GovernorVLA: Adaptive Embodied Reasoning for Long-tail Autonomous Driving

Matthew Foutter[1], Jadelynn Dao[2,*], Yunshan Wang[3,*], Daniele Gammelli[4], Marco Pavone[4,5]

***Abstract*—Foundation Models (FMs) offer promise to ground Autonomous Vehicle (AV) perception and decision-making in broad world-knowledge, bridging the limitations of in-domain data through zero-shot generalization to the long-tail of operational experience. *Embodied reasoning*, wherein a vision-language policy articulates a structured and intelligible Chain-of-Thought [1] before an embodied decision, has been shown to enhance policy efficacy [2], in particular, on the hardest decisions [3]. However, two challenges limit embodied reasoning in mature end-to-end autonomy stacks: 1) Chain-of-Thought reasoning is computationally expensive, requiring thousands of tokens to terminate [4], and 2) indiscriminate, laborious reasoning is unnecessary and may be detrimental in mundane decisions [5]. In this work, we train an open-source Vision-Language Action (VLA) model to reason over difficult driving decisions — identified by high trajectory error from our base policy — reducing mean Average Displacement Error (ADE) by 19.1% compared to the base policy in the long-tail of the vehicle's experience. We present a hybrid policy composed of our base and reasoning-based policy by learning a governor from the base policy's latent space that anticipates failure with an AUROC of 0.86 on a withheld driving environment, routing to the reasoning-based policy to reduce mean ADE by 6.4% and 0.5% on the long-tail and whole vehicle's experience, respectively, at the cost of adding 4.4% to mean latency.**

***Index Terms*—Vision-Language Action, Embodied Reasoning**

I. Introduction

Autonomous Vehicles (AV) have made tremendous progress towards large-scale deployment in recent years e.g., Waymo One [6], through the careful integration of Artificial Intelligence (AI) into the autonomy stack. In light of the nominal performance enabled by these data-driven algorithms, a persistent challenge is endowing these systems with the capability to demonstrate trustworthy behaviors in the *long-tail* of the vehicle's operational experience [7] where data is sparse and expert supervision is limited. Foundation Models (FMs) [8], namely, Vision-Language Models (VLMs) [9]–[11], have emerged as a promising prior for generalist robot policy learning in low-data regimes given the strong semantic grounding and vision-language alignment inherited through self-supervised training on a corpus of human knowledge. More recently, the robotics community has explored scaling the computational budget of VLMs, eliciting an intermediate Chain-of-Thought (CoT) — herein referred to as *embodied reasoning* — effectively increasing the model's depth [12] before predicting an embodied decision. In practice, this operational paradigm improves downstream policy efficacy by e.g., exploring different solution strategies [4], [13] and enabling a step-by-step reasoning framework [1], drawing inspiration from human decision-making processes. Nevertheless, the conclusive performance gains observed through embodied reasoning are countervailed by a latency penalty at inference due to the auto-regressive sampling procedure underlying State-of-the-Art (SoTA) FMs.

Therefore, guided by prior research in FMs demonstrating that reasoning provides the strongest gain on challenging queries [3], [5], [15]–[18] and the associated latency required at inference, we aim to endow an autonomous navigation agent with embodied reasoning capabilities at critical decision points where a straightforward, quick-thinking architecture is likely to struggle. We propose to use the offline trajectory error of a base Vision-Language Action (VLA) policy as a measure of task difficulty, suggesting the existence of a non-trivial decision. Online, we learn a routing algorithm to preemptively catch failure from the base policy and activate embodied reasoning for intelligible and trustworthy performance gains, observing generalization to new driving environments and unseen road conditions. Concretely, we offer two key contributions:

1) We filter mundane state transitions from a robotics dataset using a scalable metric, namely, the Average Displacement Error (ADE), which doubles as a supervision signal to trigger embodied reasoning at inference. With these high-impact key-frames, we collect CoT annotations from a SoTA VLM [19] augmented with domain priors and train an open-source VLM [14] to reason over environmental context before committing to a predicted trajectory offering 19.1% reduction in mean ADE among difficult driving scenarios compared to the base policy.
2) We train a 3-layer MLP to recognize the signature of failure from the latent space of the base policy, achieving an AUROC of 0.86 on failure detection in a withheld driving environment. Online, we use the routing algorithm to govern adaptive reasoning, observing a 6.4% decrease in mean ADE in the *long-tail* of the robot's experience and a 0.5% decrease in mean ADE against the base policy across the full population of unseen road conditions.

