AB-DualVLA: Task-Routed Cloud–Edge Vision–Language–Action Policies

Yuhe Wen, Jaeho Jung, Quan Gan, Jehwan Choi, Duy-Linh Nguyen, Kang-Hyun Jo
Dept. of Electrical, Electronic and Computer Engineering, University of Ulsan, Ulsan, Korea
{wenyuhe, jjho0514, ganquan, ndlinh301}@mail.ulsan.ac.kr; {cjh1897, acejo}@ulsan.ac.kr

Abstract

We study how cloud-side and edge-side VLAs can be organized into a task-routed manipulation system. ***AB-DualVLA*** *(Attention-Bottleneck DualVLA) keeps Pi0 and SmolVLA frozen, trains a token-level fusion student path over extracted intermediate tokens, and selects among paths at the task level. On LIBERO-Object (10 tasks) the system reaches 89.5% closed-loop success at 10 episodes per task per seed, averaged over multiple seeds. The 79M student attains low open-loop fit, but this does not transfer to standalone closed-loop performance; the student is therefore used as one of the routed paths rather than a replacement for either backbone. We report the experimental results in open-loop fit, latency, and fallback behavior metrics.*

1. Introduction

Vision-language-action (VLA) policies such as Pi0, RT-2, OpenVLA, Octo, and RoboMamba have improved language-conditioned robot manipulation [2, 3, 9, 14, 17], leveraging cross-embodiment pre-training corpora such as Open X-Embodiment [18]. Compact VLA variants including SmolVLA, TinyVLA, and CEED-VLA make edge-side deployment more practical [22, 23, 25]. A robot may have access to both a strong cloud policy and a compact local policy, with different latency, availability, and task strengths. Prior cloud-edge and split-inference work studies distributed computation across devices and hybrid SLM/LLM serving [6, 7, 19]. Related embodied-AI work has explored edge perception [4], pre-trained visual representations [24], hierarchical diffusion policies [15], and dynamics-aligned flow matching [10]; in contrast, we ask whether heterogeneous frozen VLA branches can be organized into a task-routed system with executable fallback paths.

AB-DualVLA builds on frozen Pi0 and SmolVLA branches without fine-tuning or modifying their forward passes. It trains a 79M token-level fusion student over intermediate tokens from both branches using MBT-style bottleneck communication [1, 8, 11, 16, 21]. At the system level, it selects among the cloud branch, edge branch, and student path according to task-level validation performance.

2. Method

Student path. The AB-DualVLA student forms a single input token sequence from projected cloud tokens $\mathbf{z}^c$, edge tokens $\mathbf{z}^e$, a robot-state token $\mathbf{s}$, and a task token $\boldsymbol{\tau}$, where $\boldsymbol{\tau}$ is a learned task-embedding lookup indexed by the 10 LIBERO-Object task identifiers. Pi0 and SmolVLA action chunks are not forward inputs; they are used only as labels or diagnostics, so the student learns to combine the two backbones' *representations* rather than ensemble their actions. From each tap layer we keep the first 8 image-conditioned tokens after the language prefix, fixed across observations, and project both streams to 512 dimensions; the tap layers (Pi0 layers 6 and 13; SmolVLA layers 8 and 18) are fixed design choices used throughout all experiments, following multi-tap practice in multimodal fusion [11, 16].

Bottleneck communication. The student has four transformer layers. The first two layers process cloud and edge streams separately; the last two use an MBT-style bottleneck with $K{=}8$ learned shared bottleneck tokens $\mathbf{b}$. Cloud and edge transformer blocks T_c and T_e share the same architecture but have separate parameters per stream. For each bottleneck layer,

$$[\tilde{z}^c, \hat{b}^c] = T_c([z^c, b]),\ [\tilde{z}^e, \hat{b}^e] = T_e([z^e, b]),\ b' = \tfrac{1}{2}(\hat{b}^c + \hat{b}^e) \tag{1}$$

where the updated $\mathbf{b}'$ feeds the next bottleneck layer. Cloud and edge tokens are not placed in the same self-attention sequence inside the bottleneck layers; cross-stream information must pass through $\mathbf{b}$. The final decoder concatenates $[\tilde{\mathbf{z}}^c, \mathbf{b}', \tilde{\mathbf{z}}^e, \mathbf{s}, \boldsymbol{\tau}]$ and uses 50 learned action queries to predict a 50×7 action chunk, following action-chunk imitation practice [5, 26]. Training uses gripper-weighted action MSE and a direction-consistency term relative to the cached cloud action prior; we augment the cached demonstration set with on-policy student rollouts re-labeled by the Pi0 prior, in the spirit of DAgger [20].

