

CLaD: Planning with Grounded Foresight via Cross-Modal Latent Dynamics

Supplementary Material

1. Evaluating Generalizability & Scalability

LIBERO benchmark [5] provides three other LIBERO suites except for LIBERO-LONG, each of which contains 10 tasks with 50 demonstrations to evaluate the model’s generalizability and scalability over certain properties:

- **LIBERO-Spatial** evaluates the agent’s ability to reason about spatial configurations about diverse layouts of objects;
- **LIBERO-Object** introduces variations in object types while maintaining the scene layouts, assessing the agent’s capacity to generalize across object instances;
- **LIBERO-Goal** evaluates the generalizability over diverse task objectives given consistent objects and layouts.

Table 1. Average success rates (%) over 50 rollouts on all suites in LIBERO benchmark. Color intensity is proportional to the performance level (thicker = higher).

Method	Spatial	Object	Goal	Long
π_0^\diamond	97.2	97.8	91.6	82.0
UniVLA	96.5	96.8	95.6	92.0
$\pi_{0.5}^\diamond$	97.6	98.4	97.2	93.2
OpenVLA $^\diamond$	98.2	98.6	97.6	93.8
CLaD (Ours)	97.3	95.7	94.3	94.7

\diamond : These results are reported in LIBERO-PRO [8].

CLaD achieves strong performance on long-horizon tasks (i.e., LIBERO-LONG) but sub-optimal results on short-horizon tasks focusing on generalization (i.e., LIBERO-Spatial, Object, Goal), compared to SOTA large VLAs [1–4] pre-trained on massive demonstrations, such as Open X-Embodiment [6]. Our results indicate that large VLA models leverage substantial pre-trained knowledge for strong in-distribution generalization. CLaD, while having less background knowledge, demonstrates particular strengths in knowledge-independent planning based on cross-modal dynamics modeling. These findings point to CLaD’s potential as a building block for scalable robotic planning systems.

2. Visualization of learned foresight.

Figure 1 shows which regions influence CLaD’s predictions using integrated gradients [7]. The gradient concentrates on task-relevant objects: heatmaps peak near targets in grasping; shift between held object and target in placement. We’ve confirmed that the predicted foresight attends broadly to object boundaries. While the resulting heatmaps (see Figure 1) do not yield precise object boundaries, they consistently highlight task-relevant objects within the current image. This

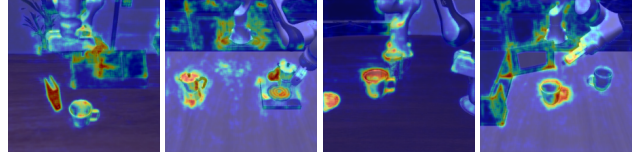


Figure 1. Pixel attribution for predicted latent foresight via Integrated Gradients. Heatmaps show pixel-level contributions toward the alignment between predicted foresight $\hat{z}^{t+\tau}$ and target embedding. Brighter regions indicate higher attribution scores. While not yielding precise object boundaries, attributions consistently highlight task-relevant objects, suggesting that the model leverages semantic features for future state prediction.

qualitative analysis confirms that the model relies on meaningful semantic cues to predict future states, demonstrating feasible grounding on observations.

The visualizations show concentrated attention around task-relevant objects: in grasping tasks, heatmaps peak near target objects; in placement tasks, attention shifts between the held object and target location. This suggests CLaD learns spatially grounded predictions rather than relying on global scene context.

3. Effect of Action-Free Data

Unlike UVA [2], CLaD’s core objective is learning shared dynamics across modalities via asymmetric cross-attention, which fundamentally requires action conditioning. Notably, CLaD already employs stochastic action masking with masking ratio $r = 0.3$ to encourage transition inference. To empirically assess action-free data utility, we evaluated Action-free, Heavy Action Mask ($r = 0.9$), and Curriculum Learning (initially action-free, then mask with $r = 0.3$) variants, but all underperformed the baseline as shown in Table 2. We hypothesize that removing action guidance introduces multi-modal ambiguity and optimization interference, confirming that our existing strategy is optimal for maintaining deterministic, grounded foresight.