II. Related Work

Reasoning Triggers in Robotics: Identifying when embodied reasoning is most beneficial requires scoring data by difficulty; existing literature filters data according to diversity [20]–[22] or influence [23], [24]. Alpamayo-R1 [3] selects auto-labeling instances where the vehicle performs a novel maneuver from ego-motion history, though these moments are difficult to identify at inference where only past motion is available. Most similar to our approach, CounterfactualVLA [5] isolates difficult scenes via the information gain of ground-truth

[1] Dept. of Mechanical Engineering, Stanford University. [2] Dept. of Computer Science, Stanford University. [3] Dept. of Engineering, Stanford University. [4] Dept. of Aeronautics and Astronautics, Stanford University. [5] NVIDIA Corp. [*] Equal authorship. Contact: {mfoutter, jadelynn, jerryw, gammelli, pavone}@stanford.edu

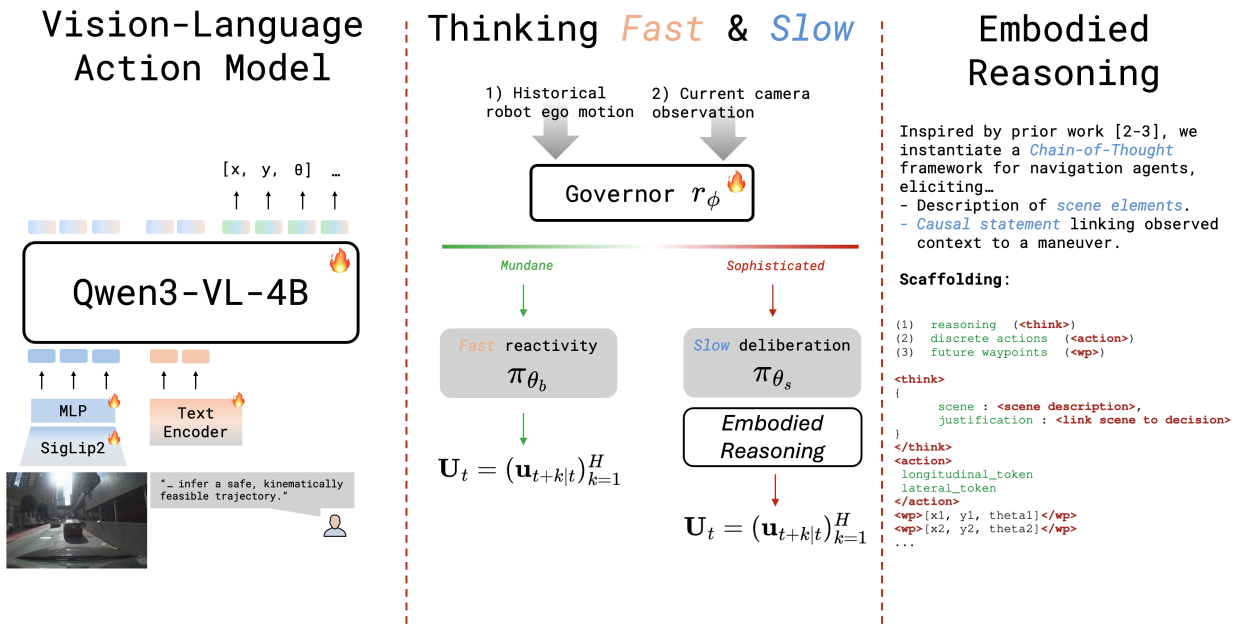

Fig. 1. **GovernorVLA Architecture Overview**. (*Left*) The base policy π_{θ_b} is instantiated as a fine-tuned Qwen3-VL-4B Vision-Language Action model [14], consuming a forward-facing camera observation and historical ego-motion to predict future relative waypoints decoded as language tokens. (*Center*) At each inference timestep, the terminal embedding of π_{θ_b} — where visual context and navigation intent are fused prior to trajectory generation — is consumed by a learned governor r_ϕ that classifies the driving decision as mundane or sophisticated, governing the activation of fast policy π_{θ_b} or the slow reasoner π_{θ_s}, respectively, akin to a Watt governor on a steam engine sensing load and adjusting the operating regime maintain stable performance. (*Right*) The slow reasoning policy π_{θ_s} elicits an embodied Chain-of-Thought structured as a scene description and causal move justification before committing to a predicted trajectory, improving trustworthiness in the long-tail of the vehicle's experience at the cost of additional inference latency.

navigation guidance, training a VLA to implicitly recognize when self-corrective reasoning improves trajectory error, akin to AutoVLA [25] and AdaThinkDrive [26]. However, each of these methods commits to a single operating point determined at training time, offering no mechanism to trade off false activations against missed detections at inference. We address this by training an explicit governor as a separable module, whose scalar output admits post-hoc threshold calibration, evaluating generalization through precision and recall on a withheld driving environment.

