

7 Appendix

7.1 Details of Environments

LIBERO-90. LIBERO is a tabletop manipulation benchmark built on the Franka robotic arm, designed to cover a broad range of complex and dexterous manipulation tasks. Each task is defined by a natural language instruction that specifies both the target object and its desired final state. The action space is 7-dimensional, consisting of the end-effector’s positional and orientational deltas, along with the gripper control force.

We conduct a comprehensive evaluation of our proposed method on the LIBERO-90 [40] subset, which includes a diverse set of language-conditioned tasks requiring precise understanding of fine-grained semantics and spatial relationships. Detailed descriptions of the task environments are provided in the supplementary materials. Many tasks involve perceptually similar objects with distinct semantic goals—for instance, multiple identical-looking containers that need to be placed at different target locations. In such scenarios, where semantic ambiguity and spatial uncertainty are tightly coupled, policies trained solely on single-view RGB input often struggle to make accurate distinctions. To ensure a fair comparison, we follow the evaluation protocol from prior work [9] and select 8 representative tasks from two representative scenarios in LIBERO-90:

- **LIVING ROOM SCENE5:** Put the red mug on the left plate.
- **LIVING ROOM SCENE5:** Put the red mug on the right plate.
- **LIVING ROOM SCENE5:** Put the white mug on the left plate.
- **LIVING ROOM SCENE5:** Put the yellow and white mug on the right plate.
- **LIVING ROOM SCENE6:** Put the chocolate pudding to the left of the plate.
- **LIVING ROOM SCENE6:** Put the chocolate pudding to the right of the plate.
- **LIVING ROOM SCENE6:** Put the red mug on the plate.
- **LIVING ROOM SCENE6:** Put the white mug on the plate.

Each task is trained using 20 expert demonstration trajectories provided by the benchmark. These tasks are characterized by language instructions that exhibit high semantic ambiguity and spatial diversity, requiring the policy to generate precise and distinct behaviors under visually similar conditions. Although these scenes may appear visually alike, they demand clearly differentiated actions, making them particularly challenging for RGB-only models to disambiguate.

Meta-World. Meta-World [41] is a simulated robotic manipulation benchmark built upon the Sawyer robotic arm. It includes a variety of tabletop tasks involving object interaction and tool usage. The action space is 4-dimensional, comprising the end-effector’s positional deltas and gripper control.

Compared to LIBERO, tasks in Meta-World feature clearly defined goals, concise language instructions, and a limited set of target objects. However, they demand fine-grained low-level control skills, such as insertion, pressing, and rotation. In this environment, the student model receives only single-view RGB images and simple textual commands, making it a suitable testbed for evaluating the effectiveness of our distillation framework in recovering spatial representations and modeling trajectories. Strong performance under this minimal-input setting further demonstrates our method’s capability in precision-critical tasks. While most tasks are short-horizon and semantically unambiguous, they still require precise spatial reasoning to satisfy geometric constraints, such as hole alignment or tool orientation. In such scenarios, even minor spatial errors or unstable trajectories from RGB-only inputs can lead to failure, highlighting the essential role of distillation in enhancing both precision and robustness.

To systematically evaluate the performance of our framework across different levels of task difficulty, we categorize 15 tasks into three groups. Each task in the training set includes 20 expert demonstration trajectories provided by the benchmark.

- **Easy (7 tasks):** `button_press`, `button_press_topdown`, `drawer_close`, `faucet_close`, `window_close`, `window_open`, `faucet_open`. These tasks involve simple control logic and clear target visibility.

- **Medium (5 tasks):** bin_picking, drawer_open, door_open, dial_turn, hammer. These require intermediate-level control of pose or stable contact interactions.
- **Hard (3 tasks):** disassemble, shelf_place, basketball. These involve structural alignment or tool use, with high precision demands and elevated failure rates.

LIBERO-LONG. To evaluate the effectiveness of our proposed **MonoLift** framework in handling long-horizon, multi-stage manipulation tasks, we conduct experiments on the LIBERO-Long benchmark [40], which corresponds to the 10 most complex tasks. Each task comprises a temporally extended sequence of interdependent subgoals, requiring the agent to perform multi-step planning and execution based solely on single-view RGB observations.

Unlike existing approaches that rely on multi-view inputs, our method uses only monocular RGB images. The LIBERO-Long tasks span diverse household environments and demand robust spatio-temporal reasoning. Task examples include:

- **KITCHEN SCENE3:** Turn on the stove and put the moka pot on it.
- **KITCHEN SCENE4:** Put the black bowl in the bottom drawer of the cabinet and close it.
- **LIVING ROOM SCENE5:** Put the white mug on the left plate and the yellow and white mug on the right plate.
- **LIVING ROOM SCENE2:** Put both the alphabet soup and the tomato sauce in the basket.
- **LIVING ROOM SCENE2:** Put both the cream cheese box and the butter in the basket.
- **LIVING ROOM SCENE6:** Put the white mug on the plate and the chocolate pudding to the right of the plate.
- **LIVING ROOM SCENE1:** Put both the alphabet soup and the cream cheese box in the basket.
- **KITCHEN SCENE8:** Put both moka pots on the stove.
- **KITCHEN SCENE6:** Put the yellow and white mug in the microwave and close it.
- **STUDY SCENE1:** Pick up the book and place it in the back compartment of the caddy.

Each task is provided with 50 expert demonstrations, enabling consistent and reproducible evaluation across methods.

7.2 Details of Baselines

We compare MonoLift against multiple baselines:

- **Single-view RGB methods (direct mapping):**

These methods illustrate that direct policy learning from monocular RGB observations is feasible, even without incorporating any form of 3D data. They serve as lightweight and deployable baselines in real-world settings.

MT-ACT [12] is an advanced Transformer-based encoder-decoder architecture that extends the Action-Chunking Transformer to multi-task learning.

RT-1 [11] is a Transformer-based multi-task policy learning framework that models actions as discrete classes by uniformly discretizing the action space into bins.

- **Single-view RGB methods (learned 3D cues):**

These methods demonstrate that even in the absence of explicit 3D input, meaningful 3D cues can be learned through predictive modeling or large-scale visual pretraining.

GROUND [9] first predicts future visual observations and then trains a goal-conditioned policy to generate the actions required to realize each frame of the synthesized video sequence.

MT-R3M [34] utilizes R3M as a pretrained visual encoder to process observation images, coupled with a GPT-style Transformer to predict actions. R3M [31], pretrained on large-scale human-centric video datasets, has been shown to improve spatial understanding and enhance the performance of downstream robotic manipulation policy learning.

- **Explicit 3D input methods:**

These methods highlight the benefits of directly incorporating 3D inputs, enabling more accurate spatial reasoning and stronger 3D priors for embodied perception and control.

3D-VLA [28] introduces a new family of 3D vision-language-action foundation models that unifies 3D perception, reasoning, and action generation via a generative world model.

SPA [2] proposes a novel representation learning framework that emphasizes the critical role of 3D spatial awareness in embodied AI. It leverages differentiable neural rendering over multi-view images to endow vision transformers with an inherent capability for spatial understanding.

7.3 Details of Implementations

Our MonoLift framework is trained on a single NVIDIA A800 GPU. The complete list of hyperparameters used in our experiments is provided in Table 3.

Table 3: List of hyperparameters.

Method	Parameter	Value
MonoLift	Learning Rate	$1e^{-4}$
	Image size	128 x 128 x 3 (LIBERO-90) 84 x 84 x 3 (Meta-World) 128 x 128 x 3 (LIBERO-Long)
	Input type	1 RGB camera view (LIBERO-90) 1 RGB camera view (Meta-World) 1 RGB camera view and robot proprioception (LIBERO-Long)
	Batch size	64
	Optimizer	Adam
	Hidden dim	256
	Attention fusion	1 layers and 4 heads
	Transfromer	8 layers and 4 heads
	Action head	2-layer MLP
	History length	5

7.4 Why Distillation Works Better: Comparison with Auxiliary Depth Regression.

To further evaluate the effectiveness of our proposed distillation strategy in introducing structural awareness, we conduct a comparative study involving two baselines: (i) *RGB-only*, which learns policies from monocular images without any structural guidance; and (ii) *AuxDepth*, which introduces structural guidance via an auxiliary depth regression objective. Empirical results on the LIBERO-90 reveal a progressive improvement across the evaluated methods: *RGB-only* \rightarrow *AuxDepth* \rightarrow *MonoLift* (53.7 \rightarrow 67.8 \rightarrow 80.8 in success rate). While *AuxDepth* offers modest gains over the RGB-only baseline, the improvements remain limited. To better understand the underlying factors behind this performance gap, we further evaluate the models using three metrics: (1) **Depth Correlation Score**: A higher value indicates stronger geometric awareness. We freeze the encoder of the trained model and append a MLP to map policy features to the depth features, allowing the MLP to remain trainable during this process. The Euclidean distance between the mapped features is computed, and its inverse is used as the score. (2) **Feature Uniformity**: A higher score indicates a more dispersed and discriminative representation. Following Wang et al. [42], it quantifies the evenness of feature distribution on the unit hypersphere using a Gaussian potential function. (3) **Temporal Difference** captures the model’s sensitivity to dynamic information. It is computed as the average Euclidean distance between feature embeddings at consecutive time steps. As shown in Figure 8, *AuxDepth* achieves a high depth correlation score, indicating that it captures structural cues through regression. However, it shows significantly lower scores in both feature uniformity and temporal difference, suggesting that its representations become overly compact and less responsive to temporal variation. This compression hampers the ability to distinguish between different states and adapt to dynamic changes. These limitations could explain the modest performance gains observed with *AuxDepth*, despite its alignment with depth features. In contrast, the results of *MonoLift* demonstrate that distillation effectively injects structural awareness without compromising feature quality.

