CSCI699 Project Proposal
VLA - Regression-Based Alternatives to Action Bin Discretization

Yumeng He, Peilin Cai, Phillip Huang, Junyi Ouyang, Siliang Zhang, Tianyi Zhou
University of Southern California

December 15, 2025

Abstract

Vanilla Vision Language Action model predicts robot actions by discretizing each action dimension into fixed tokens and training with next-token prediction. This fits language modeling, but introduces quantization error, jittery control, and vocabulary limits. We propose a continuous-action alternative that keeps the existing vision language backbone and replaces the tokenized action head with the following: (i) a direct regression head that transforms a learned action query into normalized continuous actions using a small MLP, optimized with MSE, RMSE, MAE, or Huber loss, and (ii) a logit-weighted expectation head that predicts a distribution over K anchor values per dimension and converts it to a single value via a probability-weighted average. Both avoid token overwrites and support parameter-efficient fine-tuning. We will conduct evaluation on our own benchmark, repeatable long-horizon trials in RobotTwin [Mu et al., 2025b], and comparisons on the LIBERO [Liu et al., 2023a] robot manipulation benchmark. Metrics include task success, trajectory smoothness, long-horizon stability, data efficiency, and inference latency, with ablations on losses, normalization, and history conditioning and comparisons to numeric embeddings such as xVal [Golkar et al., 2023] and FoNE [Zhou et al., 2025a]. The study is designed to test whether continuous outputs reduce quantization artifacts and match or exceed the success rate of the vanilla VLA under equal compute.

1 Introduction

Motivation. The current OpenVLA pipeline discretizes each action dimension into 256 bins and maps them to tokens for prediction by the LLM backbone [Kim et al., 2024]. While this allows integration with the tokenizer, it introduces quantization error and limits precision: subtle differences in continuous actions are collapsed into the same bin. Furthermore, the Llama tokenizer only reserves 100 special tokens, forcing OpenVLA to overwrite 256 rare tokens with action tokens, which can interfere with language modeling.

Other numerical embedding methods, such as xVal [Golkar et al., 2023] and Abacus embeddings [McLeish et al., 2024], also rely on discretization or digit-wise decomposition. These methods are effective in arithmetic reasoning benchmarks but are suboptimal for robotic control, where smooth and high-resolution action prediction is critical.

Proposed Alternatives. We replace bin-based discretization with regression-style formulations that produce continuous values:

1. **Direct Regression.** Each action dimension is predicted as a scalar, optimized with root mean squared error (RMSE): $\mathcal{L}_{\text{RMSE}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(a_i - \hat{a}_i)^2}$, where a_i is the ground-truth action and $\hat{a}_i$ is the model prediction.

2. **Logit-Weighted Expectation.** Define a discrete set of anchor values $\{v_1, \ldots, v_K\}$ and interpret output logits as probabilities. The predicted action is the expectation: $\hat{a} = \sum_{i=1}^{K} p_i \cdot v_i, \quad p_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$, where z_i are the logits. The training objective combines MSE on $\hat{a}$ with entropy regularization.

Expected Benefits.

- **Higher Precision:** Removes coarse quantization for smoother trajectories.
- **Reduced Token Overhead:** Eliminates the need to overwrite rare tokens.
- **Improved Generalization:** Continuous regression captures diverse action distributions.
- **Compatibility:** The logit-weighted method maintains token alignment with the LLM backbone.

Implementation Plan. We will replace the action-tokenization module with either a regression head or a logit-weighted expectation layer, and compare three variants: (1) baseline binning, (2) direct regression, and (3) logit-weighted expectation (4) Abacus (5)xVAL , etc. Evaluation metrics include task success rate, smoothness of control (variance of consecutive deltas), and training stability.

2 Related Work

Vision-Language Models (VLMs). Modern VLMs learn transferable multimodal representations from web-scale image–text pairs and can be adapted to a broad set of downstream tasks. *CLIP* established scalable contrastive pretraining for zero-shot transfer [Radford et al., 2021]; *Flamingo* introduced an interleaved vision–text architecture that supports in-context few-shot learning [Alayrac et al., 2022]; *BLIP-2* bridged frozen vision encoders and LLMs using a lightweight querying transformer [Li et al., 2023]; and *LLaVA* demonstrated visual instruction tuning for multimodal assistants [Liu et al., 2023c]. These VLMs are frequently used as backbones or initializations for robotic-oriented VLAs.

Vision-Language Action (VLAs). Recent VLAs unify visual perception and language grounding to directly output robot actions, enabling end-to-end manipulation from instruction to control. OpenVLA provides an open 7B baseline trained on $\sim 970k$ episodes from the Open X-Embodiment (OXE) corpus, facilitating cross-embodiment transfer and parameter-efficient adaptation [Collaboration, 2023]. SmolVLA explores the complementary direction of compact, resource-friendly VLAs with competitive performance [Shukor et al., 2025].

Extensions to VLAs. Beyond vanilla tokenized actions, *Optimized Fine-Tuning (OFT)* combines parallel decoding, action chunking, and continuous action regression with a simple L1 objective, yielding large speedups and higher success on LIBERO; real-robot, bimanual evaluations further show gains over from-scratch policies [Kim et al., 2025]. These results motivate hybrid or continuous action heads that reduce quantization error and enable higher control rates as a principled alternative to purely discrete tokenization.

Benchmarks. For language-conditioned manipulation, LIBERO offers 130 procedurally generated tasks across four suites and has become a de-facto standard for VLA fine-tuning and generalization [Liu et al., 2023b]. Complementing this, the dual-arm RoboTwin & RoboTwin 2.0 provides a scalable, domain-randomized bimanual benchmark (50 tasks, 731 objects) with a generative digital-twin pipeline [Mu et al., 2025a, Chen et al., 2025]. Beyond these, VIMA-Bench [Jiang et al., 2022], RL-Bench [James et al., 2019], CALVIN [Mees et al., 2021], and ManiSkill3 [Tao et al., 2024] support simulator-style online evaluation, while OXE [Collaboration, 2023] and BridgeData V2 [Walke et al., 2023] offer large-scale offline real-robot distributions.

3 Proposed Method and Pipeline

Overview. We propose a regression-oriented reformulation of the action decoding stage in OpenVLA. Instead of discretizing each action dimension into 256 bins, we explore continuous prediction heads that improve precision and smoothness of control. Specifically, we compare three alternatives—*xVal embeddings*, *logit-weighted sum*, and *direct regression*—with an optional enhancement using Fourier Number Embeddings (FoNE). Our pipeline minimally modifies the architecture: only the decoding head is replaced, followed by lightweight finetuning of the model.

Action Decoding Alternatives. We design and evaluate the following decoding strategies:

1. **xVal.** Following [Golkar et al., 2023], each number n is represented as the elementwise product of its embedding and its numerical value. For action prediction, the LLM outputs a representation that directly scales with the action value, bridging embedding-based representations and continuous outputs.

2. **Logit-Weighted Sum.** Define a fixed set of anchor values $\{v_1, \ldots, v_K\}$ for each action dimension. The model produces logits $\{z_i\}$, normalized into probabilities $p_i = \mathrm{softmax}(z_i)$. The predicted action is the expectation:
$$\hat{a} = \sum_{i=1}^{K} p_i \cdot v_i.$$
This approach retains a token-like discrete structure while outputting continuous values, and can be regularized with entropy terms for sharper predictions.

3. **Direct Regression.** A lightweight MLP regression head maps the hidden states of the LLM to continuous scalars. Training minimizes root mean squared error (RMSE):
$$\mathcal{L}_{\mathrm{RMSE}} = \sqrt{\tfrac{1}{N} \sum_{i=1}^{N} (a_i - \hat{a}_i)^2}.$$

4. **FoNE-Enhanced Regression (Optional).** To further improve precision, we replace raw scalars with Fourier Number Embeddings [Zhou et al., 2025b]. The regression head outputs a Fourier-encoded representation, which is decoded back into scalars, allowing smoother extrapolation and better generalization.