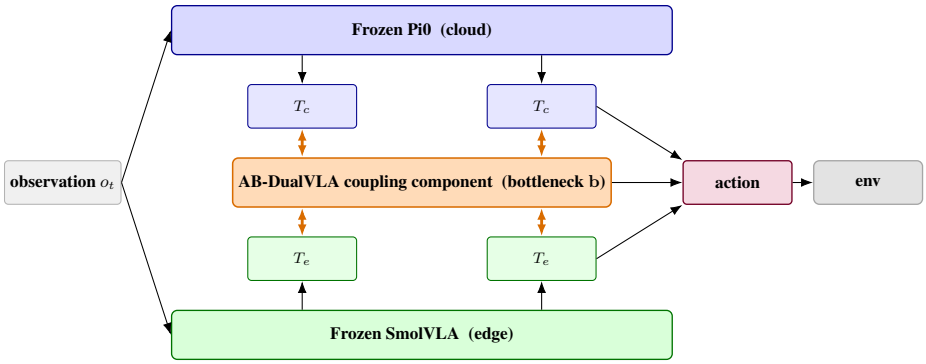

Figure 1. **AB-DualVLA overview.** Frozen Pi0 and SmolVLA branches run normally. Mid-layer tokens are tapped into the AB-DualVLA student, where cloud and edge streams communicate only through bottleneck **b**; the system then routes among available action paths.

System routing and fallback. The full AB-DualVLA system uses task-conditioned routing among Pi0, SmolVLA, and the student path. The router is a validation-based task lookup, not a learned confidence estimator. If cloud is unavailable, the system falls back to the edge branch; if edge is unavailable, it falls back to the cloud branch.

3. Experiments

We evaluate on LIBERO-Object [13] (10 tasks) using closed-loop success rate (SR); recent LIBERO-PRO [27] introduces additional robustness perturbations beyond memorization, which we leave to future work. Each branch uses its released inference setting, and all three paths (Pi0, SmolVLA, the student) are evaluated at 10 episodes per task on the same task list, with results averaged over multiple random seeds. Per-task routing decisions for both the two-branch baseline and the AB-DualVLA system are selected on a validation seed set disjoint from the reported test rollouts.

Table 1. **Closed-loop SR on LIBERO-Object** (10 ep/task, multi-seed). System score is task-level routed, not a mean of component rows.

Method / mode	Role	Ep./task	SR
Pi0	cloud branch	10	79.6
SmolVLA	edge branch	10	85.0
AB-DualVLA student (alone)	student path	10	60.0
Two-branch routing (Pi0+SmolVLA)	2-branch baseline	10	87.6
AB-DualVLA system (3-branch)	**task-routed**	10	**89.5**

Under the same task-level validation routing protocol, a Pi0+SmolVLA two-branch baseline reaches 87.6% SR. The integrated three-branch AB-DualVLA system reaches **89.5%**; the additional headroom is contributed by the student path on tasks where neither backbone alone is best, despite its lower standalone SR (60.0%).

Worked examples. The system shares one input/output pipeline across tasks: RGB observation $o_t \in \mathbb{R}^{256\times256\times3}$, robot state $s_t \in \mathbb{R}^8$, and task index τ. Mid-layer taps from Pi0 (L6/L13, 16×2048-d) and SmolVLA (L8/L18, 16×960-d) project to 512-d and are fused inside the student via the $K{=}8$ bottleneck; the decoder emits a 50×7 action chunk. We illustrate one strong and one challenging task.

(i) t1 "pick up the cream cheese and place it in the basket" – Pi0 alone: **74%**; SmolVLA alone: **10/10 (100%)**; AB-DualVLA system: **10/10 (100%)**. Despite Pi0's $\sim 7\times$ larger parameter count, it underperforms SmolVLA, indicating that scale alone does not determine closed-loop success on this benchmark.

(ii) t0 "pick up the alphabet soup and place it in the basket" – Pi0 alone: **76%**; SmolVLA alone: **60%**; AB-DualVLA student alone: **62%**; AB-DualVLA system: **76%**. All three backbones struggle; routing cannot exceed the strongest available branch when no single backbone is strong on a task.

Student diagnostics. We report open-loop fitting separately as a diagnostic for the student path.

Table 2. **Open-loop action fit.** Validation MSE vs. demonstrations.

Predictor	Params	Val MSE
Pi0 action prior	3.5B	0.0453
SmolVLA action prior	0.5B	0.2289
AB-DualVLA student	**79M**	**0.0099**

The 79M student obtains lower validation MSE than either backbone action prior, but this does not transfer to closed-loop SR: open-loop MSE on cached states does not penalize compounding errors under autoregressive execution [20], motivating the routed-path role of the student rather than standalone use.