Routing Among Specialist Policies: The deployment of generalist robot policies in a real-time environment necessitates a balance between predictive accuracy and inference latency. One potential path forward in the community is to scale the policy's computational budget with query difficulty. Existing literature confronts rare, sophisticated decisions by delegating control to a conservative fallback policy by e.g., utilizing the latent embedding space of FMs or the model's own uncertainty to detect anomalies [27]–[29]. Our work builds on this paradigm by preemptively detecting failure and routing to a high-capacity navigation policy, attempting to improve closed-loop performance without sacrificing task completion.

Recent efforts in adaptive reasoning outside of robotics focus on model and inference strategy selection from a heterogeneous pool [30]–[32] to improve response quality. Most similar to our approach, [33]–[36] route queries between models of varying capacity to balance e.g., user preferences or answer correctness, and the financial cost associated with querying closed-source API models. However, in embodied applications, such as autonomous driving, dynamic multi-agent interactions demand real-time responsiveness *and* trustworthiness – creating a natural tension that is most acute in rare, safety-critical scenarios. We build on this emerging paradigm in closed-loop control by instantiating a hybrid policy with two fine-tuned 4B parameter VLAs employing different reasoning strategies, where routing is governed by anticipating trajectory difficulty and inference latency, rather than API expenditures.

III. Problem Formulation

In this work, we consider an autonomous navigation agent operating with discrete-time dynamics:

$$\mathbf{x}_{t+k} = f(\mathbf{x}_{t+k-1}, \mathbf{u}_{t+k|t}), \quad k = 1, \dots, H \tag{1}$$

where $\mathbf{x}_t \in \mathcal{X} \subseteq \mathbb{R}^n$ and $\mathbf{u}_t \in \mathcal{U} \subseteq \mathbb{R}^m$ are the state and control input, respectively, and H denotes the planning horizon. Concretely, the state is $\mathbf{x}_t = (x_t, y_t, \theta_t) \in \mathcal{X} \subseteq \mathbb{R}^3$, comprising longitudinal position, lateral position, and heading. The agent outputs navigation commands through a sequence of H relative waypoints over a fixed horizon $T_h = H \cdot t_s$, with $t_s = 0.25\,\text{s}$ and $H = 24$, giving $T_h = 6\,\text{s}$:

$$\mathbf{U}_t = \left(\mathbf{u}_{t+1|t},\ \mathbf{u}_{t+2|t},\ \dots,\ \mathbf{u}_{t+H|t}\right) \in \mathcal{U}^H, \tag{2}$$

where $\mathbf{u}_{t+k|t} = (\Delta x_{t+k|t},\ \Delta y_{t+k|t},\ \Delta\theta_{t+k|t}) \in \mathcal{U} \subseteq \mathbb{R}^3$ is the relative waypoint k steps ahead, expressed in the body frame of $\mathbf{x}_t$ corresponding to time $t + k\,t_s$. We assume access to a dataset of expert navigation demonstrations collected across two geographic regions i.e., $\mathcal{D} = \mathcal{D}_{\text{US}} \cup \mathcal{D}_{\text{DE}}$, where each regional subset $\mathcal{D}_r = \{\tau^{(i)}\}_{i=1}^{|\mathcal{D}_r|}$, $r \in \{\text{US}, \text{DE}\}$, consists of trajectories of the form:

$$\tau^{(i)} = \left(o_t^{(i)},\ \mathbf{x}_t^{(i)}\right)_{t=1}^{T},$$

where $o_t^{(i)} \in \mathbb{R}^{H_{\text{img}} \times W_{\text{img}} \times C}$ is a forward-facing RGB camera image and $T = T_f \cdot f$ is the trajectory length, with final time $T_f = 20\,\text{s}$ and sample frequency $f = 30\,\text{Hz}$. Our objective is to derive a policy π_θ capable of trustworthy navigation in the *long-tail* of the robot's experience as measured by the open-loop ADE relative to ground-truth, without incurring excessive latency on nominal inputs.