Pipeline. Our proposed pipeline consists of four stages:

1. **Input Encoding:** Task instructions and observations are encoded into hidden representations via a pretrained multimodal backbone.

2. **Backbone Processing:** A Llama-based decoder integrates contextual and multimodal information.

3. **Action Decoding Head:** Instead of a discretization-based action tokenizer, one of the proposed heads (xVal, logit-weighted sum, regression, or FoNE-enhanced regression) predicts continuous actions.

4. **Finetuning:** We finetune the new decoding head and optionally the final transformer block on paired $(\text{instruction}, \text{observation}, \text{action})$ data, keeping most of the backbone frozen.

Evaluation Plan. We will compare the following decoding strategies: (1) baseline 256-bin discretization, (2) xVal, (3) logit-weighted sum, (4) direct regression, and (5) FoNE-enhanced regression. Metrics include task success rate, smoothness of action trajectories (variance of consecutive deltas), training stability, and vocabulary/token efficiency.

4 Experiment

4.1 Benchmark design

To evaluate continuous-action VLA models, we will develop a benchmark focused on multi-step manipulation tasks. Experiments will be conducted in the RobotTwin simulator [Mu et al., 2025b], which enables repeatable execution of robot trajectories with precise logging of control smoothness, failure points, and recovery attempts. We will additionally compare against the LIBERO benchmark suite [Liu et al., 2023a], a widely used evaluation framework for manipulation tasks with natural language instructions. Benchmark metrics will include task success rate, trajectory smoothness, long-horizon stability, data efficiency, inference latency, and safety violations (e.g., collisions or excessive forces). By incorporating both simulation and real-world grounded datasets, the benchmark will measure not only accuracy but also deployability under realistic robotic constraints.

We will test on different loss functions (ways the model measures and learns from its errors, such as MSE, RMSE, MAE, or Huber), normalization strategies (how we scale or adjust the action values), and history conditioning (how much of the past sequence of actions the model considers), along with comparisons to recent numeric embedding methods such as xVal [Golkar et al., 2023] and FoNE [Zhou et al., 2025a]. These experiments will help us understand whether continuous-action models can actually perform better than token-based models when both use the same amount of compute and data.

4.2 Application to robotics

Beyond controlled benchmarks in simulated environments, we will also test the continuous-action VLA models for robotic manipulation in real-world environments. By eliminating discretization artifacts, we hope these models can generate smoother, safer, and more efficient trajectories, which can benefit applications such as manufacturing, household assistance, and medical robotics, where precision and safety are critical. The proposed heads should also support parameter fine-tuning, making them practical for adapting pretrained VLA models to specific datasets with limited additional supervision.

For the evaluation in simulations, we will use RobotTwin [Mu et al., 2025b] for long-horizon reproducibility and LIBERO [Liu et al., 2023a] as a standardized manipulation dataset, while for real-world environment, we will use the same datasets for the simulated environment with paired robot demonstrations. The real-world evaluation will also be assessed using the same measures of task success rate, trajectory efficiency, latency, and safety. Our approach aims to close the gap between language reasoning and the fine-grained motor control required for real-world robotic autonomy.

5 Work Assignment

1. Benchmark Design: Yumeng He, Phillip Huang, Junyi Ouyang
2. Training Strategy: Peilin Cai, Tianyi Zhou, Siliang Zhang
3. Application to Robotics: Yumeng He, Phillip Huang, Junyi Ouyang, Peilin Cai, Tianyi Zhou, Siliang Zhang

References

[Alayrac et al., 2022] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. (2022). Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[Chen et al., 2025] Chen, T., Chen, Z., Chen, B., et al. (2025). Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. https://robotwin-platform.github.io/.

[Collaboration, 2023] Collaboration, O. X.-E. (2023). Open x-embodiment: Robotic learning datasets and rt-x models.

[Golkar et al., 2023] Golkar, S., Pettee, M., Eickenberg, M., Bietti, A., Cranmer, M., Krawezik, G., Lanusse, F., McCabe, M., Ohana, R., Parker, L., et al. (2023). xval: A continuous number encoding for large language models. *arXiv preprint arXiv:2310.02989*.

[James et al., 2019] James, S., Ma, Z., et al. (2019). Rlbench: The robot learning benchmark & learning environment.

[Jiang et al., 2022] Jiang, Y., Zhu, Y., Fan, L., et al. (2022). Vima: General robot manipulation with multimodal prompts.

[Kim et al., 2025] Kim, M. J., Finn, C., and Liang, P. (2025). Fine-tuning vision-language-action models. https://openvla-oft.github.io/.

[Kim et al., 2024] Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al. (2024). Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*.

[Li et al., 2023] Li, J., Li, D., Savarese, S., and Hoi, S. C. H. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

[Liu et al., 2023a] Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., and Stone, P. (2023a). Libero: Benchmarking knowledge transfer for lifelong robot learning.

[Liu et al., 2023b] Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., and Stone, P. (2023b). Libero: Benchmarking knowledge transfer for lifelong robot learning. In *NeurIPS 2023 Datasets and Benchmarks Track*.

[Liu et al., 2023c] Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023c). Visual instruction tuning.

[McLeish et al., 2024] McLeish, S., Bansal, A., Stein, A., Jain, N., Kirchenbauer, J., Bartoldson, B., Kailkhura, B., Bhatele, A., Geiping, J., Schwarzschild, A., et al. (2024). Transformers can do arithmetic with the right embeddings. *Advances in Neural Information Processing Systems*, 37:108012–108041.

[Mees et al., 2021] Mees, O., Hermann, L., Rosete-Beas, E., and Burgard, W. (2021). Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks.

[Mu et al., 2025a] Mu, Y., Chen, T., Chen, Z., et al. (2025a). Robotwin: Dual-arm robot benchmark with generative digital twins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Mu et al., 2025b] Mu, Y., Chen, T., Chen, Z., Peng, S., Lan, Z., Gao, Z., Liang, Z., Yu, Q., Zou, Y., Xu, M., Lin, L., Xie, Z., Ding, M., and Luo, P. (2025b). Robotwin: Dual-arm robot benchmark with generative digital twins. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 27649–27660.

[Radford et al., 2021] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

[Shukor et al., 2025] Shukor, M., Aubakirova, D., Capuano, F., Kooijmans, P., Palma, S., Zouitine, A., Aractingi, M., Pascal, C., Russi, M., Marafioti, A., Alibert, S., Cord, M., Wolf, T., and Cadene, R. (2025). Smolvla: A vision-language-action model for affordable and efficient robotics. https://huggingface.co/blog/smolvla.

[Tao et al., 2024] Tao, S., Xiang, F., Shukla, A., Qin, Y., Hinrichsen, X., Yuan, X., Bao, C., Lin, X., Liu, Y., kai Chan, T., Gao, Y., Li, X., Mu, T., Xiao, N., Gurha, A., Huang, Z., Calandra, R., Chen, R., Luo, S., and Su, H. (2024). Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai.

[Walke et al., 2023] Walke, H., Black, K., Lee, A., Kim, M. J., Du, M., Zheng, C., Zhao, T., Hansen-Estruch, P., Vuong, Q., He, A., Myers, V., Fang, K., Finn, C., and Levine, S. (2023). Bridgedata v2: A dataset for robot learning at scale.

[Zhou et al., 2025a] Zhou, T., Fu, D., Soltanolkotabi, M., Jia, R., and Sharan, V. (2025a). Fone: Precise single-token number embeddings via fourier features.

[Zhou et al., 2025b] Zhou, T., Fu, D., Soltanolkotabi, M., Jia, R., and Sharan, V. (2025b). Fone: Precise single-token number embeddings via fourier features. *arXiv preprint arXiv:2502.09741*.