# A Steerable Vision-Language-Action Framework for Autonomous Driving

Tian Gao*[1], Catherine Glossop*[2], Kyle Stachowicz[2], Timothy Gao[2], Celine Tan[2], Oier Mees[2], Yuejiang Liu[1], Sergey Levine[2], Dorsa Sadigh[1] and Chelsea Finn[1]
[1]Stanford University [2]UC Berkeley

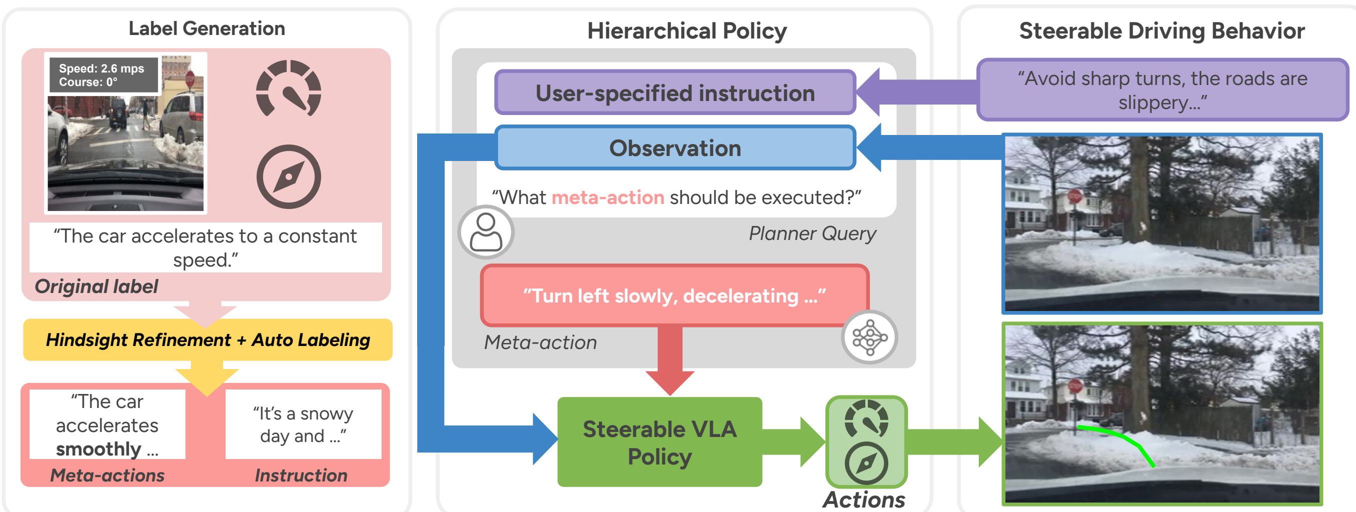



Fig. 1: We present SteerVLA, a framework for training steerable driving VLAs using hindsight label refinement and a flexible hierarchical architecture. We demonstrate that our policy can follow language instructions and reason over visual references and user instructions.

***Abstract*—For autonomous vehicles to be truly useful, they must move beyond fixed rules to understand nuanced human intent and adapt to diverse scenarios. We introduce a hierarchical vision-language-action (VLA) framework for promptable autonomous driving. Our framework combines a high-level planner that reasons over high-level specifications of desired behaviors – *"I'm late for work, get me there as fast as possible"* – to generate intermediate language commands with a low-level policy that grounds these intermediate commands into trajectory-level actions. To generate diverse paired language data from driving datasets without structured language labels, we propose a label refinement pipeline that makes use of off-the-shelf VLMs applied to hindsight data to generate a "preference function" aligning high-level user specifications with their corresponding intermediate- and low-level commands. We evaluate our framework against both real and simulated driving datasets, using the Berkeley DeepDrive dataset and the CARLA simulator, respectively, and find that it provides a highly steerable driving policy that is responsive to user prompts without compromising driving performance.**

## I. Introduction

Autonomous driving has seen great progress in recent years, with the advent of end-to-end learned behaviors enabling increasingly flexible behaviors [11, 19]. However, current driving models aim to achieve a single "nominal" behavior, offering limited support for user customization and lacking the ability to be steered by the user via nuanced language instructions.

In the real world, different users have different preferences for how a car should act in a given scenario. For example, passengers who are running late for a flight might want their car to act very differently from a passenger who is carrying a full cup of hot coffee. We propose that driving policies should be able to reason over both visual information and freeform user-specified instructions to produce steerable behaviors, just like their human counterparts. This allows a model to follow user preferences (e.g., *"drive cautiously, I'm carrying a cup of hot coffee"*) as well as specific commands (e.g., *"cut in behind the black truck to take this exit"*). Traditional self-driving models rely on rule-based reasoning, with language following restricted to a limited set of high-level routing commands (e.g., *"turn left"*), which makes following instructions that require implicit reasoning difficult.

On the other hand, learning-based methods that leverage vision-language models (VLMs) [9] or fine-tune vision-language-action models (VLAs) [3, 11] endow the model with semantic reasoning but have largely focused on interpreting or narrating the behavior of a vehicle rather than improving their

* Equal contribution
Correspondence to: tiangao@stanford.edu

instruction-following capabilities. In this work, we present a hierarchical vision-language-action framework for instruction-following and user-aligned autonomous driving. Our key insight is that a hierarchical architecture enables the decoupling of reasoning and acting into two policies. We preserve the VLM's strong semantic priors in the high-level planner, which interfaces with the low-level policy through *meta-actions* such as *"turn left cautiously"*. The low-level policy then grounds these meta-actions into continuous control trajectories.