IV. Proposed Hybrid Policy

We decompose the inference policy into two modes: a *fast* baseline policy π_{θ_b} and a *slow* reasoning architecture π_{θ_s}, inspired by the "Fast" and "Thinking" variants of common agentic FMs [37], [38]. The two modes may share weights ($\theta_b = \theta_s$) or maintain independent parameters ($\theta_b \neq \theta_s$), depending on the deployment setting. The two modes also share the same environmental context, allowing us to route instantaneously between each mode at inference by conditioning auto-regressive generation on a natural language instruction, e.g., l_b activates the baseline and l_s activates the reasoning architecture. At each inference timestep t, the policy conditions on the current camera image o_t and a history of $H_{\text{past}} = \lfloor T_{\text{hist}}/t_s \rfloor + 1$ relative waypoints, where $T_{\text{hist}} = 2\,\text{s}$:

$$\mathcal{I}_t = \left(o_t,\ \left(\mathbf{u}_{t-k|t} \right)_{k=0}^{H_{\text{past}}} \right). \tag{3}$$

Fig. 1 provides an overview of the proposed architecture. At each timestep, the shared environmental context $\mathcal{I}_t$ is passed to the fast policy π_{θ_b}, which produces an embedding from the final layer corresponding to the last input token. The embedding is consumed by the routing algorithm r_ϕ, which determines whether the fast policy's prediction is sufficient or whether the slow reasoning policy π_{θ_s} should be invoked to produce an embodied chain-of-thought $\mathbf{z}$ before waypoint generation. The following subsections detail each component in turn: the fast policy (§IV-A), the slow reasoning policy (§IV-B), the routing algorithm (§IV-C), and the data curation pipeline that supervises both (§IV-D).

A. Thinking Fast

We directly sample future relative waypoints for low-level execution from the *fast* baseline policy, an open-source VLM [14] fine-tuned to predict action maneuvers, conditioned on the input context and a basic navigation instruction:

$$\mathbf{U}_t \sim \pi_{\theta_b}\left(\cdot \mid \mathcal{I}_t,\, l_b \right), \quad \mathbf{U}_t = (\mathbf{u}_{t+k|t})_{k=1}^{H}, \tag{4}$$

where the waypoints are decoded directly as language tokens [39], [40], but importantly, our algorithm does not preclude alternative decoding strategies [3], [41].

B. Thinking Slow

We instantiate the *slow* reasoning policy by post-training the baseline policy with embodied chain-of-thought priors to justify the vehicle's maneuver, enhancing downstream action prediction. Explicitly, at inference we sample the embodied chain-of-thought and relative waypoints in series from the reasoning policy π_{θ_s} conditioned on the input context and a reasoning-based navigation instruction:

$$\mathbf{z} \sim \pi_{\theta_s}\left(\cdot \mid \mathcal{I}_t,\, l_s \right), \tag{5}$$

$$\mathbf{U}_t \sim \pi_{\theta_s}\left(\cdot \mid \mathcal{I}_t,\, l_s,\, \mathbf{z} \right), \quad \mathbf{U}_t = \left(\mathbf{u}_{t+k|t} \right)_{k=1}^{H}, \tag{6}$$

where $\mathbf{z}$ articulates a scene description and justification for the selected move statement in natural language.

C. The Governor: Anticipating Critical Decisions

In order to dynamically route between the *fast* baseline and the *slow* reasoning policy, we learn a routing algorithm r_ϕ as a shallow Multi-Layer Perceptron (MLP) on the final-layer embedding of the last input token provided to π_{θ_b}:

$$r_\phi(\mathcal{I}_t, l_b) = \sigma\Big(f_\phi(h_{\text{last}}^{(L)}(\mathcal{I}_t, l_b)) \Big) \in [0, 1], \tag{7}$$

where $h_{\text{last}}^{(L)}(\mathcal{I}_t, l_b) \in \mathbb{R}^d$ is the final transformer layer's embedding representation at sequence length L i.e., right before auto-regressive generation begins, and $f_\phi : \mathbb{R}^d \to \mathbb{R}$ is a 3-layer MLP. Then, we operationalize this binary classifier by calibrating a decision boundary $\alpha \in [0, 1]$ on a withheld evaluation set, isolating mundane and sophisticated decisions in a previously unseen environment, which can be used to shape the model's output distribution and associated prediction latency:

$$\mathbf{U}_t \sim \begin{cases} \pi_{\theta_b}(\cdot \mid \mathcal{I}_t,\, l_b), & r_\phi \leq \alpha, \\ \pi_{\theta_s}(\cdot \mid \mathcal{I}_t,\, l_s,\, \mathbf{z}),\ \mathbf{z} \sim \pi_{\theta_s}(\cdot \mid \mathcal{I}_t,\, l_s), & r_\phi > \alpha. \end{cases} \tag{8}$$

