**CubicVLA: Efficient Robot Action Representation via Cubic Trajectory Parameterization**

Keywords: Vision-Language-Action models, tokenization techniques, dimension reduction

Large vision-language models (VLMs), particularly transformer-based architectures, have demonstrated strong generalization capabilities through pre-training on large-scale image-text datasets. Recently, VLMs have been adapted for robotics by fine-tuning them on action prediction tasks, giving rise to vision-language-action (VLA) models that enable end-to-end continuous control. However, these transformer-based models inherently operate in a discrete token space—they consume discrete inputs and produce discrete outputs. This introduces a fundamental challenge of how to tokenize the continuous action space such that it can be effectively represented and predicted in a discrete format.

The most straightforward method would be to apply naive binning; however, such approach performs poorly when applied in high control frequency scenarios. Alternative approaches leverage the Discrete Cosine Transform (DCT), which decomposes a continuous trajectory into a sum of cosine basis functions, yielding a more compact representation than naive binning. These limitations highlight the need for an action representation that is both compact and expressive. To this end, we propose CubicVLA, a novel representation that models continuous actions through cubic interpolation of sparse control points.

CubicVLA parameterizes continuous robot actions as Piecewise Cubic Hermite curves reconstructed from a sparse set of control points. To further reduce sequence length, we apply byte-pair encoding (BPE) to the discretized control points, compressing repeated patterns across trajectories. The method is simple, non-learned, and fully compatible with existing token-based architectures. Compared to baseline discretization methods, CubicVLA achieves lower reconstruction error, more stable policy rollouts, and superior performance in high-frequency control scenarios. Overall, this representation effectively captures the structure of high-frequency continuous trajectories while minimizing redundancy.

To evaluate the performance of our method against existing baselines, we evaluated on action token length and reconstruction error when applied to manipulation tasks. Specifically, we evaluated using the LIBERO-100 (simulation-based manipulation), iDP3 (real-world humanoid manipulation), BiPlay (bimanual manipulation), and Bridge (diverse tabletop manipulation) datasets.

Our results demonstrate that CubicVLA consistently achieves the highest action token length compression ratios across all datasets, with particularly strong gains on high-frequency tasks such as BiPlay (34.15×) and LIBERO-90 (15.01×). When evaluated on reconstruction error, the cubic representation also yields lower error on LIBERO-90 and iDP3, datasets with particularly higher control frequencies. To assess downstream effectiveness, we integrated our tokenization method into a VLA-based imitation learning policy and deployed it in simulation on the LIBERO-90 manipulation benchmark. The cubic method achieved a success rate of 89.36%, surpassing DCT by 3.32% and the vanilla (uncompressed) baseline by 1.56%. Finally, we successfully validated our approach on a Franka Emika Panda robot performing real-world pick-and-place tasks, demonstrating the practicality and effectiveness of our solution beyond simulation.