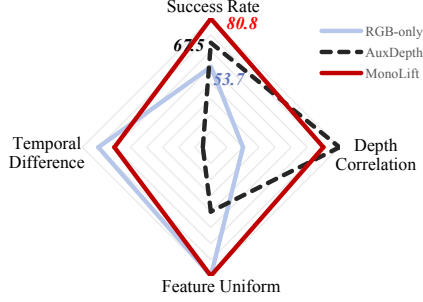


Figure 8: Comparison of feature-level metrics across different methods.

7.5 MonoLift Outperforms the Teacher Model

We evaluate the teacher model trained and tested directly with RGB and depth inputs, without any distillation. As shown in Table 4, our proposed MonoLift consistently outperforms the teacher across all benchmarks. This improvement can be attributed to two key factors: (i) The proposed tri-level distillation effectively transfers structured spatial, temporal, and action-level knowledge from the depth-guided teacher to the student model; (ii) The teacher model, which directly uses RGB+Depth during inference, performs worse due to per-frame errors introduced by predicted depth that accumulate over long sequences. In contrast, our method avoids relying on predicted depth at inference and instead transfers high-level spatiotemporal and behavioral representations rather than low-level pixel details, thereby mitigating depth noise and enhancing overall robustness.

Table 4: Comparison between the RGB+Depth teacher and our MonoLift.

Method	Libero-90 (%)	Libero-Long (%)	Meta-World (%)
RGB+Depth Teacher	70.3±3.7	53.0±2.4	85.4±2.6
MonoLift (ours)	80.8±3.3	71.7±1.1	87.8±2.3

7.6 Effect of Teacher Model Quality on Student Performance

To further investigate how the quality of the teacher model influences student learning, we evaluate several teacher variants equipped with pretrained depth estimators of different model scales from *Depth Anything v2* [14]. As shown in Table 5, improvements in the teacher’s depth estimation accuracy consistently lead to stronger student policies after distillation. As the capacity of the teacher’s depth estimator increases from small to large, the student’s policy exhibits a clear and consistent improvement. These results reveal a significant positive correlation between teacher quality and student effectiveness. When the teacher possesses more accurate 3D geometric perception, the student benefits from richer spatial representations and depth cues.

Table 5: Influence of teacher model quality on student performance.

Teacher Configuration	NYU- δ_1 \uparrow	NYU-AbsRel \downarrow	Success Rate (%)
No depth (RGB-only baseline)	—	—	53.7
With Depth Estimator (S)	0.961	0.073	67.5 (+13.8)
With Depth Estimator (B)	0.977	0.063	70.0 (+16.3)
With Depth Estimator (L)	0.984	0.056	80.8 (+27.1)

7.7 Effect of Policy Head Architecture

To examine the influence of the policy head architecture on both performance and efficiency, we compare the lightweight MLP head adopted in MonoLift with a diffusion-style policy head inspired by Diffusion Policy [26]. Both variants employ the same backbone and training configuration, differing solely in the design of the policy head.

As summarized in Table 6, the diffusion-style head achieves a higher success rate on *Libero-Spatial* compared to the MLP head. However, the improvement is not consistent across all tasks, and the diffusion head incurs substantial inference latency, increasing the runtime from 18.1 ms (MLP) to 221.6 ms. These findings suggest that while more expressive policy heads can enhance spatial reasoning in certain scenarios, they also come with a considerable computational cost. Notably, recent advances in efficient diffusion inference [43, 44, 45] may further narrow this computational gap, and incorporating such techniques into MonoLift represents a promising avenue for future research.

Table 6: Comparison of different policy heads in MonoLift.

Policy Head	Libero-Goal (%)	Libero-Spatial (%)	Libero-Object (%)
MonoLift (MLP)	85.3±1.2	78.6±0.9	96.3±1.2
MonoLift (Diffusion Policy)	80.0±1.1	82.5±0.8	96.2±1.4

7.8 Comparison of Our Distillation Framework with Feature and Sequential Distillation

To validate the effectiveness of the proposed distillation framework, we compare two alternative settings: (1) Feature Distillation, which transfers spatial representations from the teacher to the student at the feature level only; and (2) Sequential Distillation, where the teacher is first trained independently and then kept frozen during the student’s training, providing fixed supervision without adaptation. As shown in Table 7, our method significantly outperforms both alternatives. Feature distillation weakens the coupling between perception and action, making it difficult to capture temporal dependencies. Sequential distillation, on the other hand, suffers from the limitation of static supervision—since the frozen teacher cannot adjust its guidance to the student’s evolving learning dynamics. Moreover, in the cross-modal distillation setting (RGB+Depth→RGB), the teacher and student operate on different input modalities, often leading to mismatched feature distributions. In contrast, our joint distillation approach enables both teacher and student to co-adapt their representations during optimization, effectively bridging the modality gap while jointly enforcing spatial, temporal, and action consistency. This dynamic teacher–student interaction leads to more stable and efficient policy learning.

Table 7: Effectiveness of our distillation framework.

Benchmark	No Distillation	Feature Distillation	Sequential Distillation	MonoLift
Libero-Long (%)	46.5 ± 1.3	61.0 ± 1.6	57.3 ± 3.6	71.7 ± 1.1