Applying this framework to off-the-shelf driving datasets, we build SteerVLA, a user-steerable hierarchical VLA for autonomous vehicles. We adapt existing datasets focused on the interpretability of driving behaviors, using hindsight information to train a highly steerable low-level policy. A powerful off-the-shelf VLM serves as our high-level planner.

We evaluate our framework using offline metrics on real-world driving datasets, including open-loop evaluations of the full pipeline. Our experiments assess both the driving performance and instruction-following ability of the model across a range of instructions and scenarios. We compare our low-level policy's performance to a state-of-the-art baseline and demonstrate clear improvements in both control quality and language-following accuracy. Extensive qualitative examples illustrate the steerability of our approach.

## II. Related Work

**Vision-Language-Action models.** Inspired by the success of pretrained vision-language models (VLMs), several works have introduced *vision-language-action* (VLA) models [4], which typically consist of a VLM backbone fine-tuned to produce robot actions, rather than language, conditioned on visual inputs and language instructions [15]. These models benefit from excellent cross-modal grounding between language and vision, enabling the transfer of internet-scale semantic knowledge from the pretraining data. Recent works have also sought to imbue VLAs with reasoning capabilities [28] to improve generalization and compositional task-following, and have introduced hierarchical structure to improve long-horizon behavior [2, 12].

**End-to-end policies for autonomous driving.** Recent work has explored a range of approaches for integrating multimodal foundation models into autonomous driving [8, 10, 25]. Some efforts leverage pretrained VLMs to provide driving systems with broad world knowledge and reasoning capabilities [20], while others have sought to develop VLA policies for driving by fine-tuning VLMs with an action head [11, 29]. These works typically focus on the low-level act of driving a car, with the "language" component of the VLA used mostly as an auxiliary learning signal or for very structured instructions (at the level of our meta-actions). In contrast to these works, we aim to develop a driving policy that is highly steerable in response to open-ended user instructions—despite the lack of a pre-existing dataset with these types of labels.

**Steerable VLA policies.** One major promise of VLA policies is that their VLM backbones implicitly have strong language-following priors—in other words, they should be highly amenable to *steering* with open-ended language instructions. Prior work [21] has shown that it is possible to fine-tune a VLA policy with steerable instructions, given sufficient data.

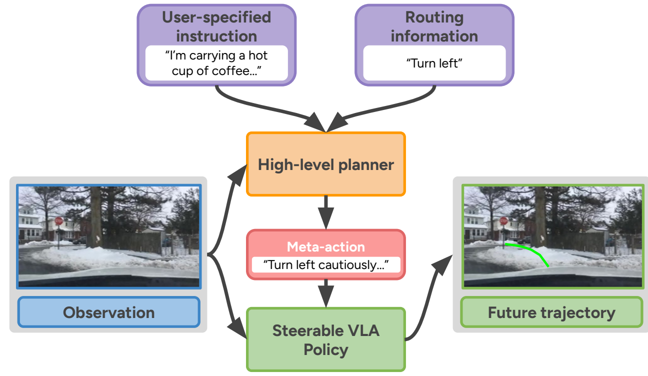



Fig. 3: **Architecture of the driving policy.** We use a flexible hierarchical architecture, which takes in an instruction, routing information, and the current observation of the vehicle and produces the future trajectory of the vehicle.

However, despite encouraging progress, existing practical VLA policies in open-world settings like driving often suffer from limited language steering. Works studying this effect have found that training on actions can degrade the internet-scale knowledge acquired during pretraining. In other words, the sudden shift from typical VLM pretraining tasks to a robot action generation task during fine-tuning can harm the network's general semantic knowledge and language-understanding capabilities [6]. We build upon prior work that utilizes a hierarchical framework [13, 16] to mitigate this shift by training a high-level policy on tasks that resemble the original pretraining distribution, and a low-level policy that interfaces with the high-level policy through an intermediate representation. In this work, we use a VLM as the high-level policy and fine-tune a VLA, with structured "meta-actions" forming an intermediate bridge between the two components.

## III. Method

To achieve steerable driving policies, we require diverse, language-labeled data that allow us to train policies capable of flexibly understanding and producing actions in line with user-specified instructions. Additionally, we require a policy architecture conducive to learning how to follow complex language prompts that demand strong reasoning abilities. To this end, we describe the two key components of our method: 1) a flexible hierarchical VLA policy architecture with meta-actions as an intermediate representation, and 2) VLM-guided hindsight labeling of driving data.

### A. Hierarchical VLA Architecture

The first hurdle to following complex language prompts is being able to understand complex language in the context of visual observations. Therefore, we use a hierarchical VLA policy architecture, where the low-level policy is fine-tuned from a powerful VLM pre-trained on internet-scale data, offering strong semantic priors for vision and language, and the high-level policy is a powerful off-the-shelf VLM. By using a hierarchical structure, the high-level policy is tasked with focusing on the *reasoning* component of the task. The low-level policy is trained to produce control commands (speed and course deltas) conditioned on the meta-actions, allowing it to focus more on *grounding* and *acting*. An overview of the architecture is provided in Fig. 3.

