drives on unfavorable terrain for more than 2 seconds.

Towards understanding $\mathcal{Q}_1$, we observe in Table I that VENTURA outperforms all other approaches in seen and unseen environments by 40% and 33% respectively on average across all tasks. This is consistent with the results in Table III, demonstrating that our method is able to plan paths that align more closely with expert behavior. Specifically, we find that our approach is able to ground diverse actions to unseen entities even under instruction ambiguity. We highlight this behavior in Fig. 5, where VENTURA correctly avoids an unseen safety vest and garden hose when told to "avoid loose clutter". Furthermore, the same model can rapidly adapt its behavior to align with more specific instructions, such as "drive over the garden hose". By comparison, while LeLAN and Convoi can follow specific instructions, they struggle to infer user intent when given ambiguous commands like "avoid loose clutter". We also observe that VENTURA is able to generate precise motion commands that respect nuanced commands like "keep a safe distance from kids". From these results, we conclude that VENTURA is significantly more effective at interpreting language commands and identifying collision-free paths across common navigation tasks.

Towards Q_2, we compare VENTURA with and without StableDiffusion weight initialization to understand how internet pre-training on non-robotics tasks transfers to robot path planning. We observe that initializing the denoising Unet with StableDiffusion improves overall performance by 47% and 128% on average across seen and unseen scenarios respectively compared to training from scratch (VENTURA-P). Fig. 5 corroborates these findings, showing that the model trained from scratch struggles to identify unseen entities, often behaving randomly for object-centric goal navigation when presented with multiple unseen options. Even with these limitations, VENTURA-P performs on par with LeLaN

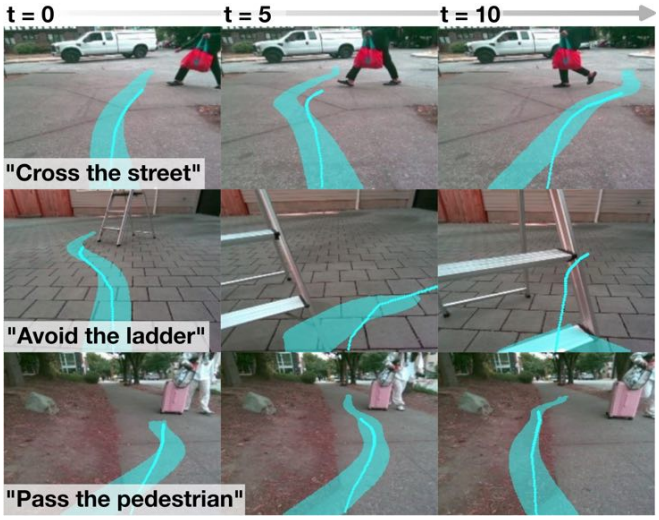

Fig. 6: Limitations of VENTURA. In the first row, our approach struggles to infer social dynamics, leading to path plans that cut in front of the pedestrian. In the second row, we test our model's ability to handle tight turns when avoiding obstacles. It is difficult to represent backwards actions as visual plans, leading to suboptimal behavior when backtracking is the only valid path. In the final row, we observe that while VENTURA is able to pass the pedestrian safely, the generated plans are often temporally inconsistent, occasionally resulting in unstable behavior.

despite being trained on far less robot data. We hypothesize that this is possible because the StableDiffusion text and image encoders are pre-trained on more diverse data sources than those used by robot foundation models, enabling better zero-shot generalization compared to learning these model components from scratch.

To understand Q_3, we vary the target object distance for the object goal navigation task (4m, 8m, 12m) and compare each model's ability to perceive and plan towards long-range entities. From Table II, we find that while existing approaches perform comparably to VENTURA in short to medium ranges, our approach separates itself for longer distances, outperforming the second best approach, Convoi, by 21.4% and 50.0% on average across seen and unseen scenarios respectively. Interestingly, we observe that the most common sources of failures are caused by an inability to localize the target object and losing sight of the target object. As seen in Fig. 5, LeLaN and Convoi plan paths that close the target distance, but neglect to consider how the future path affects the visibility of the target object. This introduces failures where the target object gradually drifts out of the field of view. In contrast, VENTURA predicts path masks directly in the image, resulting in precise, long range plans that reduce the likelihood of these kinds of myopic decisions.

VIII. Limitations and Future Work

While VENTURA inherits open-set semantic knowledge from pre-trained foundation models, it does not enable generalization to novel motion primitives. This limits our model's ability to follow complex motion patterns not seen in the training data, such as "circling around the house". Furthermore, it is difficult to capture motion dynamics with visual plans, which are important for scenarios depicted in Fig. 6 with dynamic agents (e.g. social navigation) or complex vehicle dynamics (e.g. offroad driving). Another promising direction to explore is extending VENTURA to reason about multiple observations and produce temporally consistent plans. This will enhance robustness in long horizon partially observable environments that require joint understanding of information from multiple viewpoints.

IX. Conclusion

In this paper, we presented VENTURA, a flexible vision-language model that repurposes pre-trained image diffusion models to plan paths that follow diverse language instructions. Our unified policy uses a pre-trained image diffusion backbone pre-trained for image generation to generate path masks (i.e. visual plans) conditioned on language commands. We train a lightweight behavior cloning policy to ground these path masks to robot actions, demonstrating its robustness and generalizability to novel environments despite limited on-robot training data. We study our approach's effectiveness across a variety of navigation environments and tasks, showing improvements of up to 33% in performance compared to SOTA in unseen settings. Based on these findings, we believe that VENTURA presents a promising direction for leveraging internet-scale priors to achieve adaptive, open-world autonomy.