D. Reasoning & Routing Algorithm Data Curation

The offline trajectory error of the baseline policy on in-distribution decisions defines a spectrum along which we hypothesize routine decisions tend to have low error and conversely sophisticated decisions tend to have high error. Therefore, for each trajectory in our training dataset $\tau^{(i)} \sim \mathcal{D}_{\text{US}}$, we augment each decision moment with the ADE from the base policy against ground-truth:

$$\tau^{(i)} = \left(o_t^{(i)},\, \mathbf{x}_t^{(i)} \right)_{t=1}^{T} \longrightarrow \tilde{\tau}^{(i)} = \left(o_t^{(i)},\, \mathbf{x}_t^{(i)},\, \epsilon_t^{(i)} \right)_{t=1}^{T}, \tag{9}$$

where $\epsilon_t^{(i)} = \frac{1}{H} \sum_{k=1}^{H} \left\| \hat{\mathbf{u}}_{t+k|t}^{(i),xy} - \mathbf{u}_{t+k|t}^{(i),xy} \right\|_2$ with $\mathbf{u}^{xy} = (\Delta x, \Delta y) \in \mathbb{R}^2$ denoting the positional component of the waypoint (heading $\Delta\theta$ is excluded), $\hat{\mathbf{U}}_t^{(i)} \sim \pi_{\theta_b}(\cdot \mid \mathcal{I}_t^{(i)})$ the predicted waypoints and $\mathbf{U}_t^{(i)}$ the ground-truth waypoints over the H-step horizon.

We isolate the *long-tail* of the robot's experience by thresholding nominal experience according to the 90th percentile of error, creating a subset of operating conditions ripe for additional CoT annotations; we auto-label the resultant top 10% of error cases with embodied CoT derived from a closed-source model [19] augmented with domain priors populating each decision with a scene description and move justification, inspired by ECoT and Alpamayo-R1 [2], [3].

Separately, we construct a binary classification dataset pairing the terminal embedding with an associated class label indicating whether the decision requires embodied reasoning at inference:

$$y_t^{(i)} = \begin{cases} 1 & \text{if } \epsilon_t^{(i)} \geq \epsilon_{90}, \\ 0 & \text{otherwise}, \end{cases} \tag{10}$$

where $\epsilon_{90} = \text{percentile}_{90}\Big(\{\epsilon_t^{(i)}\} \Big)$ is the 90th percentile of ADE computed across all time-steps and trajectories in $\mathcal{D}_{\text{US}}$. The classification dataset is then:

$$\mathcal{D}_{\text{US, cls}} = \left(h_{\text{last}}^{(L)}\left(\mathcal{I}_t^{(i)} \right),\ y_t^{(i)} \right)_{i,\,t}, \tag{11}$$

which is used to train the routing algorithm to identify the signature characteristics of high error from the base policy.

TABLE I

EFFICACY OF VARIOUS ROUTING STRATEGIES ON THE GERMANY TRANSFER SET (n=26,371). WE HIGHLIGHT THAT ADE_{90} IS EVALUATED ON THE ORACLE'S ACTIVATION SUBSET AND PARENTHETICAL VALUES DENOTE RELATIVE CHANGE VERSUS π_{θ_b}. BEST IN CLASS IS BOLDED.

	ADE (m) ↓			ADE_{90} (m) ↓			
Routing Strategy	Mean	Median	P90	Mean	Median	Latency (s) ↓	Activation (%)
π_{θ_b} only	4.15	3.15	8.64	11.10	9.75	3.53 ± 0.15	0.00
π_{θ_s} only	4.52 (+8.9%)	3.56 (+13.0%)	9.17 (+6.2%)	8.98 (–19.1%)	8.41 (–13.7%)	4.63 ± 0.58 (+31.2%)	100.00
Random	4.20 (+1.3%)	3.20 (+1.8%)	8.72 (+0.9%)	10.80 (–2.7%)	9.63 (–1.2%)	3.68 ± 0.46 (+4.4%)	14.00
Mahalanobis	4.25 (+2.3%)	3.28 (+4.0%)	8.73 (+1.1%)	10.39 (–6.4%)	9.42 (–3.4%)	3.96 ± 0.64 (+12.4%)	39.88
AESOP [27]	4.21 (+1.5%)	3.21 (+1.8%)	8.75 (+1.3%)	10.79 (–2.8%)	9.62 (–1.4%)	**3.67 ± 0.45 (+4.0%)**	12.89
Ours	**4.13 (–0.5%)**	**3.14 (–0.4%)**	**8.68 (+0.5%)**	**10.38 (–6.4%)**	**9.43 (–3.3%)**	3.68 ± 0.46 (+4.4%)	14.02
Oracle	3.85 (–7.2%)	3.05 (–3.0%)	7.30 (–15.5%)	8.98 (–19.1%)	8.41 (–13.7%)	3.68 ± 0.46 (+4.4%)	14.05