Variants	Heavy Mask	Action-free	Curriculum	CLaD
Avg. SR (%)	88.2	90.8	85.1	94.7

Table 2. Ablation with action-free training variants.

4. Qualitative Results

Demonstrations on LIBERO-LONG Figure 2, 3, 4, and 5 visualize representative rollouts for all 10 tasks in LIBERO-LONG. Green circles mark successful task execution, while red crosses show failures. We provide a supplementary video demonstrating LIBERO-LONG tasks.

References

- [1] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0 : A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. [1](#)
- [2] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025. [1](#)
- [3] Physical Intelligence et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv*, 2025.
- [4] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. [1](#)
- [5] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023. [1](#)
- [6] Abby O’Neill, Abdul Rehman, Abhiram Maddukuri, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *International Conference on Robotics and Automation*, pages 6892–6903. IEEE, 2024. [1](#)
- [7] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. [1](#)
- [8] Xueyang Zhou et al. Libero-pro: Towards robust and fair evaluation of vision-language-action models beyond memorization. *arXiv*, 2025. [1](#)



Figure 3. Demonstrations on task 4, 5, 6 of LIBERO-LONG.



Figure 4. Demonstrations on task 7, 8, 9 of LIBERO-LONG.

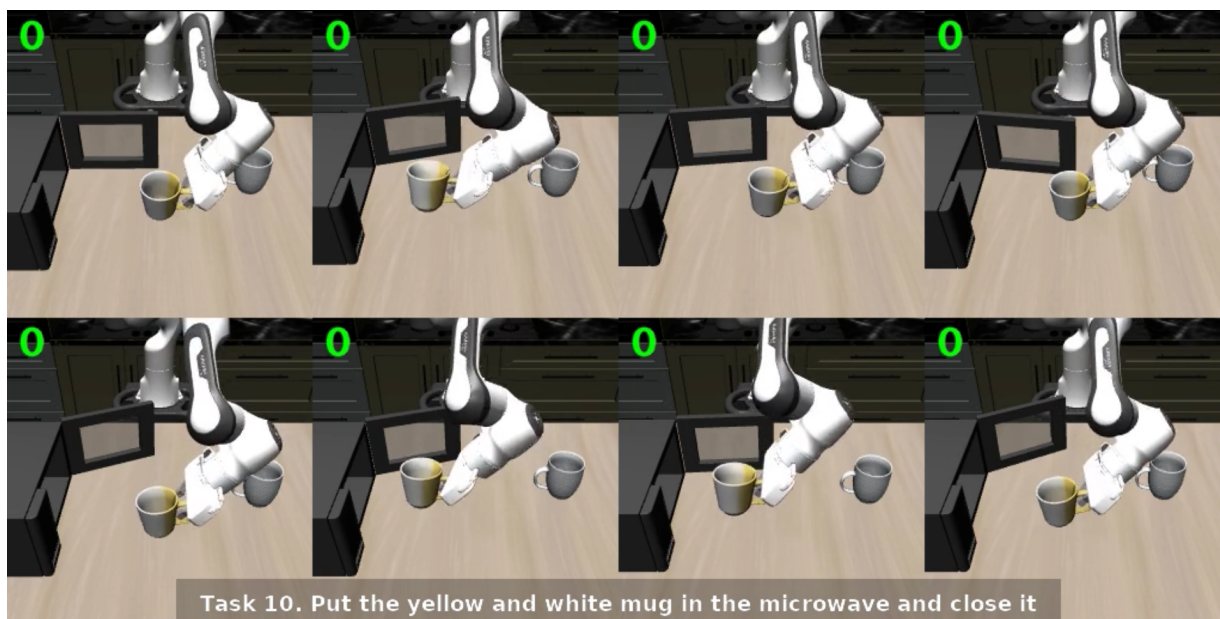


Figure 5. Demonstrations on task 10 of LIBERO-LONG.