References

[1] Tomáš Rouček et al. "Darpa subterranean challenge: Multi-robotic exploration of underground environments". In: *International Conference on Modelling and Simulation for Autonomous Systems*. Springer. 2019, pp. 274–290.

[2] Lynne E Parker and John V Draper. "Robotics applications in maintenance and repair". In: *Handbook of industrial robotics* 2 (1998), pp. 1023–1036.

[3] Arthur Zhang et al. "CREStE: Scalable Mapless Navigation with Internet Scale Priors and Counterfactual Guidance". In: *Robotics: Science and Systems (RSS)*. 2025.

[4] Anthony Brohan et al. "RT-1: Robotics Transformer for Real-World Control at Scale". In: *Robotics: Science and Systems XIX* (2023).

[5] Brianna Zitkovich et al. "Rt-2: Vision-language-action models transfer web knowledge to robotic control". In: *Conference on Robot Learning*. PMLR. 2023, pp. 2165–2183.

[6] Kevin Black et al. "π: A Vision-Language-Action Flow Model for General Robot Control". In: *CoRR* (2024).

[7] Adarsh Jagan Sathyamoorthy et al. “Convoi: Context-aware navigation using vision language models in outdoor and indoor environments”. In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2024, pp. 13837–13844.

[8] Chenguang Huang et al. “Visual Language Maps for Robot Navigation”. In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2023, pp. 10608–10615.

[9] Soroush Nasiriany et al. “PIVOT: Iterative Visual Prompting Elicits Actionable Knowledge for VLMs”. In: ().

[10] Matthew Chang et al. “GOAT: GO to Any Thing”. In: *Robotics: Science and Systems*. 2024.

[11] Noriaki Hirose et al. “LeLaN: Learning A Language-Conditioned Navigation Policy from In-the-Wild Video”. In: *Conference on Robot Learning*. PMLR. 2025, pp. 666–688.

[12] Dhruv Shah, Błażej Osiński, Sergey Levine, et al. “LM-Nav: Robotic Navigation with Large Pre-Trained Models of Language, Vision, and Action”. In: *Conference on Robot Learning*. PMLR. 2023, pp. 492–504.

[13] Alec Radford et al. “Learning transferable visual models from natural language supervision”. In: *International conference on machine learning*. PmLR. 2021, pp. 8748–8763.

[14] Jiazhao Zhang et al. “NaVid: Video-based VLM Plans the Next Step for Vision-and-Language Navigation”. In: *Robotics: Science and Systems (RSS)*. 2024.

[15] Robin Rombach et al. “High-resolution image synthesis with latent diffusion models”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695.

[16] Dhruv Shah et al. “GNM: A General Navigation Model to Drive Any Robot”. In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2023, pp. 7226–7233.

[17] Ajay Sridhar et al. “Nomad: Goal masked diffusion policies for navigation and exploration”. In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2024, pp. 63–70.

[18] Pascal Roth et al. “Viplanner: Visual semantic imperative learning for local navigation”. In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2024, pp. 5243–5249.

[19] Luisa Mao et al. “Pacer: Preference-conditioned all-terrain costmap generation”. In: *IEEE Robotics and Automation Letters* (2025).

[20] Matt Schmittle et al. “Long Range Navigator (LRN): Extending robot planning horizons beyond metric maps”. In: *RSS 2025 Workshop on Resilient Off-road Autonomous Robotics*.

[21] Daeun Song et al. “Vl-tgs: Trajectory generation and selection using vision language models in mapless outdoor environments”. In: *IEEE Robotics and Automation Letters* (2025).

[22] Dhruv Shah et al. “ViNT: A Foundation Model for Visual Navigation”. In: *7th Annual Conference on Robot Learning*.

[23] Oriane Siméoni et al. “Dinov3”. In: *arXiv preprint arXiv:2508.10104* (2025).

[24] Lucas Beyer et al. “PaliGemma: A versatile 3B VLM for transfer”. In: *CoRR* (2024).

[25] An-Chieh Cheng et al. “NaVILA: Legged Robot Vision-Language-Action Model for Navigation”. In: *RSS*. 2025.

[26] Jihan Yang et al. “Thinking in space: How multimodal large language models see, remember, and recall spaces”. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 10632–10643.

[27] Jing Liang et al. “Dtg: Diffusion-based trajectory generation for mapless global navigation”. In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2024, pp. 5340–5347.

[28] Matthijs TJ Spaan. “Partially observable Markov decision processes”. In: *Reinforcement learning: State-of-the-art*. Springer, 2012, pp. 387–414.

[29] Jose Barreiros et al. “A careful examination of large behavior models for multitask dexterous manipulation”. In: *arXiv preprint arXiv:2507.05331* (2025).

[30] Nikita Karaev et al. “CoTracker3: Simpler and Better Point Tracking by Pseudo-Labelling Real Videos”. In: *CoRR* (2024).

[31] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[32] Chelsea Finn et al. “Deep spatial autoencoders for visuomotor learning”. In: *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2016, pp. 512–519.

[33] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.

[34] Matt Schmittle. “Off-Road Navigation Under Sensing Uncertainty”. PhD thesis. 2025.

[35] Bingxin Ke et al. “Marigold: Affordable Adaptation of Diffusion-Based Image Generators for Image Analysis”. In: *arXiv preprint arXiv:2505.09358* (2025).

[36] Josh Achiam et al. “Gpt-4 technical report”. In: *arXiv preprint arXiv:2303.08774* (2023).