**High-level planner.** The high-level policy, or "planner," is tasked with interpreting a complex instruction and reasoning about which meta-action should be taken at the current time step to follow it. We instantiate the high-level planner as a powerful VLM [22], leveraging its strong semantic priors to generate a suitable meta-action that captures both the global and local nuances of the instruction. We structure the query to the VLM as a visual question-answering problem by providing the current observation and speed, and prompting the model to produce an appropriate meta-action based on a few in-context examples.

**Low-level VLA policy.** Once a meta-action has been generated, the steerable low-level policy predicts actions that align with the desired behavior. To this end, we train a meta-action-conditioned VLA policy on the BDD-X dataset [14] (see Section III-B for details on generating meta-action labels) using PaliGemma [1] as the backbone for the VLA. We follow the recipe from [15], using special tokens to represent discretized actions one dimension at a time. Unlike OpenVLA, we also predict an open-loop *action chunk* [5, 7] which enables smooth temporally-correlated actions and decreases compute requirements. The policy takes as input the current front camera image observation of the vehicle and the current speed. The output is a chunk of 6 timesteps each including delta speed and course (steering angle) over the next three seconds at a frequency of 2 Hz, normalized based on the dataset statistics [3, 24].

### B. Generating Diverse Synthetic Labels for Driving Data.

While driving datasets with language labels exist, they often consist of short-horizon trajectory descriptions with limited detail and do not capture higher-level driver intentions—such as those in the BDD-X dataset [14]. However, to enable fine-grained language following, we require detailed meta-action labels. To address this, we perform a VLM refinement step to determine the "style" and "motion extent" of the driving behavior for each trajectory chunk. We leverage future trajectory speed and course information through the benefit of *hindsight*, using information that is unavailable to the final policy at inference time, but accessible during annotation. For example, we transform the original label "the car rolls through the stop sign" into the more fine-grained "the car rolls through the stop sign with a slight right turn, accelerating gradually, driving normally." As a result, we obtain a dataset of (meta-action, action chunk, observations) tuples that can be used to fine-tune our low-level VLA policy. For detailed prompts provided to the VLM and example refinements, see Appendix A.

## IV. Experiments

Our experiments answer the following questions:

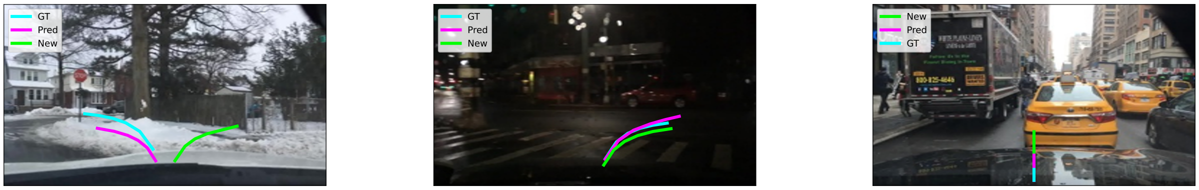
- How accurately does SteerVLA predict driving trajectories given free-form language instructions?
- How well can it follow diverse language instructions via meta-actions?
- Does our automatic meta-action annotation provide effective supervision?
- How effectively does the high-level planner generate meta-action plans?

### A. Experimental Setup

**Data.** We train the VLA policy on the BDD-X training split [14], which provides high-level natural language descriptions of driver behavior. We filter out sequences with corrupted or missing GPS data, resulting in approximately 16,000 training frames and 2,000 test frames. We evaluate the models on the test set, which contains unseen language instructions and novel driving scenes. To improve language-conditioned learning, we refine BDD-X descriptions using GPT-4o [17]; these refined descriptions serve as language instructions for the VLA policy.

**Evaluation protocols.** To assess trajectory prediction accuracy, we report Average Displacement Error (ADE) and Final Displacement Error (FDE) at 1s, 2s, and 3s prediction horizons [18], along with Root Mean Square Error (RMSE) for future speed and course angle. To evaluate instruction-following capability, we conduct a blind manual evaluation over 20 rollouts per model. Human annotators determine whether each predicted trajectory aligns with the given language instruction, without knowledge of which model generated it. To assess the high-level planner, we compare predicted meta-actions against ground-truth annotations using standard language generation metrics: BERTScore, BLEU, and ROUGE-L. We also evaluate the full pipeline, where the high-level planner generates meta-action plans that serve as language instructions for the VLA policy.

**Generating user-specified instruction labels.** While some datasets include labels at the level of meta-actions, we are unaware of any that provide labels at the user-specified instruction level. However, such labels are essential for evaluating the full instruction-following pipeline, in which a high-level planner generates language commands that the low-level policy must execute. To generate these labels, we once again leverage hindsight labeling of trajectories. As shown in Fig. 5, we provide a summarized description of the vehicle's behavior in natural language, derived from the refined BDD-X labels described in Section III-B. We then query Gemini 2.0 Flash to predict a high-level command or routing instruction from a fixed set (e.g., *turn right, turn left, move forward, stop/slow down*) [26], akin to the guidance provided by an in-car navigation system, along with a *persona* capturing the driver's likely motivation or situational context. This process yields a dataset containing trajectories labeled with both short-horizon meta-actions and longer-horizon, user-oriented instructions. The prompting details are provided in Appendix A.




(a) **Original:** "The car makes a smooth left turn, decelerating then accelerating, with normal driving style." **New:** "The car makes a smooth right turn, decelerating then accelerating, with normal driving style."

(b) **Original:** "The car accelerates steadily while making a smooth, wide right turn, reflecting a normal driving style.", **New:** "The car decelerates steadily while making a smooth, wide right turn, reflecting a normal driving style."

(c) **Original:** "The car accelerates slowly and steadily forward, maintaining a straight course, driving normally.", **New:** "The car accelerates quickly and steadily forward, maintaining a straight course, driving normally."

Fig. 4: **Qualitative language following performance across various scenes.** *GT* denotes the ground-truth trajectory from the dataset. *Pred* represents the predicted trajectory conditioned on the original meta-action command. *New* shows the predicted trajectory in response to a newly specified meta-action command. (a) and (b) evaluate the VLA's ability to follow coarse-grained instruction, such as turning left vs. right or accelerating vs. decelerating. (c) evaluates fine-grained instruction following, involving subtle distinctions like accelerating quickly vs. slowly.

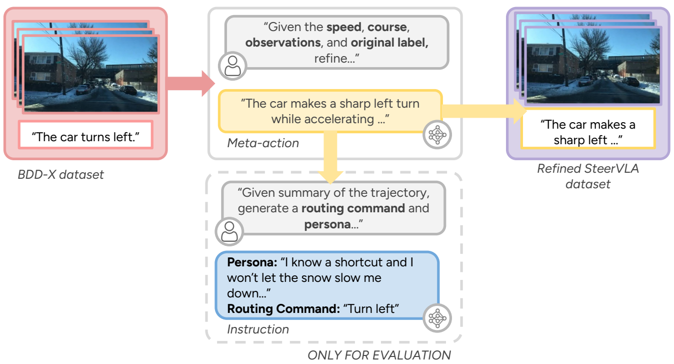



Fig. 5: **An overview of the label refinement and evaluation dataset auto-labeling pipeline.** We leverage trajectory information in hindsight and a powerful VLM to perform large-scale meta-action label refinement and generation of an evaluation dataset.

### *B. Low-Level VLA Policy Evaluation*

To evaluate trajectory prediction, we compare SteerVLA with the DriveGPT4 baseline [27], using the same BDD-X train/test splits. Since DriveGPT4 does not take language input, we also report SteerVLA's performance without language instructions. As shown in Table I, SteerVLA significantly outperforms DriveGPT4 in both speed and turning angle prediction.

To evaluate instruction-following, we conduct an ablation study across three settings: (i) without language instructions (SteerVLA w/o lang), (ii) with raw BDD-X instructions (SteerVLA w/ lang), and (iii) with refined instructions generated by our meta-action autolabeling pipeline (SteerVLA w/ refined lang). As shown in Table II, incorporating language consistently improves trajectory prediction. While the refined instructions yield only modest improvements in ADE/FDE over raw instructions, they lead to significantly better performance in human evaluation, as shown in Table III.

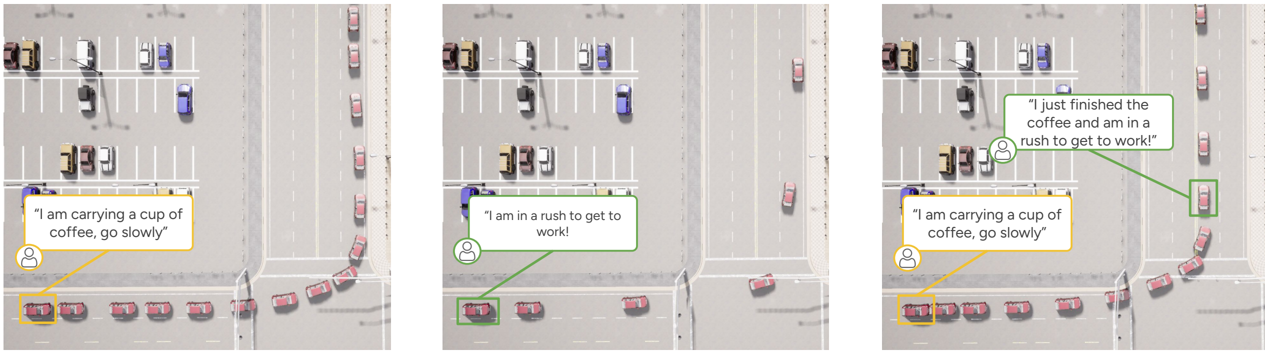
To better capture instruction adherence, we conduct a manual evaluation on 20 rollouts per model. For each rollout, human annotators assess whether the predicted trajectory aligns with the given instruction. Unlike ADE/FDE—which are strict L2-based metrics measuring deviation from ground-truth trajectories—human evaluation directly assesses whether the control behavior matches the intended semantics of the instruction. This is particularly important in cases where the predicted trajectory deviates from the ground truth but still satisfies the instruction. Our refined instructions explicitly encode such behavioral cues, enabling more expressive and interpretable control, which in turn results in more instruction-aligned driving behavior.

| Method | Speed (m/s) RMSE↓ | Turning angle (degree) RMSE↓ |
|---|---|---|
| DriveGPT4 [27] | 1.30 | 8.98 |
| SteerVLA w/o lang | 0.57 | 2.39 |
| SteerVLA w lang | **0.53** | **2.16** |

TABLE I: **Comparison of our VLA policy (with and without language instructions) and DriveGPT4 on trajectory prediction.** SteerVLA significantly outperforms DriveGPT4 in both speed and turning angle prediction.

### *C. High-Level Planner Evaluation*

We use Gemini 2.0 Flash as the VLM to perform zero-shot high-level planning. To quantitatively assess the quality of the generated meta-action plans, we report BERTScore, BLEU, and ROUGE-L against ground-truth meta-actions in Table IV. We also evaluate the full pipeline, in which the high-level planner generates meta-action plans that serve as language instructions for the VLA policy. As shown in Table II, using planner-generated meta-actions improves performance over the no-language baseline, although it underperforms compared to manually labeled instructions. This performance gap likely stems from occasional inaccuracies or ambiguities in the zero-shot plans. Detailed results are provided in Appendix A. As




(a) When prompted with "I am carrying a cup of coffee, go slowly", SteerVLA predicts cautious meta actions and executes the turn at a reduced speed.

(b) When prompted with "I am in a rush to get to work!", SteerVLA predicts aggressive behavior and successfully executes a sharp turn.

(c) Our policy can also adapt its behavior on the fly, responding to each of the behaviors described in (a) and (b).

Fig. 6: **Qualitative evaluation in CARLA.** We evaluate SteerVLA across task variations, demonstrating its ability to infer user intent and adapt its behavior accordingly.

| Method | ADE (m) ↓ | | | FDE (m) ↓ | | |
|---|---|---|---|---|---|---|
| | 1s | 2s | 3s | 1s | 2s | 3s |
| SteerVLA w/o lang | 0.45 | 1.08 | 1.97 | 0.67 | 2.13 | 4.33 |
| SteerVLA w/ lang | 0.40 | 0.98 | 1.77 | 0.60 | 1.92 | 3.89 |
| SteerVLA w/ refined lang | **0.39** | **0.96** | **1.75** | **0.59** | **1.90** | **3.86** |
| SteerVLA w/ planner | 0.43 | 1.04 | 1.89 | 0.66 | 2.05 | 4.16 |

TABLE II: **Trajectory prediction accuracy at 1s, 2s, and 3s horizons.** Incorporating language improves prediction accuracy, with refined instructions yielding slightly better ADE/FDE than raw instructions. Though modest, these gains reflect more semantically aligned control behaviors, as supported by human evaluation in Table III. Meta-actions generated by the high-level planner also enhance performance over the no-language baseline, though a gap remains compared to using manually refined instructions.

| Method | All (%) | Turns (%) | Speed Changes (%) |
|---|---|---|---|
| SteerVLA w/o lang | 11/20 | 2 / 20 | 4 / 20 |
| SteerVLA w/ lang | 16/ 20 | 7 / 20 | 9 / 20 |
| SteerVLA w/ refined lang | **18 / 20** | **15 / 20** | **16/20** |

TABLE III: **Human evaluation of instruction adherence across behavior types.** "All" includes 20 uniformly sampled rollouts; "Turns" and "Speed Changes" focus on instructions involving turning or speed modulation. Language improves adherence, with refined instructions yielding the highest alignment.

future work, we plan to fine-tune a VLM such as Gemma 3 [23] to serve as a dedicated high-level planner with improved task grounding.

### D. Closed-Loop Results in the CARLA Simulator

To evaluate the closed-loop capabilities of SteerVLA, we conduct a qualitative analysis in the CARLA simulator, as shown in Fig. 6. We demonstrate that the policy can be effectively steered by user-specified instructions, adapting to changing user preferences in real time and exhibiting diverse behaviors.

We collect 4,000 trajectories in CARLA, each lasting 10–30 seconds. A rule-based annotator assigns one of ten meta-actions to each interval, while CARLA agent parameters are varied to emulate *aggressive*, *normal*, and *cautious* driving styles. The map, weather, and spawn points are randomized to maximize scenario diversity. We follow the same training pipeline used for the refined BDD-X dataset.

For inference, we use Gemini as the high-level planner, which receives the egocentric image stream, persona, current vehicle state (speed, steering, throttle, brake), and dialogue history. The planner is queried for a new meta-action after each action chunk is executed (prompts are provided in Appendix A). We use CARLA's Ackermann control interface to translate the actions into simulator commands.

| Method | BERTScore | BLEU | Rouge-L |
|---|---|---|---|
| Gemini 2.0 Flash | 0.45 | 0.05 | 0.30 |

TABLE IV: Evaluation of the meta-action command quality generated by the high-level planner using BERTScore, BLEU, and ROUGE-L metrics.

## V. Discussion

We present SteerVLA, a hierarchical vision-language-action (VLA) model for autonomous driving that addresses the challenge of generating steerable low-level driving behavior from nuanced, high-level user specifications. By decomposing the problem into a high-level language-based reasoning step and a low-level action generation step—and using structured meta-actions as the interface between them—SteerVLA leverages powerful vision-language model (VLM) priors to interpret behavioral instructions in language space before producing raw control actions.

To train this hierarchical policy, we introduce a novel autolabeling pipeline that generates plausible high-level behavior specifications and meta-action annotations from unlabeled self-driving datasets. This enables SteerVLA to respond effectively to complex, unstructured language prompts, *including those unseen during training*.

**Limitations and Future Work.** While our early results are promising, the current version of SteerVLA has several limitations. First, the quality of autolabeling is constrained by the capabilities of the underlying VLM. Although labeling based on video snippets would be ideal, current VLMs still struggle with dynamic, temporally grounded reasoning compared to static scene understanding. In future work, we aim to bootstrap driving-specific dynamic reasoning capabilities into the labeling pipeline.

Second, the model's flexibility is currently limited by the predefined meta-action space, which serves as the sole interface between the high-level and low-level policies. We plan to investigate training a unified, end-to-end "chain-of-though" policy that jointly models high-level intent and low-level execution.

Lastly, we see an opportunity to incorporate techniques such as reinforcement learning from human feedback (RLHF) to improve the alignment of the high-level planner with user preferences and downstream driving behavior. We hope that future extensions of SteerVLA will build upon these directions to enhance its adaptability and human-aligned decision-making.

## References


[1] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. Paligemma: A versatile 3b vlm for transfer, 2024. URL https://arxiv.org/abs/2407.07726.

[2] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. $\pi_0$: A vision-language-action flow model for general robot control, 2024. URL https://arxiv.org/abs/2410.24164.

[3] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. *arXiv preprint arXiv:2307.15818*, July 2023.

[4] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023. URL https://arxiv.org/abs/2307.15818.

[5] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In *Robotics: Science and Systems XIX*. Robotics: Science and Systems Foundation, July 2023. ISBN 978-0-9923747-9-2.

[6] Danny Driess, Jost Tobias Springenberg, Brian Ichter, Lili Yu, Adrian Li-Bell, Karl Pertsch, Allen Z. Ren, Homer Walke, Quan Vuong, Lucy Xiaoyang Shi, and Sergey Levine. Knowledge insulating vision-language-action models: Train fast, run fast, generalize better, 2025. URL https://arxiv.org/abs/2505.23705.

[7] Zipeng Fu, Tony Z. Zhao, and Chelsea Finn. Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low-Cost Whole-Body Teleoperation. *arXiv preprint arXiv:2401.02117*, January 2024.

[8] Haoxiang Gao, Zhongruo Wang, Yaqian Li, Kaiwen Long, Ming Yang, and Yiqing Shen. A Survey for Foundation Models in Autonomous Driving. *arXiv preprint arXiv:2402.01105*, September 2024.

[9] Noriaki Hirose, Catherine Glossop, Ajay Sridhar, Dhruv Shah, Oier Mees, and Sergey Levine. Lelan: Learning a language-conditioned navigation policy from in-the-wild videos. In *Conference on Robot Learning*, 2024.

[10] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-Oriented Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023.

[11] Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, Yin Zhou, James Guo, Dragomir Anguelov, and Mingxing Tan. EMMA: End-to-End Multimodal Model for Autonomous Driving. *arXiv preprint arXiv:2410.23262*, November 2024.

[12] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization.

[13] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. \$\pi_{0.5}\$: A Vision-Language-Action Model with Open-World Generalization. *arXiv preprint arXiv:2504.16054*, April 2025.

[14] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles, 2018. URL https://arxiv.org/abs/1807.11546.

[15] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An Open-Source Vision-Language-Action Model. In *8th Annual Conference on Robot Learning*, September 2024.

[16] Yi Li, Yuquan Deng, Jesse Zhang, Joel Jang, Marius Memmel, Raymond Yu, Caelan Reed Garrett, Fabio Ramos, Dieter Fox, Anqi Li, Abhishek Gupta, and Ankit Goyal. Hamster: Hierarchical action models for open-world robot manipulation, 2025. URL https://arxiv.org/abs/2502.05485.

[17] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez,

Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

[18] Tran Phong, Haoran Wu, Cunjun Yu, Panpan Cai, Sifa Zheng, and David Hsu. What truly matters in trajectory prediction for autonomous driving? *Advances in Neural Information Processing Systems*, 36:71327–71339, 2023.

[19] Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. Gaia-2: A controllable multi-view generative world model for autonomous driving, 2025. URL https://arxiv.org/abs/2503.20523.

[20] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. DriveLM: Driving with Graph Visual Question Answering. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 256–274, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72943-0.

[21] Laura Smith, Alex Irpan, Montserrat Gonzalez Arenas, Sean Kirmani, Dmitry Kalashnikov, Dhruv Shah, and Ted Xiao. STEER: Flexible Robotic Manipulation via Dense Language Grounding. *arXiv preprint arXiv:2411.03409*, November 2024.

[22] Gemini Team. Gemini: A family of highly capable multimodal models.

[23] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

[24] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo,

You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An Open-Source Generalist Robot Policy. *arXiv preprint arXiv:2405.12213*, May 2024.

[25] Yuping Wang, Shuo Xing, Cui Can, Renjie Li, Hongyuan Hua, Kexin Tian, Zhaobin Mo, Xiangbo Gao, Keshu Wu, Sulong Zhou, Hengxu You, Juntong Peng, Junge Zhang, Zehao Wang, Rui Song, Mingxuan Yan, Walter Zimmer, Xingcheng Zhou, Peiran Li, Zhaohan Lu, Chia-Ju Chen, Yue Huang, Ryan A. Rossi, Lichao Sun, Hongkai Yu, Zhiwen Fan, Frank Hao Yang, Yuhao Kang, Ross Greer, Chenxi Liu, Eun Hak Lee, Xuan Di, Xinyue Ye, Liu Ren, Alois Knoll, Xiaopeng Li, Shuiwang Ji, Masayoshi Tomizuka, Marco Pavone, Tianbao Yang, Jing Du, Ming-Hsuan Yang, Hua Wei, Ziran Wang, Yang Zhou, Jiachen Li, and Zhengzhong Tu. Generative AI for Autonomous Driving: Frontiers and Opportunities. *arXiv preprint arXiv:2505.08854*, May 2025.

[26] Yiran Xu, Xiaoyin Yang, Lihang Gong, Hsuan-Chu Lin, Tz-Ying Wu, Yunsheng Li, and Nuno Vasconcelos. Explainable object-induced action decision for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9523–9532, 2020.

[27] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee. K. Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model, 2024. URL https://arxiv.org/abs/2310.01412.

[28] Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*, 2024.

[29] Xingcheng Zhou, Xuyuan Han, Feng Yang, Yunpu Ma, and Alois C. Knoll. OpenDriveVLA: Towards End-to-end Autonomous Driving with Large Vision Language Action Model. *arXiv preprint arXiv:2503.23463*, March 2025.

# APPENDIX

Listing 1: BDDX refinement prompt.

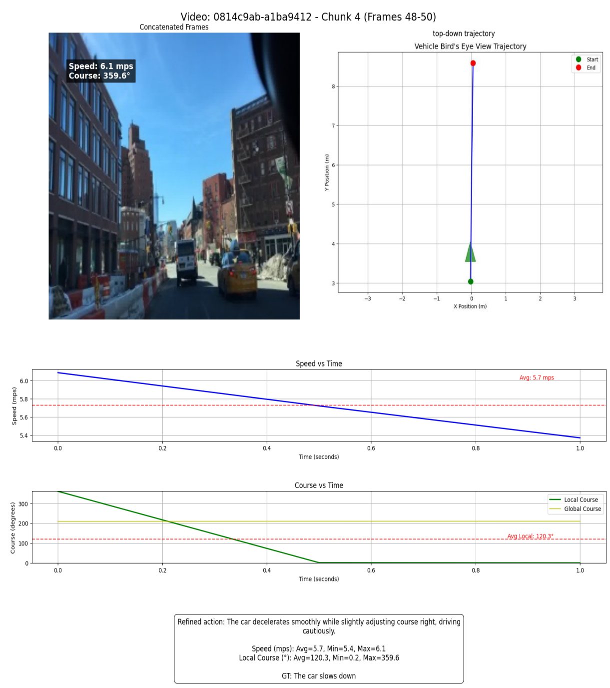
```
# Driving Behavior Refinement Prompt

You are an expert in vehicle dynamics and driving
    behavior analysis. Your task is to interpret
    natural language descriptions of driving
    behavior by analyzing vehicle ego state data (
    speed and course over time). Your response must
    include two parts:

1. **Ego State Analysis** - a brief explanation of
    observed speed and course trends over time.
2. **Refined Driving Behavior Description** - a more
     specific version of the original description,
    enhanced with motion extent and driving style.

You are an expert in vehicle dynamics and driving
    behavior analysis. Your task is to interpret and
     refine natural language descriptions of driving
     behavior by analyzing vehicle ego state data (
    speed and course over time) to produce a **
    precise and nuanced behavior summary**. Your
    output should describe:

1. **Ego State Analysis** - a brief explanation of
    observed speed and course trends over time.
2. **Refined Driving Behavior Description** - a more
     specific version of the original description,
    enhanced with a meaningful modifier  _(e.g., **
    smooth turning**, **wide turn**, **abrupt stop
    **, **steady lane keeping**)_ and a **driving
    style**, reflecting the driver’s attitude or
    intent
   _(e.g., **cautiously**, **normally**, **
       aggressively**)_

---

## Input Format

**Driving Description:**
INSERT_BEHAVIOR_DESCRIPTION

**Ego Vehicle States:**
INSERT_EGO_STATE_SEQS

These ego states reflect how the vehicle moved
    during the described behavior.

> **Note:**
> - **Course increasing** --> vehicle is turning **
    right**
> - **Course decreasing** --> vehicle is turning **
    left**

---

## Output Guildlines

Your response should contain two sections:

### 1. Ego State Analysis

Analyze the speed and course sequence:
- Describe speed patterns: Is the vehicle
    accelerating, decelerating, or maintaining speed
    ?
- Describe course patterns: Is the vehicle turning
    sharply, smoothly, or going straight?
- Mention time duration and total changes in course
    or speed.
```

```
-

### 2. Refined Driving Behavior Description

Produce a single, natural-language sentence that:
- Refines the driving description with motion extent
     (e.g., *smooth*, *sharp*, *wide*, *slight*)
- Adds driving style (e.g., *cautiously*, *normally
    *, *aggressively*)
- Grounding the refinement in the observable
    patterns of the ego vehicle states

---

## Notes

- The refined description must not exceed **20 words
    **.
- Use **speed trends** to judge acceleration or
    deceleration patterns.
- Use **course change patterns** to assess turning
    sharpness or trajectory smoothness.
- If the style cannot be confidently inferred,
    default to **"normally"**.
- Use **natural, human-readable language**-avoid
    unnecessary technical jargon.

---
## Output Format (REQUIRED)

Respond **only** with a valid JSON object in the
    following structure (do not include any other
    text outside the JSON block):

'''json
{
  "ego_state_analysis": "<Short paragraph analyzing
      speed and course trends>",
  "refined_description": "<One complete sentence
      with refined behavior and driving style within
       20 words>"
}
''',
```

Listing 2: Example High-level VLM planner prompt.

```
Prompt:
You are an autonomous driving assistant. Your task
    is to generate a driving behavior plan based on:
A front-view camera image
The current speed of the vehicle
A high-level driving command (e.g., move forward,
    stop, turn left, turn right)
A persona describing the driver's intent or external
     conditions (e.g., cautious driving due to rain)
Inputs:
Image: See Fig. 5
High-level command: turn left
Persona: It's snowing, so I'm being careful to avoid
     slipping.
Output:
Produce a driving behavior plan (no more than 20
    words) that includes:
Speed behavior - Will the vehicle accelerate,
    maintain speed, or decelerate?
Heading behavior - Describe the expected heading
    change (e.g., continue straight, turn slightly
    right, make a sharp left).
Driving style - Reflect the persona (e.g.,
    cautiously, smoothly, assertively).
Respond with a single natural language sentence
    summarizing the driving behavior.
Example Output:
"The car decelerates smoothly and prepares for a
```


(a) Starting with the label "The car accelerates slowly", we can augment with additional information from the vehicle's states to get the label "The car rolls through the stop sign with a slight right turn, accelerating gradually, driving normally."




(b) Starting with the label "The car slows down", we can augment with additional information from the vehicle's states to get the label "The car decelerates smoothly while slightly adjusting course right, driving cautiously."

Fig. 7: Examples of refining the BDD-X labels to train a more steerable low-level policy.

```
    slight right turn, driving normally.",
```

Listing 3: Example High-level VLM planner output.

```
Output: ``The car will cautiously decelerate, making
     a slow, wide left turn due to snowy conditions
    .'',
```

where the ground truth is “The car makes a smooth left turn, decelerating then accelerating, with normal driving style.”,

Fig. 8: The input image of the high-level planner.

Listing 4: Persona generation prompt

```
# Driving Behavior Interpretation Prompt

You are an expert in interpreting driving behavior.
    Given a natural language description of a
    vehicle's behavior, extract two things:

1. **High-Level Command**  select one of the
    following discrete options:
   - `Move forward`
   - `Stop/Slow down`
   - `Turn left`
   - `Turn right`

2. **Persona**  write a vivid, one-sentence first-
    person description of the driver's likely
    motivation or situation. The persona should
    reflect the internal reasoning or external
    circumstances influencing how they drive. Use
    natural language that includes emotional or
    situational cues (e.g., urgency, responsibility,
     distractions, time pressure, purpose of the
    trip). Avoid generic or purely factual
    statements-make the driver feel like a real
    person in a specific moment.

3. **Reasoning**  Provide a brief explanation
    connecting the driving behavior description, the
     dashcam view, your selected persona, and your
    chosen high-level instruction. Explain how these
     elements logically support each other.


---

## Notes

### For Persona:
- The persona must be plausible based on both the
    actions taken by the vehicle and the
    surroundings of the vehicle.
- Otherwise, if it is not definitive whether the
    surroundings fit the description (e.g. the
    behavior describes a baby in the car, but a baby
     would not be visible from a dashcam), the
    option is fine to propose.
- The persona must align with the style (e.g.
    aggressive, cautious, normal) of the driving
    description.
- The persona must differ from the examples of
    possible personas.
- The persona must provide a long-horizon reason for
     the car behavior over its whole trajectory (and
     therefore must NOT be dependent on things like
    stop signs, traffic lights)
- Assume that the driver is experienced.
- The persona should describe a legal scenario.
    However, do not include any legal jargon or
    references to the law in the language of the
    persona.

### For High-Level Instruction:
- For the high level command, base your selection **
    only** on the textual driving description and
    the dashcam view.
- For the high level command, turning is defined as
    a full turn at intersections.
- If the car is moving leftward or rightward because
     it is simply following a curve in the road or
    slightly adjusting within the lane, this should
    be categorized as either moving forward or
    slowing/stopping.
- Possible explanations for a car moving forward
    include "Traffic light is green", "Follow
    traffic", and "Road is clear".
- Possible explanations for a car stopping/slowing
    include "Traffic light", "Traffic sign", "
    Obstacle ahead"
- Possible explanations for a car turning left
    include "On the left-turn lane", and "Traffic
    light allows"

### For Reasoning:
- Connect the behavior description, dashcam visual
    elements, persona motivation, and instruction
    choice
- Explain how the persona logically leads to the
    observed driving behavior
- Reference specific elements from both the text
    description and visual scene


---

## Input Format
Driving Behavior Description: <description here>

---

## Output Format
Router Command: <one of: move forward | stop | turn
    left | turn right>
Persona: <one-sentence persona in first person>
Reasoning: <brief explanation connecting behavior,
    image, persona, and instruction>

---

## Examples of possible personas
I'm trying to avoid slipping because the weather
    conditions are not the best for driving.
The car is driving on an open road, so I am speeding
     quickly through the streets.
I'm an uber driver and my passenger is prone to
    carsickness.
My wife is giving birth, so I'm trying to get to the
     hospital as quickly as possible.
```

```
It’s 8:55 AM and I’m going to be late for a very
    important meeting.
My baby is sleeping in the back seat, and I’m
    driving gently so that I don’t wake them up.
There are many pedestrians around, so I’m making
    sure to drive carefully.
I am going to be on the highway for a while, so I’d
    like to use the leftmost lane.

---

## Now, process the following:
Driving Behavior Description: {refined_annotation}
```