V. EXPERIMENTS & DISCUSSION

With a hybrid policy π_θ governed by our learned routing algorithm r_ϕ, we benchmark our approach against alternative routing criteria to ablate key design choices. In the following section, we describe the evaluation dataset, associated baseline algorithms and compute infrastructure. Then, we analyze our algorithm's performance in Table I with discussion to explore the research question:

RQ *What supervision signal is most effective for governing adaptive reasoning?*

A. Experimental Setup

We evaluate π_θ and Governor r_ϕ on the withheld environment D_{DE}, with 27k driving decisions across 2250 scenes, to characterize out-of-distribution generalization. We baseline against: the base policy π_{θ_b} and reasoner π_{θ_s} applied naively at every decision; two self-supervised anomaly detectors, AESOP [27] and Mahalanobis distance, calibrated to a 5% false positive rate on $\mathcal{D}_{\text{US}}$ using scene embeddings and the base policy's terminal embedding, respectively; an oracle with ground-truth trajectory error; and random activation at 14%, matching the oracle rate. All experiments are conducted on a GeForce RTX 5090 GPU with the vLLM inference package.

B. Experimental Results & Discussion

Trajectory error, latency and activation rate statistics for the hybrid policy on $\mathcal{D}_{\text{DE}}$ are reported in Table I, alongside isolated statistics on the subset of decisions exceeding the calibrated US 90th percentile ADE threshold.

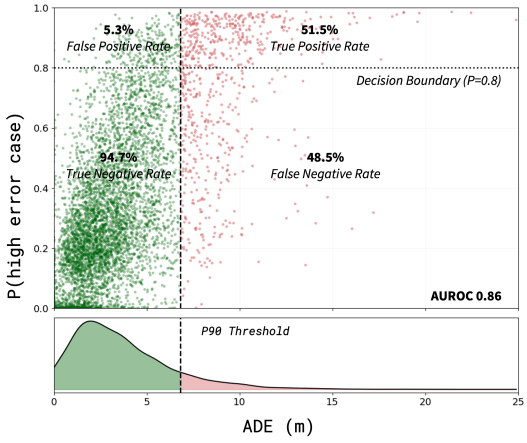

Fig. 2. **Router Generalization**. Outlier probability predicted from π_{θ_b} embeddings vs. ground-truth ADE on the Germany transfer set, demonstrating reliable discrimination between nominal and long-tail driving scenarios.

An oracle governor establishes the theoretical ceiling in Table I: activating the reasoner on only difficult decisions achieves a 7.2% reduction in mean ADE at a 4.4% latency cost (14% activation rate). While randomly routing to the reasoner at the same activation rate erases this gain — producing a 1.3% *increase* in mean ADE — largely due to the base policy's relative strength in straightforward scenarios. Both self-supervised anomaly detectors — AESOP [27] and the Mahalanobis distance — support a supervised learning approach: neither the scene embeddings nor the base policy's latent geometry naturally separates mundane and sophisticated decisions, causing performance gains in the long-tail to be diluted by false positives.

In practice, our governor r_ϕ — trained on the base policy's latent space, where environmental context and navigation intent are naturally fused prior to trajectory generation — is the only strategy to simultaneously reduce mean ADE on both the long-tail and full population, recovering a 6.4% and 0.5% reductions, respectively, at a 4.4% latency cost (14% activation rate) compared to π_{θ_b}. We find that the baseline algorithms degrade population ADE through spurious activations, underscoring the strength of a supervised router in the policy's latent-space over self-supervised anomaly detection for judicious adaptive reasoning. As visualized in Figure 2, the governor achieves an AUROC of 0.86 on $\mathcal{D}_{\text{DE}}$ with a 51.5% true positive rate at a 5.3% false positive rate — a deliberate operating point that prioritizes precision over recall, as spurious activations of the reasoning policy are particularly costly given the base policy's strength and the abundance of mundane decisions across the full population.

VI. CONCLUSIONS

We presented GovernorVLA, a hybrid navigation policy that routes between a fast base policy and a slow reasoning policy by anticipating trajectory difficulty from the base policy's terminal embedding. Trained entirely on US driving data, the governor generalizes to a withheld German driving environment, achieving a 6.4% reduction in mean ADE on difficult decisions and a 0.5% reduction across the full population at a modest 4.4% latency overhead — suggesting that a policy's own latent geometry is a reliable and deployable proxy for task difficulty in the long-tail of real-world driving experience.

REFERENCES

[1] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *arXiv preprint arXiv:2201.11903*, 2022.

[2] M. Zawalski, W. Chen, K. Pertsch, O. Mees, C. Finn, and S. Levine, "Robotic control via embodied chain-of-thought reasoning," *arXiv preprint arXiv:2407.08693*, 2024.

[3] NVIDIA and Y. W. et al., "Alpamayo-R1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail," *arXiv preprint arXiv:2511.00088*, 2025.

[4] D. Guo, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu *et al.*, "Deepseek-r1 incentivizes reasoning in llms through reinforcement learning," *Nature*, vol. 645, no. 8081, pp. 633–638, 2025.

[5] Z. Peng, W. Ding, Y. You, Y. Chen, W. Luo, T. Tian, Y. Cao, A. Sharma, D. Xu, B. Ivanovic, B. Li, B. Zhou, Y. Wang, and M. Pavone, "Counterfactual VLA: Self-reflective vision-language-action model with adaptive reasoning," *arXiv preprint arXiv:2512.24426*, 2025.

[6] L. Di Lillo, T. Gode, X. Zhou, M. Atzei, R. Chen, and T. Victor, "Comparative safety performance of autonomous- and human drivers: A real-world case study of the waymo driver," *Heliyon*, vol. 10, no. 14, p. e34379, 2024.

[7] R. Xu *et al.*, "Wod-e2e: Waymo open dataset for end-to-end driving in challenging long-tail scenarios," *arXiv preprint arXiv:2510.26125*, 2025.

[8] R. B. et al., "On the opportunities and risks of foundation models," *ArXiv*, 2021. [Online]. Available: https://crfm.stanford.edu/assets/report.pdf

[9] S. B. et al., "Qwen3-vl technical report," *ArXiv*, vol. abs/2511.21631, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:283262018

[10] A. A. et al., "The llama 4 herd: Architecture, training, evaluation, and deployment notes," *ArXiv*, vol. abs/2601.11659, 2026. [Online]. Available: https://api.semanticscholar.org/CorpusID:284910371

[11] A. R. et al., "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:231591445

[12] G. Feng, B. Zhang, Y. Gu, H. Ye, D. He, and L. Wang, "Towards revealing the mystery behind chain of thought: A theoretical perspective," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://openreview.net/forum?id=qHrADgAdYu

[13] J. Hu, Y. Zhang, Q. Han, D. Jiang, X. Zhang, and H.-Y. Shum, "Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. [Online]. Available: https://openreview.net/forum?id=NFM8F5cV0V

[14] S. B. et al., "Qwen3-vl technical report," *arXiv preprint arXiv:2511.21631*, 2025.

[15] T. L. et al., "Measuring faithfulness in chain-of-thought reasoning," *ArXiv*, vol. abs/2307.13702, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259953372

[16] N. Muennighoff *et al.*, "s1: Simple test-time scaling," *arXiv preprint arXiv:2501.19393*, 2025.

[17] X. Chen, J. Xu, T. Liang, Z. He, J. Pang, D. Yu, L. Song, Q. Liu, M. Zhou, Z. Zhang, R. Wang, Z. Tu, H. Mi, and D. Yu, "Do NOT think that much for 2+3=? on the overthinking of long reasoning models," in *Forty-second International Conference on Machine Learning*, 2025. [Online]. Available: https://openreview.net/forum?id=MSbU3L7V00

[18] T. e. a. Ye, "Limo: Less-is-more reasoning via difficulty-gated chain-of-thought annotation," in *Conference on Language Modeling (COLM)*, 2025. [Online]. Available: https://arxiv.org/abs/2502.03387

[19] Google DeepMind, "Gemini 3 pro model card," https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf, December 2025, model card.

[20] J. Hejna, C. A. Bhateja, Y. Jiang, K. Pertsch, and D. Sadigh, "Remix: Optimizing data mixtures for large scale imitation learning," in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=fIj88Tn3fc

[21] J. Hejna, S. Mirchandani, A. Balakrishna, A. Xie, A. Wahid, J. Tompson, P. R. Sanketi, D. Shah, C. Devin, and D. Sadigh, "Robot data curation with mutual information estimators," *ArXiv*, vol. abs/2502.08623, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:276287599

[22] Y. Zhang, Y. Xie, H. Liu, R. Shah, M. Wan, L. Fan, and Y. Zhu, "Scizor: Self-supervised data curation for large-scale imitation learning," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

[23] C. Agia, R. Sinha, J. Yang, R. Antonova, M. Pavone, H. Nishimura, M. Itkina, and J. Bohg, "Cupid: Curating data your robot loves with influence functions," in *Proceedings of The 9th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, J. Lim, S. Song, and H.-W. Park, Eds., vol. 305. PMLR, 27–30 Sep 2025, pp. 2907–2932. [Online]. Available: https://proceedings.mlr.press/v305/agia25a.html

[24] H. Lee, T. Min, J. Kim, S. Kang, F. Liu, L. Pinto, and K. Lee, "Quality over quantity: Demonstration curation via influence functions for data-centric robot learning," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2026, accepted. [Online]. Available: https://arxiv.org/abs/2603.09056

[25] Z. Zhou, T. Cai, Y. Zhao, Seth Z.and Zhang, Z. Huang, B. Zhou, and J. Ma, "Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning," *arXiv preprint arXiv:2506.13757*, 2025.

[26] Y. Luo, F. Li, S. Xu, Z. Lai, L. Yang, Q. Chen, Z. Luo, Z. Xie, S. Jiang, J. Liu, L. Chen, B. Wang, and Z.-X. Yang, "Adathinkdrive: Adaptive thinking via reinforcement learning for autonomous driving," *ArXiv*, vol. abs/2509.13769, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:281332471

[27] R. Sinha, A. Elhafsi, C. Agia, M. Foutter, E. Schmerling, and M. Pavone, "Real-time anomaly detection and reactive planning with large language models," in *Robotics: Science and Systems*, 2024.

[28] A. Elhafsi, R. Sinha, C. Agia, E. Schmerling, I. A. D. Nesnas, and M. Pavone, "Semantic anomaly detection with large language models," *Auton. Robots*, vol. 47, no. 8, p. 1035–1055, Oct. 2023. [Online]. Available: https://doi.org/10.1007/s10514-023-10132-6

[29] A. Z. Ren, A. Dixit, A. Bodrova, S. Singh, S. Tu, N. Brown, P. Xu, L. Takayama, F. Xia, J. Varley, Z. Xu, D. Sadigh, A. Zeng, and A. Majumdar, "Robots that ask for help: Uncertainty alignment for large language model planners," in *Proceedings of the Conference on Robot Learning (CoRL)*, 2023.

[30] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=B1ckMDqlg

[31] Z. Chen, J. Li, P. Chen, Z. Li, K. Sun, Y. Luo, Q. Mao, M. Li, L. Xiao, D. Yang, Y. Ban, H. Sun, and P. S. Yu, "Harnessing multiple large language models: A survey on llm ensemble," 2025. [Online]. Available: https://arxiv.org/abs/2502.18036

[32] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: scaling to trillion parameter models with simple and efficient sparsity," *J. Mach. Learn. Res.*, vol. 23, no. 1, Jan. 2022.

[33] W. Jitkrittum, H. Narasimhan, A. S. Rawat, J. Juneja, C. Wang, Z. Wang, A. Go, C.-Y. Lee, P. Shenoy, R. Panigrahy, A. K. Menon, and S. Kumar, "Universal model routing for efficient llm inference," 2025. [Online]. Available: https://arxiv.org/abs/2502.08773

[34] D. Ding, A. Mallick, S. Zhang, C. Wang, D. Madrigal, M. D. C. H. Garcia, M. Xia, L. V. S. Lakshmanan, Q. Wu, and V. Rühle, "Best-route: Adaptive llm routing with test-time optimal compute," 2025. [Online]. Available: https://arxiv.org/abs/2506.22716

[35] L. Chen, M. Zaharia, and J. Zou, "FrugalGPT: How to use large language models while reducing cost and improving performance," *Transactions on Machine Learning Research*, 2024, featured Certification. [Online]. Available: https://openreview.net/forum?id=cSimKw5p6R

[36] I. Ong, A. Almahairi, V. Wu, W.-L. Chiang, T. Wu, J. E. Gonzalez, M. W. Kadous, and I. Stoica, "RouteLLM: Learning to route LLMs from preference data," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=8sSqNntaMr

[37] Google AI, "Gemini api thinking documentation," 2026. [Online]. Available: https://ai.google.dev/gemini-api/docs/thinking

[38] OpenAI, "Gpt-5.1: A smarter, more conversational chatgpt," https://openai.com/index/gpt-5-1/, November 2025, accessed: 2026-04-09.

[39] M. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[40] J.-J. Hwang, R. Xu, H. Lin, W.-C. Hung, J. Ji, K. Choi, D. Huang, T. He, P. Covington, B. Sapp, Y. Zhou, J. Guo, D. Anguelov, and M. Tan, "Emma: End-to-end multimodal model for autonomous driving," *arXiv preprint arXiv:2410.23262*, 2024.

[41] P. Intelligence *et al.*, "$\pi_{0.5}$: A vision-language-action model with open-world generalization," *arXiv preprint arXiv:2504.16054*, 2025.