Latency and degradation. When the student path is selected, AB-DualVLA avoids the full Pi0 flow decoder [12] and runs in 131 ms per action chunk, vs. 891 ms for Pi0 and 428 ms for SmolVLA on the same single-GPU benchmark. The 131 ms covers the student forward and backbone partial forwards but excludes cloud–edge network round-trip.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[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. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

[3] Anthony Brohan et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning (CoRL)*, 2023.

[4] Aditya Chakravarty. Multi-step guided diffusion for image restoration on edge devices: Toward lightweight perception in embodied AI. In *CVPR Embodied AI Workshop*, 2025.

[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 (RSS)*, 2023.

[6] Zixu Hao, Huiqiang Jiang, Shiqi Jiang, Ju Ren, and Ting Cao. Hybrid SLM and LLM for edge-cloud collaborative inference. In *Proceedings of the Workshop on Edge and Mobile Foundation Models (EdgeFM)*, 2024.

[7] Shijing Hu, Zhihui Lu, Xin Xu, Ruijun Deng, Xin Du, and Qiang Duan. LAECIPS: Large vision model assisted adaptive edge-cloud collaboration for IoT-based embodied intelligence system. *arXiv preprint arXiv:2404.10498*, 2024.

[8] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J. Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver IO: A general architecture for structured inputs and outputs. In *International Conference on Learning Representations (ICLR)*, 2022.

[9] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag 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 *Conference on Robot Learning (CoRL)*, 2024.

[10] Dohyeok Lee, Jung Min Lee, Munkyung Kim, Seokhun Ju, Seungyub Han, Jin Woo Koo, and Jungwoo Lee. Dynamics-aligned flow matching policy for robot learning. In *CVPR Embodied AI Workshop*, 2025.

[11] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, 2023.

[12] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

[13] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2023.

[14] Jiaming Liu, Mengzhen Liu, Zhenyu Wang, Pengju An, Xiaoqi Li, Kaichen Zhou, Senqiao Yang, Renrui Zhang, Yandong Guo, and Shanghang Zhang. RoboMamba: Efficient vision-language-action model for robotic reasoning and manipulation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[15] Yiyang Lu, Yufeng Tian, Zhecheng Yuan, Xianbang Wang, Pu Hua, Zhengrong Xue, and Huazhe Xu. H3DP: Triply-hierarchical diffusion policy for visuomotor learning. In *CVPR Embodied AI Workshop*, 2025.

[16] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[17] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. In *Robotics: Science and Systems (RSS)*, 2024.

[18] Open X-Embodiment Collaboration. Open X-Embodiment: Robotic learning datasets and RT-X models. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[19] Pratyush Patel, Esha Choukse, Chaojie Zhang, Aashaka Shah, Íñigo Goiri, Saeed Maleki, and Ricardo Bianchini. Splitwise: Efficient generative LLM inference using phase splitting. In *Proceedings of the 51st Annual International Symposium on Computer Architecture (ISCA)*, 2024.

[20] Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

[21] Michael Ryoo, A. J. Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. TokenLearner: Adaptive space-time tokenization for videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[22] Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, Simon Alibert, Matthieu Cord, Thomas Wolf, and Remi Cadene. SmolVLA: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*, 2025.

[23] Wenxuan Song, Jiayi Chen, Pengxiang Ding, Yuxin Huang, Han Zhao, Donglin Wang, and Haoang Li. CEED-VLA: Consistency vision-language-action model with early-exit decoding. *arXiv preprint arXiv:2506.13725*, 2025.

[24] Nikolaos Tsagkas, Andreas Sochopoulos, Duolikun Danier, Sethu Vijayakumar, Chris Xiaoxuan Lu, and Oisin Mac Aodha. On the use of pre-trained visual representations in visuo-motor robot learning. In *CVPR Embodied AI Workshop*, 2025.

[25] Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Kun Wu, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, Yaxin Peng, Feifei Feng, and Jian Tang. TinyVLA: Towards fast, data-efficient vision-language-action models for robotic manipulation. *arXiv preprint arXiv:2409.12514*, 2024.

[26] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems (RSS)*, 2023.

[27] Xueyang Zhou, Yangming Xu, Guiyao Tie, Yongchao Chen, Guowen Zhang, Duanfeng Chu, Pan Zhou, and Lichao Sun. LIBERO-PRO: Towards robust and fair evaluation of vision-language-action models beyond memorization. *arXiv preprint arXiv:2510.03827*, 2025.