Visual-Geometry Diffusion Policy
Robust Generalization via Complementarity-Aware Multimodal Fusion

Yikai Tang[1*], Haoran Geng[1*†], Sheng Zang[2], Pieter Abbeel[1†], Jitendra Malik[1†]
[1]University of California, Berkeley, [2]Nanyang Technological University
* Equal Contribution, † Equal Advising

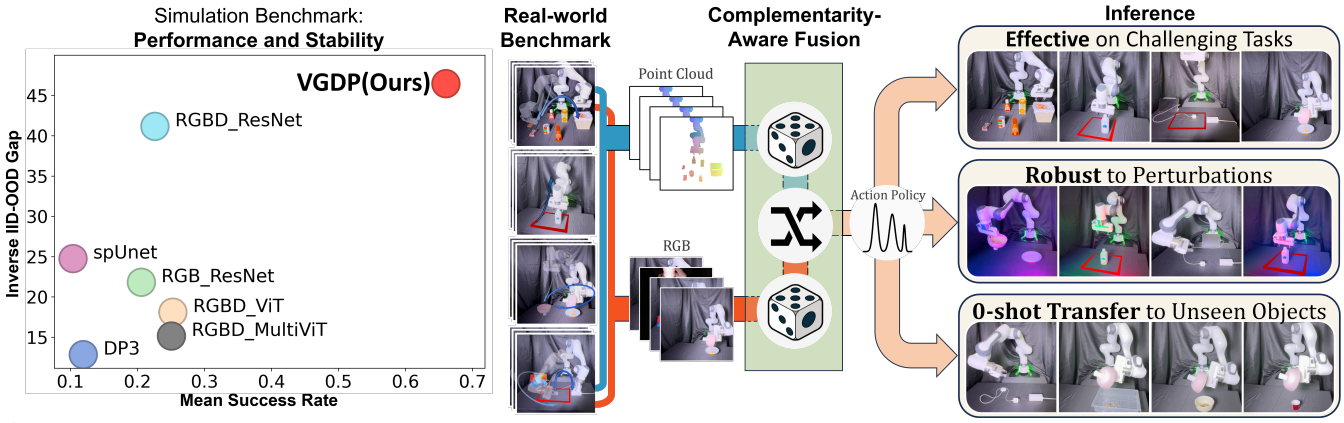


Fig. 1: **Visual-Geometry Diffusion Policy (VGDP)** is an imitation learning method that fuses 3D observations with 2D images through a *Complementarity-Aware Fusion Module*, which uses modality-wise dropout to enforce balanced use of RGB and geometry. This design yields substantial improvements in average performance, generalization, and robustness. VGDP is extensively evaluated in both simulation and the real world, covering a wide range of tasks and both visual and spatial randomizations.

***Abstract*— Imitation learning has emerged as a crucial approach for acquiring visuomotor skills from demonstrations, where designing effective observation encoders is essential for policy generalization. However, existing methods often struggle to generalize under spatial and visual randomizations, instead tending to overfit. To address this challenge, we propose Visual-Geometry Diffusion Policy (VGDP), a multimodal imitation learning framework built around a *Complementarity-Aware Fusion Module* where modality-wise dropout enforces balanced use of RGB and point-cloud cues, with cross-attention serving only as a lightweight interaction layer. Our experiments show that the expressiveness of the fused latent space is largely induced by the enforced complementarity from modality-wise dropout, with cross-attention serving primarily as a lightweight interaction mechanism rather than the main source of robustness. Across a benchmark of 18 simulated tasks and 4 real-world tasks, VGDP outperforms seven baseline policies with an average performance improvement of 39.1%. More importantly, VGDP demonstrates strong robustness under visual and spatial perturbations, surpassing baselines with an average improvement of 41.5% in different visual conditions and 15.2% in different spatial settings.**

I. INTRODUCTION

Imitation learning has become a key paradigm in robotics, enabling robots to efficiently acquire complex manipulation skills directly from expert demonstrations without expensive online learning [1] [2] [3] [4] [5] [6]. Recent advances in imitation learning have extended its applicability from simple manipulation primitives to diverse, long-horizon visuomotor tasks, supported by more robust visual representations and scalable policy architectures. [7], [8] Despite this progress, real-world deployment inevitably exposes robots to diverse sensor noise and continual distribution shifts—spanning appearance, illumination, viewpoint, and workspace layout—under which robust generalization remains elusive. Image-based encoders are notoriously brittle to changes in appearance, illumination, and viewpoint [9]–[12]. Existing approaches either lack the capacity to accurately fit complex action distributions, or they overfit to narrow demonstration data and consequently fail to generalize. Merely scaling demonstrations is unsustainable; hundreds of real-world demos can take days to collect, while the scaling law yields diminishing accuracy gains as data grows [1], [13]. The question of how to produce data-efficient representations that are simultaneously informative and generalizable remains unresolved. Previous work relies largely on the concatenation of early or late features [14], [15] or attaching RGB images as attributes to points [16]–[18], which often yields modest or unstable gains under distribution shift.

To tackle this problem, we propose the **Visual-Geometry Diffusion Policy (VGDP)**, which integrates a *Complementarity-Aware Fusion Module* to actively prevent modality collapse and promote balanced use of RGB and geometric cues. In this layer, RGB and point-cloud features are first projected into a shared embedding space and then passed through a **modality-dropout** mechanism that randomly removes entire sensory streams during training. This dropout-driven constraint forces the policy to learn modality-invariant and complementary structure, making the fused rep-

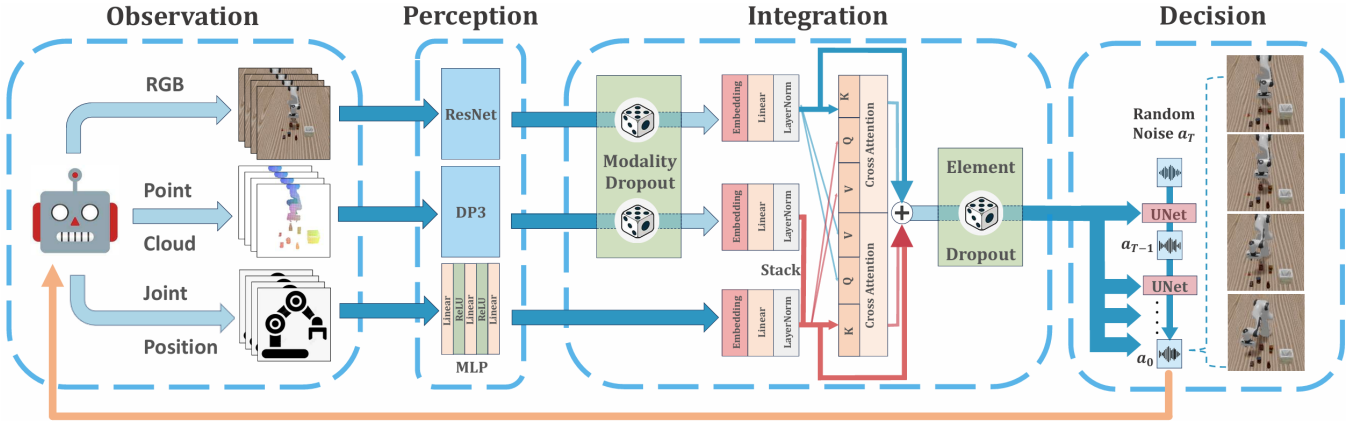


Fig. 2: **Visual-Geometry Diffusion Policy Overview.** (a) **Observation**: The environment is captured by a single-view RGB-D camera together with robot joint states. (b) **Perception**: Each modality is independently encoded by a dedicated encoder, ensuring comprehensive semantic representation. (c) **Integration**: The richly represented features learn cross-modal dependencies and contextual relationships via a cross-attention layer, wrapped with modality-wise and element-wise dropout to enforce balanced modality utilization and feature activation. (d) **Decision**: Conditioned on the fused feature, a noised action is denoised to either provide loss for end-to-end training or output an action during evaluation.

resentation far more expressive than what architecture alone can offer. A lightweight bidirectional cross-attention layer then exchanges information between the surviving modalities, with residual connections and element-wise dropout providing additional stability before the fused feature is passed to the decision head.

Finally, the action is denoised from random noise conditioned on this expressive multimodal representation. This formulation jointly exploits the semantic detail of RGB inputs and the geometric precision of point clouds, markedly improving the diffusion policy's generalization and its resilience to distributional shifts.

To systematically assess the capability of VGDP, we conduct experiments on a unified RoboVerse benchmark [19], consisting of 18 simulation tasks covering three levels of randomization and four real-world tasks, covering a wide range of generalization challenges. As demonstrated by our experimental results, VGDP exhibits the following features:

1) **Visual Generalizability.** VGDP is not only highly robust to new visual conditions when trained with randomization in materials, lights, and camera positions, but also shows a strong ability to **operate under out-of-distribution conditions** even when only trained on a fixed scene. It exceeds the other six baselines with an average increase in success rate of **41.5%** in visually randomized scenes.
2) **Spatial Generalizability.** The fused observation feature provides the VGDP with a more **in-depth understanding of the scenario**, facilitating its ability to generalize smoothly across a wide workspace, showing an average rise in the success rate of **15.2%** in spatially randomized tasks.
3) **Precise Control Ability.** VGDP significantly prevails over other methods in tasks that require precise control, highlighting the importance of including an information-rich RGB representation. For the selected tasks that require control precision no more than 1 centimeter, VGDP demonstrates a success rate increase of **18.77%**.

Conclusively, our method exhibits strong generalization across varied scenes, highlighting the central importance of the modality-dropout mechanism in shaping resilient and modality-invariant representations.

II. RELATED WORK

Imitation Learning. Imitation learning (IL) learns visuomotor control from demonstrations, avoiding costly online exploration [20]. Recent work explores generative or diffusion-based policies for robust action denoising [1], data-efficient 3D-structure-aware policies [3], and large-scale demonstrations for diverse skill acquisition [13], [21]. Yet generalization under lighting, viewpoints, or layout changes remains a central challenge, as even small visual drifts cause significant performance drops. Common remedies—domain randomization, diverse demonstrations, and visual pretraining—alleviate but cannot fully bridge the gap, as low-level approaches in data augmentation or pretraining cannot transcend the representational limits of high-level observation encoders. To address this, our method provides a stronger visual representation that fully unleashes the potential of imitation learning.

Multimodal Visual Encoders. Nowadays, fusing RGB images with point clouds has become a mainstream approach for robotic perception and manipulation. Multimodal visual encoders combine RGB's rich semantics with 3D geometry's metric structure [22]. Progress ranges from early point-wise fusion such as PointFusion [14], to sequential semantic painting exemplified by PointPainting [16], and to recent robot imitation learning methods like FPV-Net [18] that condition point-cloud encoders on visual cues for robust

policy learning. Despite these advances, challenges such as modality dominance [23], [24], cross-sensor misalignment [16], and supervision imbalance persist [25], motivating our method, which introduces a geometry-aware shared feature space and leverages attention mechanisms to achieve more balanced multimodal integration.

Visual Representation in Robot Learning. Across IL and reinforcement learning, robots consume RGB, RGB-D, and point clouds, with different trade-offs between semantic detail and geometric invariance. Recent systems employ cross-modal attention or structured latent spaces to integrate these cues for policy learning [16]. Yet principled 2D–3D fusion remains uncommon: many methods still rely on a single modality or naïve concatenation that underutilizes complementary strengths. We therefore propose to explicitly align modalities and regularize their balanced use, enabling stronger representations and better out-of-distribution generalization.

III. METHOD

VGDP outputs actions a from random noise $\mathbf{a}_T \sim \mathcal{N}(0, I)$ conditioned on observation c. We learn such a denoising process with a denoising network D, as illustrated in Equation (1).

$$\mathbf{a} = \mathrm{D}_\theta(\mathbf{a}_T, \mathbf{c}) \tag{1}$$

Our goal is to learn an informative and generalizable observation representation c, and we believe the key lies in fusing the naturally complementary visual and 3D observation space. Following such a design, we build VGDP as demonstrated in Figure 2 with the following components: (a) **Perception**, (b) **Integration**, and (c) **Decision**.

A. *Perception*

The perception module consists of three independent encoders that respectively process the image, the geometry, and the low-dimensional features. For simplicity, we use a single-view RGB-D camera for both visual encoders.

- **Image Encoder**: High-dimensional RGB observations can be encoded by a variety of backbones. For data efficiency and real-time latency, we adopt a *ResNet-18* [26] image encoder. Concretely, a 256×256 RGB frame is processed by ResNet-18 followed by global average pooling to produce a 512-dimensional global feature used in downstream fusion.
- **3D Encoder**: 3D scenes can be parameterized in many ways—RGB-D [27], point clouds [28] [29], voxels [30], implicit fields [31], even 3D Gaussians [32]. We adopt a point-cloud, geometry-first representation for its robustness and ease of generalization. From a single RGB-D frame, depth is back-projected to a metric cloud, then cropped to the task workspace. To balance coverage and compute, Farthest Point Sampling (FPS) is applied for down-sampling the point cloud to 4096 points. The DP3 encoder [1], a light-weight yet effective adaptation of PointNet, distills the down-sampled points into a compact 64-dimensional global feature.
- **Low-dim Encoder**: For encoding robot state, we use a light-weight MLP network as shown in Figure 2 to map the joint positions to a higher-dimensional space.

B. *Integration*

Our integration module implements a *Complementarity-Aware Fusion Layer* that combines RGB images, point clouds, and low-dimensional robot states into a unified policy representation. Rather than relying on architecture alone, the layer is designed to prevent modality collapse and encourage balanced use of complementary cues.

At its core is a **modality-dropout** operation: with probability p=0.2, the RGB or 3D branch is randomly disabled before fusion. This structured perturbation serves as the primary regularizer of VGDP, forcing the policy to develop modality-invariant and complementary structure and making the fused representation far more expressive and robust to missing or corrupted observations [33], [34].

After dropout, surviving features are projected into a shared 256-dimensional space, where a lightweight bidirectional cross-attention layer performs minimal interaction between modalities. Residual connections preserve unimodal competence, and a final element-wise dropout further discourages co-adaptation, producing a stable and well-balanced latent representation that is passed to the decision head.

C. *Decision*

VGDP adopts a diffusion-policy [1] head for action selection. Starting from noised action a_T, the policy iteratively denoises an action sequence conditioned on the fused context $\mathbf{c}$. Here, $\mathbf{c}$ denotes the *shared latent representation* produced by the integration module, aggregating RGB images, point-clouds, and low-dimensional state features.

Training. The action is noised for a random step T, as shown in Equation (2):

$$a_t = \sqrt{1-\beta_t}\, a_{t-1} + \sqrt{\beta_t}\, \epsilon_t \tag{2}$$

where β_t is a hyperparameter, and ϵ_t is Gaussian Noise. In the diffusion step t, the network predicts the noise as

$$\hat{\boldsymbol{\epsilon}} = \epsilon_\theta(\mathbf{a}_t,\, t \,|\, \mathbf{c})\,, \tag{3}$$

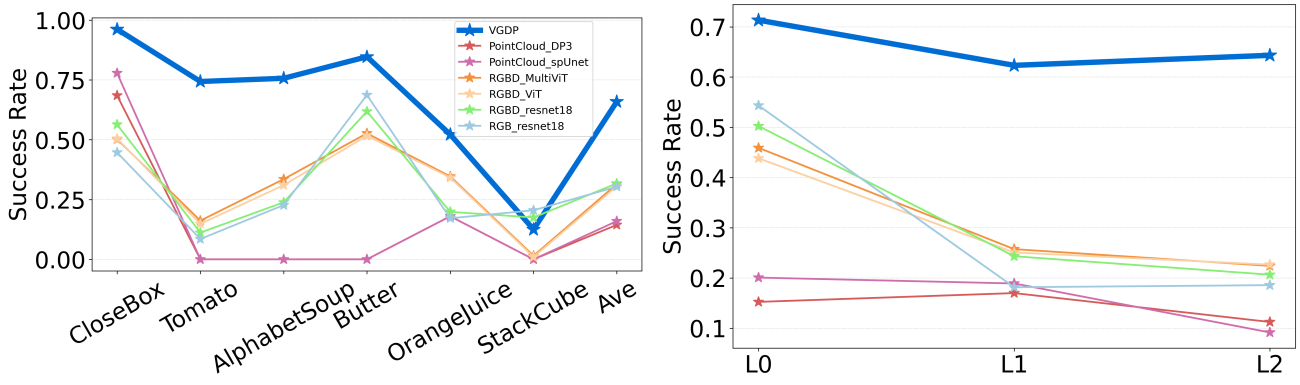
and is trained with the standard L2 objective;

$$\mathcal{L}_{\text{diff}} = \mathbb{E}\Big[\big\|\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}\big\|_2^2\Big]. \tag{4}$$

Inference. At inference, we initialize $\mathbf{a}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and repeatedly condition on $\mathbf{c}$ to denoise toward an executable action sequence.

IV. SIMULATION EXPERIMENTS

We conduct comprehensive simulation benchmarking and ablations across 18 simulated tasks comparing 7 observation encoders.



(a) Performance across tasks

(b) Performance across randomization levels

Fig. 3: **Comparative Analysis of Encoder Performance.** Figure 3a shows how different encoders perform across tasks, with each task score averaged over the three randomization levels. Figure 3b depicts their performance across randomization levels, where each level score is averaged over all tasks.

Relative Dispersion	**VGDP**	RGB_resnet18	RGBD_resnet18	RGBD_ViT	RGBD_MultiViT	PointCloud_DP3	PointCloud_spUnet
Cross-task ↓	**30%**	52.80%	48.48%	45.49%	43.86%	126.19%	133.33%
Cross-randomization-level ↓	**2.53%**	32.72%	35.01%	45.28%	37.23%	5.06%	15.43%

TABLE I: **Cross-task and cross-randomization stability comparison.** We evaluate the stability with relative dispersion of success rates across-tasks or randomization levels. Lower relative fluctuation means more stable performance across different tasks or observation distributions.

Task	**VGDP**	RGB_resnet18	RGBD_resnet18	RGBD_ViT	RGBD_MultiViT	PointCloud_DP3	PointCloud_spUnet
CloseBox	**0.96**	0.35	0.53	0.48	0.49	0.58	0.75
StackCube	0.03	**0.1**	0.06	0	0.01	0	0
AlphabetSoup	**0.79**	0.22	0.24	0.61	0.48	0	0
Butter	**0.88**	0.72	0.62	0.68	0.69	0	0
OrangeJuice	**0.53**	0.17	0.18	0.31	0.33	0.07	0.08
Tomato	**0.72**	0.08	0.11	0.13	0.14	0	0
Average	**0.65**	0.27	0.29	0.32	0.33	0.11	0.14

TABLE II: **Cross-task performance comparison of simulation benchmark.** VGDP prevails over all others on almost all tasks, with an average absolute improvement of **40.11%** and a smaller variance. VGDP also achieves high success rate(72%) for task that all others nearly fail(averaged performance $\leq$15%).

Level	**VGDP**	RGB_resnet18	RGBD_resnet18	RGBD_ViT	RGBD_MultiViT	PointCloud_DP3	PointCloud_spUnet
L0	**0.68**	0.46	0.43	0.44	0.48	0.12	0.19
L1	**0.64**	0.17	0.23	0.36	0.32	0.13	0.15
L2	**0.64**	0.19	0.21	0.20	0.20	0.08	0.08

TABLE III: **Cross-randomization-level performance comparison of simulation benchmark.** Each entry reports the average success rate of an encoder at a given randomization level, averaged over all tasks. VGDP exhibits near-invariant performance while maintaining high performance, virtually unaffected by scene randomization; meanwhile, performance of other encoders is either low across all randomization levels, or drops rapidly when randomization enhances.

A. Benchmark Design and Setup

Baseline Models. We compare VGDP against six baselines that span 2D, RGBD and 3D representations and various network families (see Table II), including ResNet18-RGB, ResNet18-RGB-D [35], ViT-RGBD, MultiViT-RGBD [36], DP3 [3] and SparseUNet [37]. For fair comparison, all methods share the same diffusion-policy head for action generation.

Simulation Benchmark. We evaluated models in simulation on 18 tasks drawn from RoboVerse. These tasks span multiple mainstream benchmarks—such as Libero-objects [38], Maniskill [39]–[41] and RLBench [42]—and are executed on diverse simulators with varying degrees of visual randomization. This setup ensures that the benchmarking results are not biased toward any specific task set or simulator.

- **Benchmark Platform and Setup.** Simulation studies have long been hampered by the sim-to-real (Sim2Real) gap. To mitigate this, we initially build our benchmark on RoboVerse [19], a unified platform that integrates widely used simulators with a consistent API and configuration, allowing apples-to-apples comparisons in diverse tasks and engines. Each task is trained with 100 demos and 200 trials evaluated, equally covering in-distribution and OOD scenes. RoboVerse supports

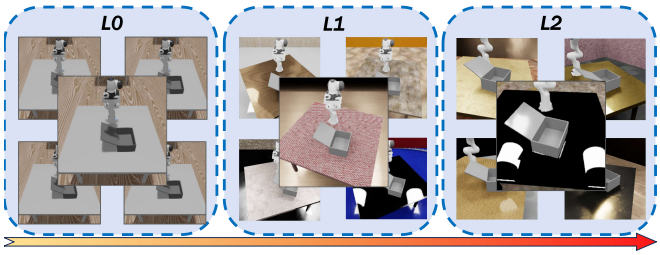


Fig. 4: **Environments under Different Randomization Levels.** The three brackets, named L0, L1 and L2, demonstrate example randomization in materials, positions and viewpoints of the same task across different randomization levels.

Level	materials	table position	camera pos
L0	✗	✗	✗
L1	✓	✓	✗
L2	✓	✓	✓

TABLE IV: **Randomization factors at different levels.** Level 1 randomizes wall, table, and object materials and rotates the table arbitrarily; Level 2 further randomizes camera positions from a predefined pool.

controlled randomization of lighting, textures/materials, specular reflectances, and camera poses; for each task, we define three randomization levels (see Table IV and Figure 4).

- **Task Suite.** Our simulation benchmark includes tasks originating from two mainstream benchmarks: Libero, RLBench and Maniskill, which include diverse tasks that require precise control, articulated-object manipulation, and deformable-object manipulation. The variance of tasks combined with randomization of scenarios makes our benchmarking both challenging and comprehensive, enabling a more faithful evaluation of policy generalization against visual and physical variations.

B. *Effective Modeling of Visual Changes*

VGDP attains the highest success rates in all three randomization levels, outperforming unimodal baselines by an average of **39.1%**, as demonstrated in Table III. It reaches **72.3%** success even on tasks where the two constituent encoders—ResNet and DP3—fail completely when trained individually ($\leq$**1%**), indicating a geometry-aware understanding that links dense RGB cues with viewpoint-invariant 3D structure. Equally important, its success rate varies only by **2.53%** between randomization levels (Figure 3b)—an stability gain $\approx 11\times$ over competitors—showing that coupling fine-grained textures with 3D geometry yields both strong **absolute performance** and **exceptional robustness.**

C. *Stable Performance across Tasks*

In addition to its robustness to visual randomization, VGDP consistently demonstrates strong performance in a wide spectrum of tasks, as detailed in Table II and Figure 3a. Across various tasks, VGDP reduces the relative dispersion of performance between tasks by **47.3%** relative to the best unimodal baseline, maintaining a high performance floor and a narrow variance, as demonstrated by Table I. RGB/RGBD policies remain sensitive to view-point noise, while 3D encoders under-sample small targets and over-smooth local details. VGDP's soft cross-modal routing selects the most reliable cue for consistent competence across heterogeneous tasks.

D. *Exceptional Generalization under Out-of-Distribution scenarios*

Apart from effectively and stably modeling widely-distributed training data, it is also crucial whether a policy can maintain such performance when evaluating on out-of-distribution scenes. Under the strongest Level-2 randomization, we estimate such ability from two perspectives: the averaged IID and OOD success rate, and the performance gap between IID and OOD scenes. VGDP delivers a **55.3%** higher mean success rate and a **59.5%** smaller IID–OOD gap than the 6 baselines, as demonstrated by Figure 1. While baselines are either low in success rate or poor at generalizing IID knowledge to OOD scenes, VGDP excels in both axes: it captures transferable representations that support robust generalization, underscoring the distinct strength of VGDP.

E. *Ablation Study*

To isolate where VGDP's improvements come from, we evaluate the key components of our fusion layer on CloseBox across all randomization levels. The results reveal a consistent pattern: **cross-attention provides a useful mechanism for exchanging information, but the major performance gains arise only when fusion is paired with complementarity-aware dropout**. Without dropout, multimodal models tend to collapse to whichever unimodal signal is easier to fit, failing to realize true multimodal benefits.

	CBL0		CBL1		CBL2	
	IID	OOD	IID	OOD	IID	OOD
VGDP	**1.00**	**0.97**	**0.94**	**0.95**	**0.95**	**0.95**
w/o Residual	0.76	0.82	0.52	0.52	0.94	0.86
w/o Dropout	0.86	0.80	0.70	0.67	**0.95**	0.93
attention→concat	0.80	0.76	0.87	0.86	0.69	0.66
attention→early fusion	0.00	0.00	0.00	0.00	0.46	0.40
DP (RGB-only)	0.81	0.52	0.40	0.20	0.42	0.32
DP3 (PC-only)	0.82	0.60	0.73	0.77	0.82	0.38

TABLE V: **Ablation study on CloseBox at three randomization levels.** DP = RGB-only Diffusion Policy; DP3 = point-cloud baseline.

1) ***Complementarity-Aware Dropout****: Enforcing Balanced Fusion:* Modality-wise dropout is the critical factor that prevents modality dominance. Removing dropout causes the fused model to closely track the stronger unimodal encoder under each randomization level (Table V).

At Level 0, RGB-only DP performs best (0.81/0.52), while DP3 lags, and the fusion model without dropout mirrors this behavior (0.86/0.80). At Level 1, DP3 becomes dominant

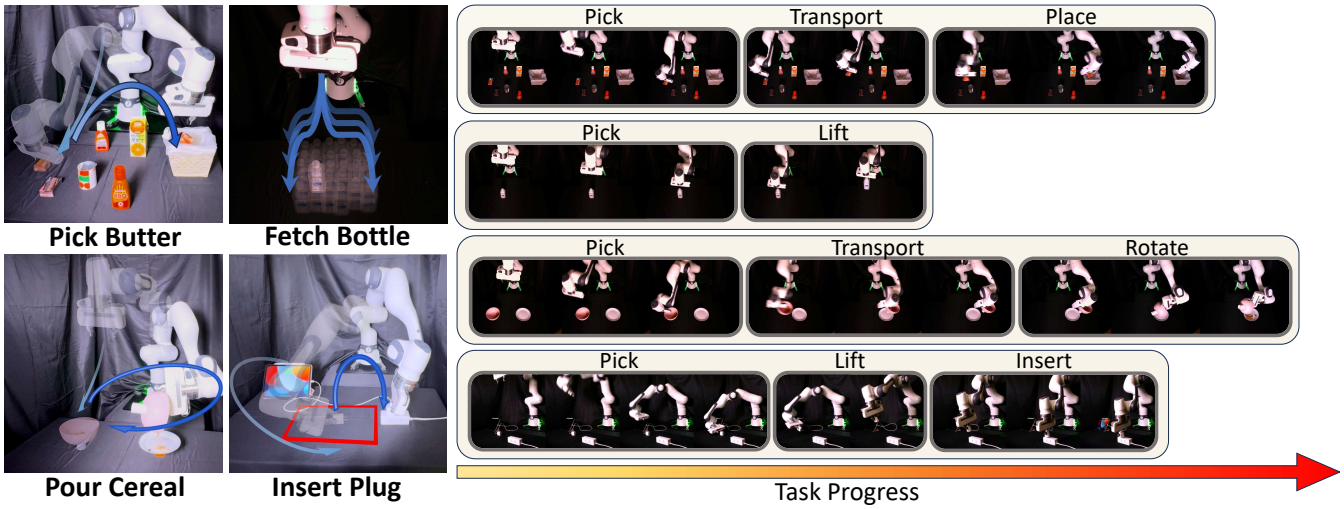


Fig. 5: **Real-world Benchmarks.** We deploy VGDP in the real world with a Franka Arm across four challenging tasks, including clutter scene manipulation (PickButter), 6-DoF cereal pouring (PourCereal), fine-grained handling over large spatial generalization (FetchBottle) and precision-driven, force-aware insertion (InsertPlug).

(0.73/0.77), and the fused model again shifts toward the point-cloud branch (0.70/0.67). At Level 2, the trend persists.

These trends show that **without dropout, cross-attention alone cannot stop the fusion module from collapsing onto the dominant modality**. Dropout counteracts this by forcing the model to rely on complementary cues, leading to a fused latent space that is significantly more expressive and substantially more robust to OOD conditions.

*2) **Fusion Architecture**: Cross-Attention vs. Concat vs. Early Fusion:* We further compare attention-based fusion with two widely used alternatives—feature concatenation and early fusion—under identical RGB and point-cloud encoders.

- **Early fusion** collapses entirely at Levels 0 and 1 (0.00), recovering slightly only under heavy viewpoint variation. This pattern mirrors DP3, indicating that the RGB branch is effectively unused.
- **Concatenation** performs consistently between DP and DP3, suggesting that passive late fusion struggles to exploit complementary structure.
- **Cross-attention** achieves the strongest fusion among the three: it improves adaptability to viewpoint shifts and allows bidirectional interaction between modalities.

However—and this is the key insight—**cross-attention delivers its full benefit only when combined with modality dropout**. With dropout, cross-attention enables rich cross-modal alignment; without it, even attention-based fusion collapses to a single dominant signal.

In summary, **cross-attention is useful, but the decisive boost in fused representation quality comes from the complementarity constraint imposed by dropout**.

V. REAL-WORLD EXPERIMENTS

A. *Benchmark Setup*

We evaluated VGDP on four real-world manipulation tasks to probe both skill competence and adaptation under various conditions. Expert demonstrations are collected via human teleoperation with Gello [43] on a Franka Emika Panda arm with a Franka gripper. All of the policies are trained for 3000 episodes on 100 demonstrations and evaluated on 25 trials, if not otherwise stated. Because real-world rollouts are time-consuming, we only benchmark one representative encoder for each of the four modalities: ResNet18 for RGB, ResNet18-RGBD for RGBD [35], DP3 for point cloud [3], and VGDP for multimodal encoders. In Figure 5, we present a brief description of our real-world tasks.

- **LiberoPickButter**: 7 everyday objects of diverse sizes, shapes and complex colors are placed on the desk, and the policy should learn to pick the butter out from the complex scene and place it into the basket. The setting follows the same task in our simulation benchmark strictly to quantitatively assess the Sim2Real gap.
- **FetchBottle**: A bottle is placed within a 39cm*33cm rectangle that covers most of the robots' operating area, and the policy is required to pick the bottle up. This task features precise control, as the cap of the bottle is only of 2 centimeters in diameter. The training data are collected with the bottle placed in 30 grid locations, while evaluation is performed on all 143 grids, as demonstrated in Figure 6.
- **PourCereal**: A bowl of cereal and a plate is placed on the desk, and the policy should learn to pick up the bowl and pour the cereal into the plate. The challenge lies in the continuous control of complex 3D orientation, where small angular errors can easily lead to spillage.
- **InsertPlug**: A tablet charging plug rests on the table within a 20 cm × 20 cm area, with a power strip on the opposite side. The policy must grasp the plug, lift it, move it above the power strip, and insert it into a socket so that the tablet turns on charging.

Task	VGDP	DP(RGB)	DP(RGBD)	DP3
Pick Butter	**100%**	84%	80%	12%
Fetch Bottle	**22.38%**	18%	2.10%	1.40%
Pour Cereal	**100%**	88%	84%	80.00%
Insert Plug	**36%**	25%	16%	0%

TABLE VI: **Real-world policy performance across four tasks.** Highest score per task is highlighted.

B. *Spatial Generalization*

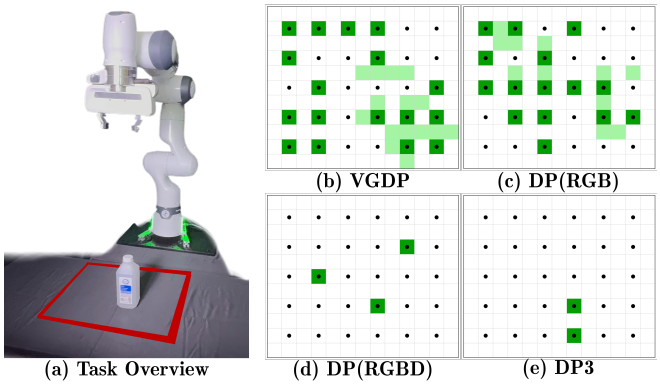
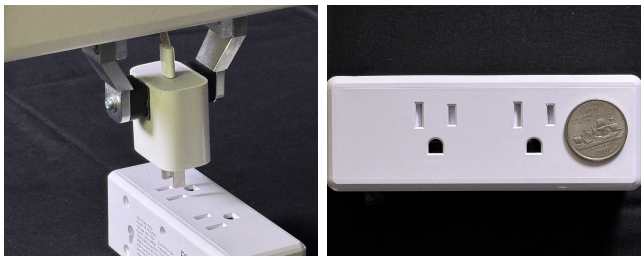


Fig. 6: **Evaluation Results for Real-world Randomization.** The grid map represents a bird's-eye-view of the workspace, where each grid corresponds to a bottle placement during evaluation. Policies are trained on expert demonstrations on the 30 dotted grids and evaluated on all 143 positions. Dark-green, light-green, and white represent IID success, OOD success, and failure, respectively.

Practical imitation learning requires transferring skills to novel spatial locations. We test this ability on a precise FetchBottle task, where the 2 centimeters bottle cap demands fine manipulation.

As demonstrated in Figure 6, VGDP and RGB encoders generalize well, whereas point-cloud and RGBD methods fail dramatically, often regressing toward familiar training positions. Such poor performance likely stems from the noisy nature of real-world depth images, which hampers 3D-based models from learning robust and broadly generalizable representations. This highlights the indispensable and robust role of VGDP as a visual representation for real-world deployment, where stable and noise-resilient features are crucial for reliable robotic perception and control.

C. *Fine-Detail Visual Cue Perception*



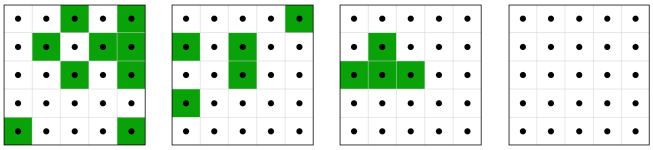
(a) Close-up view of the insertion process.

(b) Socket shown with a quarter for scale.

Fig. 7: **Key visual challenges in the InsertPlug task.** (a) The robot must insert the plug in a strictly vertical direction into a narrow socket on every attempt, despite perturbations from varying grasp poses and the plug's wobbling during the forceful insertion process. (b) The size of the socket is compared with a quarter to provide intuitive scale.

InsertPlug reveals a qualitatively different challenge from our other tasks. As demonstrated in Figure 7, successful insertion requires the robot to drive the plug downward along an almost perfectly vertical axis, with enough force to overcome socket friction. Under this physical regime, even slight lateral deviation produces large torque that immediately triggers safety stops. Because the gripper never grasps the plug in a consistent pose, every trial begins with a different plug–socket orientation—making any memorized trajectory fundamentally unstable.

We evaluate the policy on 25 evenly distributed test points across the workspace. As demonstrated in Figure 8, VGDP achieves the most consistent success across all test locations, outperforming RGB and RGBD and dramatically surpassing DP3. Surprisingly, DP3 often fails at the grasping stage itself. In contrast, VGDP and RGB/RGBD policies at least reach the alignment phase, with VGDP showing the highest robustness across varied grasp poses.



(a) VGDP (b) DP(RGB) (c) DP(RGBD) (d) DP3

Fig. 8: **Evaluation Results for Insert Plug.** The grid map represents a bird's-eye-view of the workspace, where each grid represents a position where the plug is placed during evaluation; policies are trained on expert demonstrations on all dotted grids and evaluated on selected 25 positions that spans evenly across the workspace; grids colored in dark-green and white represent success and failure, respectively.

D. *Zero-Shot Transfer under Cross-Object, Cross-Device, and Cross-Scene Shifts*

To evaluate generalization beyond the training distribution [44], we measure VGDP's robustness under six challenging zero-shot transfer conditions that span semantic variation, geometric variation, and severe appearance changes. These transfers cover three real-world tasks:

- **PourCereal Object Transfer**: replacing the target plate with (1) a large semi-transparent drawer, (2) a medium bowl, (3) a small yogurt cup.
- **InsertPlug Device Transfer**: (4) relocating the iPad to a novel position, (5) replacing the iPad entirely with an AirPods charger.
- **FetchBottle Visual–Spatial Transfer**: (6) testing under strong color-cycling illumination and unseen placement positions.

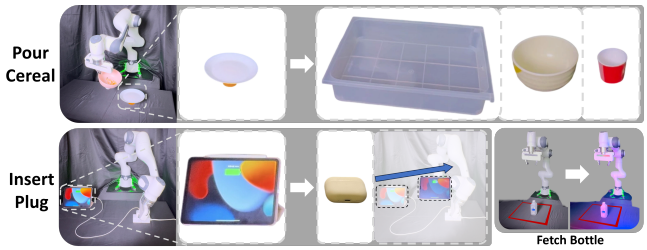
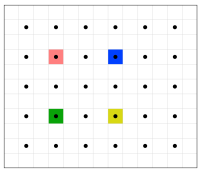


Fig. 9: **Zero-shot transfer settings.** Illustration of the transfer conditions used in our real-world evaluation. (Up): In Pour Cereal, we replace the plate with containers of various size and shape including: a drawer, a bowl and a yogurt cup. (Down Left): In Insert Plug, we either replace the IPad with an airpods case or move the IPad to a new position. (Down Right) In Fetch Bottle, we replace the original lighting with a strong color-cycling illumination.

Results. VGDP achieves a **perfect 8/9 zero-shot success rate**, while unimodal baselines fail inconsistently across the six settings. Table VII and Figure 10 summarizes the results.

	Drawer	Bowl	Yogurt	Position	AirPods	Light Strip
VGDP	✓	✓	✓	✓	✓	✓
DP(RGB)	✓	✓	✗	✓	✗	✗
DP(RGBD)	✓	✗	✗	✗	✓	✗
DP3	✓	✓	✗	✗	✗	✓

TABLE VII: **Zero-shot transfer across unseen objects, devices, and visual conditions.** VGDP succeeds on all six transfer scenarios, while unimodal baselines exhibit modality-specific weaknesses.



VGDP	✗	✓	✓	✓
DP(RGB)	✗	✗	✗	✗
DP(RGBD)	✗	✗	✗	✗
DP3	✗	✗	✗	✓

Fig. 10: **Spatial zero-shot transfer in FetchBottle.** VGDP succeeds on three of four OOD positions, while unimodal baselines fail consistently.

Unified Analysis. Together, these six transfer settings span a broad spectrum of real-world distribution shifts: (1) *semantic* (new containers, new devices), (2) *geometric* (different receptacle shapes and insertion hardware), (3) *visual* (harsh RGB lighting changes), and (4) *spatial* (unseen placements and workspace offsets).

The results in Table VII show clear differences across policies. RGB and RGBD policies succeed on the Drawer and Bowl settings but fail consistently on Yogurt, AirPods, and Light-Strip illumination. Their performance also decreases under unseen placement shifts. DP3 handles Drawer and Bowl reliably and succeeds under Light-Strip variation, yet fails on Yogurt, AirPods, and unseen-position transfer. VGDP, in contrast, succeeds in all six transfer settings, including the challenging Yogurt container, the reflective AirPods charger, and the strong color-cycling illumination.

These results indicate that VGDP is the only policy that maintains reliable zero-shot transfer across all types of distribution shift.

VI. CONCLUSIONS

We present VGDP, a multimodal policy centered on a *Complementarity-Aware Fusion Module* that enforces balanced and redundant use of RGB and 3D geometry. It dynamically integrates information from distinct modalities through a bidirectional cross-attention layer, while encouraging balanced representations across modalities via lightweight yet effective dropout and encoding. Across 18 simulated and four real-world tasks featuring appearance shifts, spatial disturbances, and fine-grained control, VGDP outperforms strong 2D/3D/RGB-D baselines by an average of **39.1%**, and attains **72.3%** success on tasks where ResNet-only and DP3-only policies fail ($\leq$1%). Ablations indicate that gains come from preserving rich RGB cues, enforcing balanced modality use, and adopting cross-attentive fusion over static combinations.

VII. LIMITATIONS & FUTURE WORK

We have benchmarked VGDP mainly on tasks from RLBench and Libero-objects, deploying only on Franka Arm with Franka Gripper, using RoboVerse as the benchmarking platform. The performance of VGDP on higher-dimensional robot state space (e.g. Dexterous hands), on wider range of benchmarks (e.g. MetaWorld [45]) and on decision backbones beyond diffusion policy (e.g. Action Chunking **??**) is yet to be discovered. In future work, we will systematically explore the performance boundaries of VGDP by evaluating it across more diverse robotic platforms, task families, and control paradigms.

ACKNOWLEDGMENT

We would like to thank Yiran Wang, Caiyi Zhang, and Feishi Wang for their helpful discussions and valuable feedback throughout this project. Pieter Abbeel holds concurrent appointments as a professor at the University of California, Berkeley and as an Amazon Scholar. This paper describes work performed at UC Berkeley and is not associated with Amazon.

REFERENCES

[1] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," 2024. [Online]. Available: https://arxiv.org/abs/2303.04137

[2] A. Posadas-Nava, A. Scorsoglio, L. Ghilardi, R. Furfaro, and R. Linares, "Action chunking with transformers for image-based spacecraft guidance and control," 2025. [Online]. Available: https://arxiv.org/abs/2509.04628

[3] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations," 2024. [Online]. Available: https://arxiv.org/abs/2403.03954

[4] K. Wu, Y. Zhu, J. Li, J. Wen, N. Liu, Z. Xu, and J. Tang, "Discrete policy: Learning disentangled action space for multi-task robotic manipulation," 2025. [Online]. Available: https://arxiv.org/abs/2409.18707

[5] M. Shridhar, L. Manuelli, and D. Fox, "Perceiver-actor: A multi-task transformer for robotic manipulation," 2022. [Online]. Available: https://arxiv.org/abs/2209.05451

[6] Z. Xu, Z. He, J. Wu, and S. Song, "Learning 3d dynamic scene representations for robot manipulation," 2020. [Online]. Available: https://arxiv.org/abs/2011.01968

[7] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, T. Jackson, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, K.-H. Lee, S. Levine, Y. Lu, U. Malla, D. Manjunath, I. Mordatch, O. Nachum, C. Parada, J. Peralta, E. Perez, K. Pertsch, J. Quiambao, K. Rao, M. Ryoo, G. Salazar, P. Sanketi, K. Sayed, J. Singh, S. Sontakke, A. Stone, C. Tan, H. Tran, V. Vanhoucke, S. Vega, Q. Vuong, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, "Rt-1: Robotics transformer for real-world control at scale," 2023. [Online]. Available: https://arxiv.org/abs/2212.06817

[8] Q. Qian, G. Zhao, G. Zhang, J. Wang, R. Xu, J. Gao, and D. Zhao, "Gp3: A 3d geometry-aware policy with multi-view images for robotic manipulation," 2025. [Online]. Available: https://arxiv.org/abs/2509.15733

[9] K. Hu, Z. Rui, Y. He, Y. Liu, P. Hua, and H. Xu, "Stem-ob: Generalizable visual imitation learning with stem-like convergent observation through diffusion inversion," 2024. [Online]. Available: https://arxiv.org/abs/2411.04919

[10] C. Wang, H. Fang, H.-S. Fang, and C. Lu, "Rise: 3d perception makes real-world robot imitation simple and effective," 2024. [Online]. Available: https://arxiv.org/abs/2404.12281

[11] S. Batra and G. Sukhatme, "Zero-shot visual generalization in robot manipulation," 2025. [Online]. Available: https://arxiv.org/abs/2505.11719

[12] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," 2021. [Online]. Available: https://arxiv.org/abs/2010.04159

[13] H.-S. Fang, H. Fang, Z. Tang, J. Liu, C. Wang, J. Wang, H. Zhu, and C. Lu, "Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot," 2023. [Online]. Available: https://arxiv.org/abs/2307.00595

[14] D. Xu, D. Anguelov, and A. Jain, "Pointfusion: Deep sensor fusion for 3d bounding box estimation," 2018. [Online]. Available: https://arxiv.org/abs/1711.10871

[15] V. A. Sindagi, Y. Zhou, and O. Tuzel, "Mvx-net: Multimodal voxelnet for 3d object detection," 2019. [Online]. Available: https://arxiv.org/abs/1904.01649

[16] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Pointpainting: Sequential fusion for 3d object detection," 2020. [Online]. Available: https://arxiv.org/abs/1911.10150

[17] R. Li, X. Li, P.-A. Heng, and C.-W. Fu, "Pointaugment: an auto-augmentation framework for point cloud classification," 2020. [Online]. Available: https://arxiv.org/abs/2002.10876

[18] A. Donat, X. Jia, X. Huang, A. Taranovic, D. Blessing, G. Li, H. Zhou, H. Zhang, R. Lioutikov, and G. Neumann, "Towards fusing point cloud and visual representations for imitation learning," 2025. [Online]. Available: https://arxiv.org/abs/2502.12320

[19] H. Geng, F. Wang, S. Wei, Y. Li, B. Wang, B. An, C. T. Cheng, H. Lou, P. Li, Y.-J. Wang, Y. Liang, D. Goetting, C. Xu, H. Chen, Y. Qian, Y. Geng, J. Mao, W. Wan, M. Zhang, J. Lyu, S. Zhao, J. Zhang, J. Zhang, C. Zhao, H. Lu, Y. Ding, R. Gong, Y. Wang, Y. Kuang, R. Wu, B. Jia, C. Sferrazza, H. Dong, S. Huang, Y. Wang, J. Malik, and P. Abbeel, "Roboverse: Towards a unified platform, dataset and benchmark for scalable and generalizable robot learning," 2025. [Online]. Available: https://arxiv.org/abs/2504.18904

[20] C. Celemin, R. Pérez-Dattari, E. Chisari, G. Franzese, L. de Souza Rosa, R. Prakash, Z. Ajanović, M. Ferraz, A. Valada, and J. Kober, "Interactive imitation learning in robotics: A survey," 2022. [Online]. Available: https://arxiv.org/abs/2211.00600

[21] H. Walke, K. Black, A. Lee, M. J. Kim, M. Du, C. Zheng, T. Zhao, P. Hansen-Estruch, Q. Vuong, A. He, V. Myers, K. Fang, C. Finn, and S. Levine, "Bridgedata v2: A dataset for robot learning at scale," 2024. [Online]. Available: https://arxiv.org/abs/2308.12952

[22] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," 2017. [Online]. Available: https://arxiv.org/abs/1706.02413

[23] J. Komorowski, M. Wysoczanska, and T. Trzcinski, "Minkloc++: Lidar and monocular image fusion for place recognition," 2021. [Online]. Available: https://arxiv.org/abs/2104.05327

[24] J. Komorowski, "Minkloc3d: Point cloud based large-scale place recognition," 2020. [Online]. Available: https://arxiv.org/abs/2011.04530

[25] S. Xu, D. Zhou, J. Fang, J. Yin, Z. Bin, and L. Zhang, "Fusionpainting: Multimodal fusion with adaptive attention for 3d object detection," 2021. [Online]. Available: https://arxiv.org/abs/2106.12449

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[27] A. Wilcox, M. Ghanem, M. Moghani, P. