2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Best of Sim and Real: Decoupled Visuomotor Manipulation via Learning Control in Simulation and Perception in Real

Jialei Huang[1,3], Zhaoheng Yin[2], Yingdong Hu[1], Shuo Wang[1], Xingyu Lin[2] and Yang Gao[1,3*]

Abstract— **Sim-to-real transfer remains a fundamental challenge in robot manipulation due to the entanglement of perception and control in end-to-end learning. We present a decoupled framework that learns each component where it is most reliable: control policies are trained in simulation with privileged state to master spatial layouts and manipulation dynamics, while perception is adapted only at deployment to bridge real observations to the frozen control policy. Our key insight is that control strategies and action patterns are universal across environments and can be learned in simulation through systematic randomization, while perception is inherently domain-specific and must be learned where visual observations are authentic. Unlike existing end-to-end approaches that require extensive real-world data, our method achieves strong performance with only 10-20 real demonstrations by reducing the complex sim-to-real problem to a structured perception alignment task. We validate our approach on tabletop manipulation tasks, demonstrating superior data efficiency and out-of-distribution generalization compared to end-to-end baselines. The learned policies successfully handle object positions and scales beyond the training distribution, confirming that decoupling perception from control fundamentally improves sim-to-real transfer.**

I. INTRODUCTION

Simulation environments provide a safe, scalable, and cost-effective platform for robot learning [6]. We can run thousands of robots in parallel, automatically reset environments, and access perfect state information in simulation, making large-scale interactive learning feasible. However, reliably deploying simulation-trained policies to the real world remains a central challenge in robot learning [1], [2], [3]. This challenge not only concerns technical feasibility but fundamentally determines whether robot learning methods can transition to practical applications. The difficulty of sim-to-real transfer arises from the entanglement of perception and control.

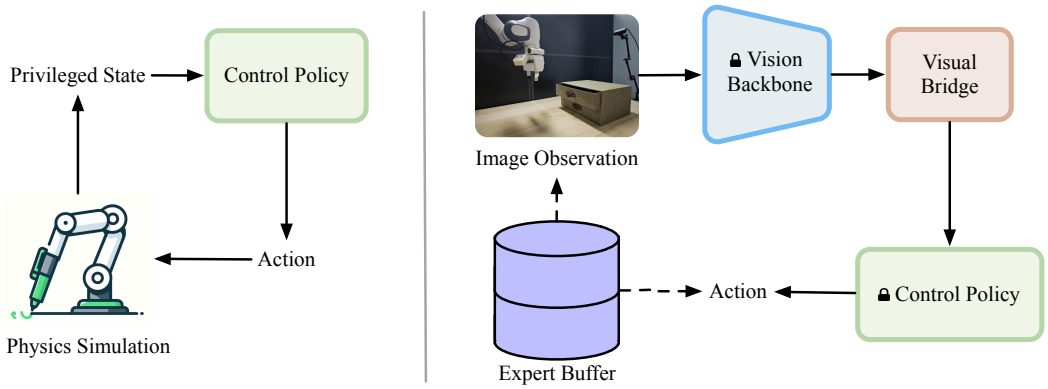
In end-to-end learning paradigms, policies must simultaneously handle visual domain shift and dynamics gap. The former arises from discrepancies in texture, lighting, and appearance between simulated and real images, while the latter stems from differences in physical parameters such as friction, mass, and contact forces between simulation and reality [1], [2], [3], [21], [22]. This dual challenge creates a compounding effect: perceptual uncertainties interact with control uncertainties, making the sim-to-real gap grow multiplicatively rather than additively. At deployment, this entanglement makes policies brittle to real-world variations, often requiring extensive fine-tuning to achieve usable performance.

Existing sim-to-real methods predominantly adopt end-to-end learning paradigms, attempting to train unified networks that directly map from pixels to actions [13], [14], [15], [16], [17], [18], [23], [24], [26]. While techniques like domain randomization have improved transfer performance [1], [2], [3], they do not address the fundamental issue of perception-control entanglement. When perception and control are learned jointly, the network must master both visual feature extraction and control strategy generation simultaneously. This coupling means that even simple control behaviors require substantial data to learn under varying visual conditions [21], [22], as the network cannot separate task-relevant control patterns from domain-specific visual features. Moreover, errors in either perception or control can propagate through the network, making failures difficult to diagnose and correct.

While privileged state learning has shown success in locomotion through teacher-student distillation [7], [19], [25], manipulation presents distinct challenges. Unlike locomotion where proprioception often suffices, manipulation critically depends on precise visual perception for object localization and grasping [8], [9], [10], [11], [12]. Rather than distilling both perception and control in simulation, we advocate learning perception only in the target domain where visual observations are authentic. Our core insight can be summarized as: **control is consistent, perception is specific**. Control strategies and action patterns are governed by consistent physical principles that remain invariant across environments, and they can therefore be effectively acquired in simulation through systematic randomization [1], [2], [3], [5], [6]. In contrast, perception is intrinsically tied to domain characteristics, since lighting conditions, textures, and visual appearances vary substantially across deployment scenarios and cannot be fully captured in simulation [8], [10], [11], [12]. This perspective motivates a rethinking of sim-to-real transfer: instead of pursuing end-to-end policies that attempt to handle all variations simultaneously, we advocate decomposing the problem and training each component in the setting where it can be learned most effectively [20].

Based on this insight, we propose **Best of Sim and Real (BSR): learn control where physics is accessible, adapt perception where visual observations are realistic.** In the first stage, we train control policies in simulation using privileged state—perfect object poses and spatial relation-

*Corresponding author
[1]Jialei Huang, Yingdong Hu, Shuo Wang and Yang Gao are with Tsinghua University, Beijing, China
[2]Zhaoheng Yin and Xingyu Lin are with University of California, Berkeley, CA, USA
[3]Jialei Huang and Yang Gao are with Shanghai Qi Zhi Institute, Shanghai, China



(a). Stage 1: Control Learning with Privileged State **(b). Stage 2: Visual Bridge Learning**

Fig. 1. Overview of our Best of Sim and Real (BSR) framework. (a) Stage 1: Control learning with privileged state in physics simulation, where the policy learns robust action patterns through systematic domain randomization. (b) Stage 2: Visual bridge learning in the real world, where a lightweight network maps image observations to the frozen control policy's input space using expert demonstrations stored in a replay buffer.

ships—allowing the policy to focus purely on learning robust control strategies through systematic randomization [1], [2], [4], [5], [6]. In the second stage, we freeze the control policy and train only a lightweight visual bridge that maps real observations to the policy's expected input space [10], [11], [12]. This decomposition transforms the complex sim-to-real problem into two well-defined subproblems: learning universal control patterns in simulation and solving a structured perception alignment task in the real world.

We validate our approach on tabletop manipulation tasks, demonstrating superior data efficiency and out-of-distribution generalization compared to end-to-end baselines [13], [14], [15], [16], [17], [18]. The learned policies generalize successfully to object positions and scales beyond the training distribution, highlighting that decoupling perception from control substantially improves sim-to-real transfer. More importantly, the learned policies exhibit strong generalization capabilities, handling object positions and scales outside the training distribution. Our main contributions are threefold: (1) proposing a new sim-to-real paradigm for manipulation that decouples perception and control, fundamentally reducing real-world data requirements; (2) designing a two-stage training framework based on privileged state that maximizes the advantages of both simulation and real environments; (3) systematic experimental validation of our method's advantages in data efficiency and generalization capability, with detailed ablation studies.

II. METHOD

A. Overview

Since control is governed by invariant physical laws while perception is tied to environment-specific factors, we reformulate sim-to-real transfer as two separable subproblems, solvable in their respective domains.

We structure our framework into two complementary stages, as illustrated in Figure 1. In the first stage (Section II-B), we exploit the perfect state observability in simulation to train a control policy that masters the geometric and dynamic patterns of manipulation tasks. This policy learns from privileged state information—precise object poses, contact states, and relative transformations—allowing it to focus purely on the control problem without the confounding effects of perceptual noise. Once trained, this policy is frozen and serves as a fixed control backbone. In the second stage (Section II-C), we address perception exclusively in the real world where visual observations are genuine. Rather than attempting to learn visual features in simulation where appearance realism is limited, we train a lightweight visual bridge network using a small number of real-world demonstrations to map camera observations to the representation space expected by the frozen control policy. This decoupling transforms the complex sim-to-real problem into a structured perception task with a clear learning target.

B. Stage 1: Control Learning with Privileged State

By providing the policy with ground-truth state information during simulation training, we enable it to focus exclusively on learning robust control strategies without perceptual uncertainties. Our privileged state representation consists of the robot's proprioceptive information including joint angles and velocities, the end-effector's 6-DoF pose, and task-relevant geometric state such as object poses expressed as relative transformations—the transformation from end-effector to object, the distance to grasp points, and approach alignments. This relative representation ensures that learned control patterns generalize across different spatial configurations. The policy network $\pi_\theta(a|s)$ then outputs action distributions for robot control.

To ensure robust transfer to the real world, we employ systematic domain randomization by incrementally intensify the extent of variation. Early in training, the policy experiences minor variations in object placement and physical parameters, allowing it to first acquire the basic task structure. As training progresses, we gradually expand both the dimensionality and magnitude of randomization. Geometric randomization encompasses variations in object positions, orientations, and scales, as well as diverse initial

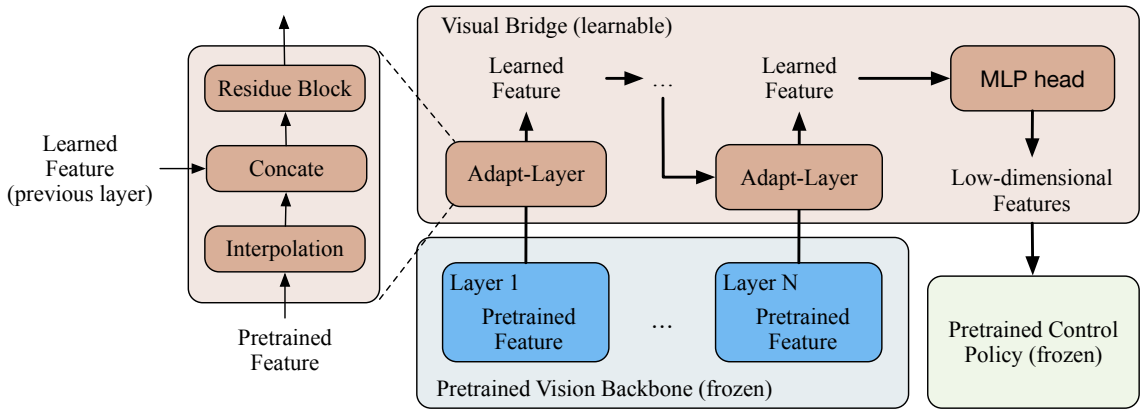


Fig. 2. Architecture of the visual bridge network. Multi-layer pretrained features from a frozen vision backbone are progressively refined through adaptive layers and residual blocks, then combined into low-dimensional features via an MLP head before being passed to the frozen control policy.

robot configurations. Physical randomization includes noise injection in observation and action spaces, variations in mass and friction coefficients, and explicit modeling of control delays. This progressive strategy ensures the policy develops robustness while maintaining stable learning dynamics. The simulation training leverages parallel simulators [6] to generate diverse experience efficiently, training the policy using PPO [5] with task-specific reward functions.

C. Stage 2: Visual Bridge Learning

At the deployment phase, we face a structured learning problem: aligning real-world visual observations with the state representation required by the fixed control policy. This formulation fundamentally changes the nature of sim-to-real transfer from an unbounded reinforcement learning problem to a supervised learning task with a clear target. The frozen control policy acts as a strong prior, embodying the geometric and dynamic knowledge necessary for manipulation, while the perception problem is reduced to learning an appropriate visual bridge between observations and the control-relevant state space.

Our visual bridge leverages pretrained visual representations to provide strong perceptual priors, crucial for data-efficient learning. We employ a frozen vision backbone, specifically a Vision Transformer pretrained with self-supervised objectives like DINOv2 [8], which provides rich visual features without task-specific training. As illustrated in Figure 2, we extract intermediate representations from multiple layers of the network. These multi-scale features capture both high-level semantic information and fine-grained spatial details, which are essential for robot manipulation.

The bridge network processes these multi-layer features through adaptive layers that project them to a common dimensionality, followed by spatial alignment through bilinear interpolation to handle varying feature map resolutions [56]. We progressively fuse these representations through shallow residual blocks, where residual connections ensure stable gradient flow [23]. Then the fused visual representation is combined with proprioceptive signals through an MLP head, producing low-dimensional features suitable for the control policy.

Training the visual bridge requires only a small set of real-world demonstrations, typically 10-20 trajectories per task. During these demonstrations, we record synchronized camera images from both third-person and wrist-mounted viewpoints, proprioceptive readings, and expert actions. The bridge is trained to minimize the L2 distance between expert actions and those produced by the frozen control policy:

$$\mathcal{L} = \mathbb{E}_{(o_t, a_t^*) \sim \mathcal{D}_{\text{demo}}} \left[\| a_t^* - \pi_\theta(f_\phi(o_t)) \|^2 \right] \tag{1}$$

where f_ϕ denotes the visual bridge, o_t represents the observation, and a_t^* is the expert action. This end-to-end supervision ensures the bridge learns to extract precisely the state information needed for successful task execution. The visual bridge contains only a small number of learnable parameters, focusing the learning problem and preventing overfitting on limited data.

D. Best of Sim and Real Pipeline

Our complete BSR pipeline coordinates the two-stage learning process to maximize both simulation efficiency and real-world performance. During the simulation phase, we establish task structure and reward design, then progressively introduce randomization following the curriculum described in Section II-B. The policy typically converges within 10-20 million environment steps through parallel simulation. Once the policy demonstrates robust performance across the full randomization distribution, we freeze all parameters and export the trained model for deployment.

The real-world deployment phase begins with collecting expert demonstrations. Our system employs a dual-camera setup with calibrated extrinsics, maintaining consistent geometry between simulation and real deployment to minimize spatial domain shift. During visual bridge training, we apply data augmentation including random crops and color jittering to improve generalization. The complete pipeline enables practical deployment of new manipulation tasks with minimal real-world data requirements, as we demonstrate in Section III.

(a). Stacking Cube (b). Opening Drawer (c). Closing Door

Fig. 3. Manipulation tasks used for evaluation. (a) **Stacking Cube**: Pick and place a cube onto a target platform, requiring precise grasp and placement within a 20×20cm workspace. (b) **Opening Drawer**: Localize the handle, grasp, and pull to open the drawer by at least 15cm, demanding accurate visual servoing and force control. (c) **Closing Door**: Push a hinged door from 90° open to fully closed while maintaining continuous contact, testing the policy's ability to handle constrained motion and contact dynamics.

III. Experiments

We design our experiments to validate three core claims: (1) decoupling perception and control fundamentally improves sim-to-real transfer efficiency, (2) learning control with privileged state in simulation provides superior spatial generalization compared to end-to-end learning, and (3) our visual bridge design enables effective transfer with minimal real-world data. We evaluate on three manipulation tasks that require precise visual-motor coordination: **Stacking Cube** (pick and stack objects with varying sizes), **Opening Drawer** (localize handle and execute constrained trajectory), and **Closing Door** (swing door requiring continuous contact control).

A. Experimental Setup

Tasks and Metrics. We evaluate our approach on three manipulation tasks illustrated in Figure 3. For Stacking Cube (Fig. 3a), robots must pick up a cube and stack it on a target platform within a 20cm×20cm workspace. Opening Drawer (Fig. 3b) requires localizing and grasping a handle, then pulling to open a drawer by at least 15cm. Closing Door (Fig. 3c) involves pushing a hinged door from 90° open to fully closed while maintaining continuous contact. We report success rate as the primary metric and additionally use a graded completion score 0-4 to capture partial task progress for detailed analysis. For Stacking Cube, the completion score indicates: 0 (failure), 1 (approach), 2 (grasp), 3 (move to target), and 4 (successful stack).

Baselines. We compare against three ablations that isolate key design choices: (1) *w/o pretrained state policy*: End-to-end learning from pixels to actions without leveraging simulation-trained control; (2) *w/o visual bridge*: Direct state regression from images instead of representation-level bridging; (3) *pure BC*: Behavior cloning without pretrained vision encoder, learning visual features from scratch. All methods use identical network capacity, training budgets, and data augmentation for fair comparison.

B. Sample Efficiency in Real World

Figure 4 presents our main result on sample efficiency. With only 10 real-world demonstrations per task, our method achieves 73.3% success on Stacking Cube, 43.3% on Opening Drawer, and 88.3% on Closing Door. In contrast, the strongest baseline (w/o pretrained state policy) only reaches 20.0%, 1.7%, and 50.0% respectively—a gap of 30-50 percentage points. This significant difference validates our core thesis: by learning control in simulation and only adapting perception in the real world, we fundamentally reduce the complexity of real-world learning.

The performance gap persists but narrows as more demonstrations become available. At $K = 80$, end-to-end learning approaches our performance on some tasks, but requires 4×-8× more data to reach the same success rate we achieve at $K = 10$-20. Notably, removing the visual bridge (green curve) severely degrades performance across all data regimes, confirming that representation-level bridging is crucial for connecting visual observations to the frozen control policy. Pure behavior cloning without pretrained encoders (red curve) shows the poorest performance, emphasizing the importance of visual priors for few-shot learning.

C. Spatial Generalization Beyond Training

A key advantage of learning universal control principles in simulation is enhanced spatial generalization. To rigorously assess this capability, we conduct extensive experiments on the Stacking Cube task, expanding the workspace from the original 20cm×20cm training region to a 40cm×40cm evaluation area, representing a fourfold increase in area. We define in-distribution (ID) performance as success within the original training region, and out-of-distribution (OOD) performance as success in the outer regions beyond the training boundary.

Figure 5 provides detailed spatial visualizations of task completion scores across the extended workspace. Our method achieves 75% ID success rate and maintains 35% OOD success rate, demonstrating meaningful generalization beyond the training distribution. The heatmap reveals that our

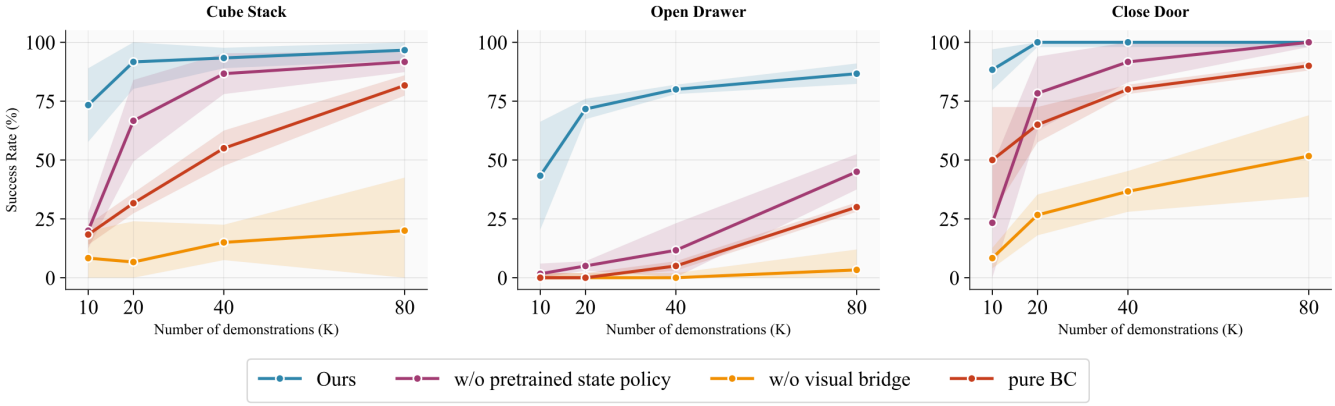


Fig. 4. Success rate as a function of real-world demonstrations. Our method achieves strong performance with just 10–20 demonstrations, while baselines require substantially more data or fail to reach comparable performance.

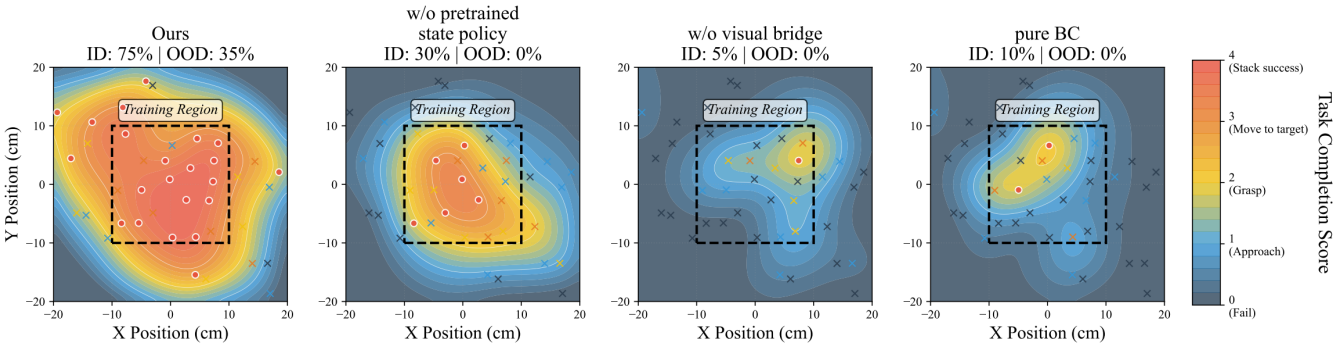


Fig. 5. Spatial visualization of task completion scores for Stacking Cube as the workspace expands from 20×20cm training region (black dashed box) to 40×40cm evaluation area. Heatmaps show interpolated completion scores (0-4 scale), with circles indicating successful trials and crosses marking failures. Our method maintains high performance (ID: 75%, OOD: 35%) across the extended workspace, while baselines show rapid degradation beyond the training boundary.

approach maintains high completion scores (3-4, indicating successful grasping and stacking) throughout most of the workspace, with gradual degradation toward extreme corners. This spatial consistency confirms that the privileged state policy learns genuine geometric and dynamic patterns rather than memorizing specific configurations.

In stark contrast, baseline methods show catastrophic failure beyond the training boundary. The end-to-end approach without pretrained state policy drops from 30% ID to 0% OOD success, with completion scores rapidly declining to 0-1 (failure to approach or grasp) outside the training region. Similarly, methods without visual bridge or pretrained encoders achieve only 5% and 10% ID success respectively, with complete failure in OOD regions. These results highlight a fundamental limitation of end-to-end learning: when perception and control are entangled, the policy overfits to the specific visual patterns seen during training, preventing generalization to new spatial configurations.

Figure 6 quantifies this generalization pattern by plotting average completion scores as a function of distance from the workspace center. The training boundary at 14.1cm (diagonal distance from center to corner of 20cm×20cm region) is marked with a dashed line. Our method exhibits graceful degradation, maintaining scores above 2.0 (successful grasping) even at 22.5cm from center. The gradual slope indicates that our decoupled approach learns distance-invariant control strategies through privileged state training.

Baselines show markedly different behavior with sharp performance cliffs near the training boundary. The w/o pretrained state policy baseline drops from 3.0 to below 0.5 beyond 15cm, indicating complete task failure. Pure BC and w/o visual bridge methods perform poorly even within the training region and show near-zero performance outside. These sharp transitions suggest that end-to-end methods learn spurious correlations between absolute positions and actions rather than relative geometric relationships that generalize.

The superior spatial generalization of our method directly stems from the decoupled training paradigm. By learning control with privileged state in simulation, the policy focuses on relative transformations—end-effector to object distances, approach angles, grasp configurations—that remain valid regardless of absolute position. The systematic domain randomization during simulation training further ensures robustness to spatial variations. Meanwhile, the visual bridge, trained with limited real demonstrations, only needs to extract these relative features from images rather than learning position-dependent control strategies.

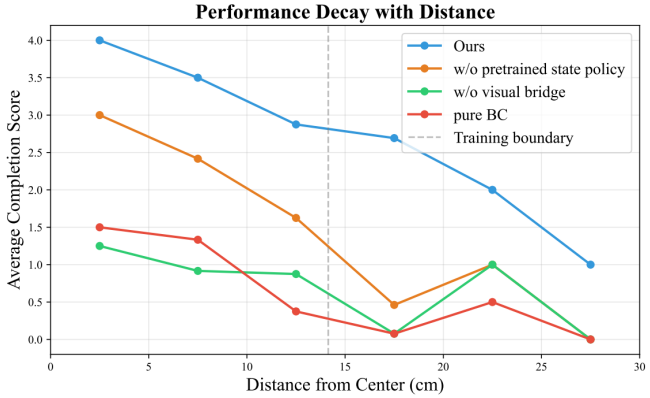


Fig. 6. Performance decay as a function of distance from the training region center. Our method shows graceful degradation with distance, maintaining an average completion score above 2.0 even at 22.5cm (beyond training boundary at 14.1cm), while baselines exhibit sharp performance cliffs.

TABLE I

ABLATION STUDY ON VISUAL BRIDGE COMPONENTS ($K = 20$)

Component	Stack	Drawer	Door	Avg
Full model	91.7	71.7	100.0	87.8
Single layer features	78.3	55.0	91.7	75.0
w/o residual connections	83.3	60.0	95.0	79.4
w/o proprioception	71.7	48.3	88.3	69.4
Direct FC mapping	65.0	41.7	85.0	63.9

D. Ablation Studies

As shown in Table I, we conduct a detailed analysis of the contribution of each visual bridge component to the model's performance. First, we observe a significant reduction in average success (12.8%) when we use only the final-layer features for the task, instead of multi-scale feature extraction. This suggests that intermediate layers play a critical role in capturing essential spatial details, which are vital for accurate manipulation across the various tasks. The reduction in success highlights the necessity of extracting spatial information at different levels of abstraction to handle complex manipulations effectively.

Next, we analyze the effect of removing residual connections, which results in an 8.4% decrease in performance. This emphasizes the role of residual connections in maintaining optimization stability, particularly with limited data. The most substantial drop (23.9%) occurs when replacing the progressive fusion architecture with direct fully-connected (FC) mapping. This outcome further validates our design choice to preserve spatial structure through gradual aggregation, as direct FC mapping fails to capture essential spatial patterns, leading to a significant degradation in task performance.

E. Discussion

Our experiments demonstrate that decoupling perception and control fundamentally changes the sample complexity of sim-to-real transfer. The 4×-8× improvement in data efficiency stems from transforming an unbounded RL problem into a structured supervised learning task. More importantly, the strong spatial generalization—maintaining 35% success rate in regions four times larger than training—confirms that simulation-trained policies with privileged state capture genuine task structure rather than memorizing configurations. The gradual performance decay with distance, as opposed to sharp cliffs in baselines, validates that our approach learns robust, distance-invariant control strategies that transfer across spatial scales. These results suggest that the key to practical sim-to-real transfer lies not in perfecting simulation fidelity, but in identifying what to learn where: control in simulation where physics is accessible, perception in the real world where appearance is authentic.

IV. RELATED WORK

A. Sim-to-Real Transfer in Robot Manipulation

Sim-to-real transfer remains a fundamental challenge in robot learning, with multiple approaches proposed to bridge the gap between simulation and real physical world. Domain randomization (DR) methods [1], [2], [3] train policies on diverse simulated environments to achieve effective transfer. Tobin et al. [1] pioneered this approach by randomizing visual parameters, while Peng et al. [2] extended it to dynamics randomization. Furthermore, Chebotar et al. [3] proposed closing the sim-to-real loop by adapting randomization parameters based on real-world experience. Recent extensions [27], [28] have shown that curriculum learning-like design in randomization improves transfer, with [29] demonstrating impressive dexterous manipulation through systematic randomization and [30] solving complex tasks like Rubik's cube.

Different from DR methods, domain adaptation techniques attempt to align simulation and real distributions through learnable strategies. Asymmetric architectures [4] separate observation and state processing for improved transfer. Recent work explores adversarial domain adaptation [31], style transfer [32], and progressive networks [33] that prevent catastrophic forgetting. Meta-learning approaches [34], [35] enable rapid adaptation, while system identification methods [36], [37] estimate real-world parameters to improve simulation fidelity.

Hybrid approaches combine multiple strategies for robustness. [38] demonstrated sim-to-sim transfer before real deployment, while [39] showed that self-supervised adaptation improves visual robotic manipulation. Despite these advances, existing methods treat perception and control as entangled problems, requiring extensive real-world data. Our work fundamentally differs by completely decoupling these components across domains, learning control with privileged state in simulation while adapting only perception in the real world with minimal demonstrations.

B. Learning with Privileged Information and Visual Representations

Privileged information has emerged as a powerful paradigm in robot learning. Kumar et al. [7] demonstrated rapid motor adaptation for legged robots through teacher-student distillation from privileged to deployable policies. Recent work [19] extended this with privileged sensing

scaffolds that guide reinforcement learning. Zeng et al. [25] applied similar concepts to visual legged locomotion, learning world models from privileged state. In manipulation, [40] leveraged privileged depth information for improved learning, while [41] used privileged force feedback for contact-rich tasks. However, these approaches typically distill all components in simulation, missing the opportunity to learn perception where visual observations are authentic.

The quality of visual representations critically impacts manipulation performance. Self-supervised pretraining has produced certain powerful foundation models. R3M [10] learns from human interaction videos, VC-1 [11] provides universal visual representations for embodied AI, while SUGAR [12] pre-trains 3D representations specifically for robotics. DINOv2 [8] offers robust features through self-supervised learning, building on the Vision Transformer architecture [9]. Recent work includes MVP [42] using masked autoencoding, CLIP [43] for vision-language alignment, and specialized representations for robotic tasks [44], [45].

Our approach uniquely leverages these pretrained representations as a bridge between real observations and simulation-trained control policies. Unlike prior work that uses pretrained models within end-to-end frameworks [13], [14], [15], [16], we employ them specifically for perception alignment, transforming the complex sim-to-real problem into a structured supervised learning task requiring only 10-20 real demonstrations.

C. End-to-End Visuomotor Learning and Modular Approaches

End-to-end visuomotor learning has achieved impressive results through scale and architectural innovations. Previous works use large-scale datasets to enable training generalizable policies: BridgeData V2 [14] provides diverse manipulation demonstrations, DROID [15] offers in-the-wild robot data, and RoboNet [46] enables multi-robot learning. Vision-language-action models have emerged as a powerful paradigm, with RT-2 [16] transferring web knowledge to robotic control, RT-1 [47] demonstrating real-world scaling, and PaLM-E [24] providing embodied multimodal capabilities. OpenVLA [26] advances general-purpose vision-language-action models, while cross-embodiment initiatives like Open X-Embodiment [17] and Octo [18] pursue generalist policies through massive data aggregation.

Architectural innovations have improved end-to-end learning efficiency. Transformer-based policies [48], [49] model long-horizon dependencies, while Perceiver-based architectures [50] handle multimodal inputs effectively. Data augmentation techniques partially address sample efficiency: RAD [21] showed simple augmentations improve RL, while DrQ-v2 [22] achieved state-of-the-art continuous control through aggressive augmentation. However, the fundamental coupling of perception and control limits their efficiency compared to our decoupled approach.

Modular approaches have explored various decomposition strategies. Rizzardo et al. [20] proposed latent prediction for non-prehensile manipulation, while [51] separated visual encoding from policy learning. Hierarchical decomposition [52], [53] addresses long-horizon tasks, and skill-based methods [54], [55] enable compositional learning. Our work advances modularity by completely decoupling perception and control across domains, achieving 4-8× better data efficiency than end-to-end baselines while maintaining strong spatial generalization—a key advantage demonstrated by our policies successfully handling workspaces four times larger than the training region.

V. CONCLUSION

In this paper, we proposed a decoupled framework for sim-to-real transfer in manipulation, which separates perception and control learning. Our key insight that control principles are universal while perception is domain-specific drives a two-stage approach. By training control policies with privileged state in simulation and adapting only perception in the real world, we transform the complex sim-to-real problem into a structured perception alignment task.

Our experiments demonstrate the effectiveness of this decoupling. With only 10-20 real-world demonstrations, our method achieves performance comparable to end-to-end approaches using $4\times$-$8\times$ more data. More importantly, the learned policies exhibit strong spatial generalization, maintaining a 35% success rate in workspaces four times larger than the training region—something end-to-end methods struggle with. The gradual performance decay and contrast with baseline methods confirm that privileged state training enables robust, distance-invariant control strategies.

Beyond data efficiency, our approach enables modularity and stability by freezing the control policy after simulation training. This allows rapid deployment to new environments without retraining the entire system, addressing a key limitation of current sim-to-real methods. The structured perception learning problem also makes debugging and improvement easier compared to opaque end-to-end systems. Future work can extend this framework to more complex tasks, such as mobile manipulation and multi-arm coordination, and explore techniques for reducing real-world data requirements even further.

ACKNOWLEDGMENT

This work was supported by the Shanghai Qi Zhi Institute & Spirit AI Innovation Program and the Tsinghua University Dushi Program.

REFERENCES

[1] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2017.

[2] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Sim-to-Real Transfer of Robotic Control with Dynamics Randomization," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2018.

[3] Y. Chebotar, A. Handa, V. Makoviychuk, M. Macklin, J. Issac, N. Ratliff, and D. Fox, "Closing the Sim-to-Real Loop: Adapting Simulation Randomization with Real World Experience," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2019.

[4] L. Pinto, M. Andrychowicz, P. Welinder, W. Zaremba, and P. Abbeel, "Asymmetric Actor-Critic for Image-Based Robot Learning," in *Proc. Robot.: Sci. Syst. (RSS)*, 2018.

[5] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," *arXiv:1707.06347*, 2017.

[6] V. Makoviychuk et al., "Isaac Gym: High Performance GPU-Based Physics Simulation for Robot Learning," in *Proc. NeurIPS Datasets and Benchmarks*, 2021.

[7] A. Kumar, Z. Fu, D. Pathak, and J. Malik, "RMA: Rapid Motor Adaptation for Legged Robots," in *Proc. Robot.: Sci. Syst. (RSS)*, 2021.

[8] M. Oquab, T. Darcet, et al., "DINOv2: Learning Robust Visual Features without Supervision," *arXiv:2304.07193*, 2023.

[9] A. Dosovitskiy et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[10] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3M: A Universal Visual Representation for Robot Manipulation," in *Proc. Conf. Robot Learning (CoRL)*, PMLR 205, 2022.

[11] A. Majumdar et al., "Where Are We in the Search for an Artificial Visual Cortex for Embodied Intelligence? (VC-1)," *arXiv:2303.18240*, 2023.

[12] S. Chen et al., "SUGAR: Pre-training 3D Visual Representations for Robotics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

[13] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion Policy: Visuomotor Policy Learning via Action Diffusion," *Int. J. Robot. Res.*, 2024.

[14] H. R. Walke et al., "BridgeData V2: A Dataset for Robot Learning at Scale," in *Proc. Conf. Robot Learning (CoRL)*, PMLR 229, 2023.

[15] A. Khazatsky et al., "DROID: A Large-Scale In-the-Wild Robot Manipulation Dataset," *arXiv:2403.12945*, 2024.

[16] A. Brohan et al., "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," *arXiv:2307.15818*, 2023.

[17] Open X-Embodiment Collaboration, "Open X-Embodiment: Robotic Learning Datasets and RT-X Models," *arXiv:2310.08864*, 2023.

[18] D. Ghosh, C. Agia, T. Z. Zhao, et al., "Octo: An Open-Source Generalist Robot Policy," in *Proc. Robot.: Sci. Syst. (RSS)*, 2024.

[19] E. S. Hu, J. Springer, O. Rybkin, and D. Jayaraman, "Privileged Sensing Scaffolds Reinforcement Learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.

[20] C. Rizzardo, F. Carlucci, and D. Calandriello, "Sim-to-Real via Latent Prediction: Transferring Visual Non-Prehensile Manipulation Policies," *Frontiers in Robotics and AI*, 2023.

[21] M. Laskin, K. Lee, A. Stooke, L. Pinto, P. Abbeel, and A. Srinivas, "Reinforcement Learning with Augmented Data (RAD)," in *Proc. NeurIPS*, 2020.

[22] D. Yarats, R. Fergus, A. Lazaric, and L. Pinto, "Mastering Visual Continuous Control: Improved Data-Augmented RL (DrQ-v2)," *arXiv:2107.09645*, 2021.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016.

[24] J. Driess et al., "PaLM-E: An Embodied Multimodal Language Model," *arXiv:2303.03378*, 2023.

[25] J. Zeng, C. Luo, Y. Hu, X. Chen, Z. Tu, and H. Su, "World Model-based Perception for Visual Legged Locomotion," *arXiv:2409.16784*, 2024.

[26] OpenVLA Team, "OpenVLA: General-Purpose Vision-Language-Action Models for Robotics," *arXiv*, 2024.

[27] J. Matas, S. James, and A. J. Davison, "Sim-to-Real Reinforcement Learning for Deformable Object Manipulation," in *Proc. Conf. Robot Learning (CoRL)*, 2018.

[28] F. Sadeghi and S. Levine, "CAD2RL: Real Single-Image Flight Without a Single Real Image," in *Proc. Robot.: Sci. Syst. (RSS)*, 2017.

[29] O. M. Andrychowicz et al., "Learning Dexterous In-Hand Manipulation," *Int. J. Robot. Res.*, vol. 39, no. 1, pp. 3-20, 2020.

[30] I. Akkaya et al., "Solving Rubik's Cube with a Robot Hand," *arXiv:1910.07113*, 2019.

[31] K. Bousmalis et al., "Using Simulation and Domain Adaptation to Improve Efficiency of Deep Robotic Grasping," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2018.

[32] S. James, P. Wohlhart, M. Kalakrishnan, and P. Abbeel, "Sim-to-Real via Sim-to-Sim: Data-efficient Robotic Grasping via Randomized-to-Canonical Adaptation Networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019.

[33] A. A. Rusu et al., "Progressive Neural Networks," *arXiv:1606.04671*, 2016.

[34] T. Yu et al., "Meta-World: A Benchmark and Evaluation for Multi-Task and Meta Reinforcement Learning," in *Proc. Conf. Robot Learning (CoRL)*, 2019.

[35] C. Finn, P. Abbeel, and S. Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017.

[36] F. Muratore, F. Treede, M. Gienger, and J. Peters, "Domain Randomization for Simulation-Based Policy Optimization with Transferability Assessment," in *Proc. Conf. Robot Learning (CoRL)*, 2018.

[37] W. Yu et al., "Sim-to-Real Transfer for Biped Locomotion," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2019.

[38] X. B. Peng, E. Coumans, T. Zhang, T. W. Lee, J. Tan, and S. Levine, "Learning Agile Robotic Locomotion Skills by Imitating Animals," in *Proc. Robot.: Sci. Syst. (RSS)*, 2020.

[39] R. Jeong, Y. Aytar, D. Khosid, Y. Zhou, J. Kay, T. Lampe, K. Bousmalis, and F. Nori, "Self-Supervised Sim-to-Real Adaptation for Visual Robotic Manipulation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2020.

[40] Y. Lee, E. S. Hu, and J. J. Lim, "IKEA Furniture Assembly Environment for Long-Horizon Complex Manipulation Tasks," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2021.

[41] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making Sense of Vision and Touch: Self-Supervised Learning of Multimodal Representations for Contact-Rich Tasks," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2019.

[42] T. Xiao et al., "Masked Visual Pre-training for Motor Control," *arXiv:2203.06173*, 2022.

[43] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021.

[44] P. Sermanet et al., "Time-Contrastive Networks: Self-Supervised Learning from Video," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2018.

[45] A. Zeng et al., "Transporter Networks: Rearranging the Visual World for Robotic Manipulation," in *Proc. Conf. Robot Learning (CoRL)*, 2020.

[46] S. Dasari et al., "RoboNet: Large-Scale Multi-Robot Learning," in *Proc. Conf. Robot Learning (CoRL)*, 2019.

[47] A. Brohan et al., "RT-1: Robotics Transformer for Real-World Control at Scale," in *Proc. Robot.: Sci. Syst. (RSS)*, 2023.

[48] A. Jaegle et al., "Perceiver IO: A General Architecture for Structured Inputs and Outputs," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[49] M. Shridhar et al., "Perceiver-Actor: A Multi-Task Transformer for Robotic Manipulation," in *Proc. Conf. Robot Learning (CoRL)*, 2022.

[50] Y. Zhu et al., "robosuite: A Modular Simulation Framework and Benchmark for Robot Learning," *arXiv:2009.12293*, 2020.

[51] K. Pertsch, Y. Lee, and J. J. Lim, "Accelerating Reinforcement Learning with Learned Skill Priors," in *Proc. Conf. Robot Learning (CoRL)*, 2020.

[52] A. Sharma, S. Gu, S. Levine, V. Kumar, and K. Hausman, "Dynamics-Aware Unsupervised Discovery of Skills," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.

[53] H. Ha, P. Florence, and S. Song, "Scaling Up and Distilling Down: Language-Guided Robot Skill Acquisition," in *Proc. Conf. Robot Learning (CoRL)*, 2023.

[54] C. Lynch et al., "Learning Latent Plans from Play," in *Proc. Conf. Robot Learning (CoRL)*, 2019.

[55] C. Finn and S. Levine, "Deep Visual Foresight for Planning Robot Motion," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2017.

[56] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2117–2125, 2017.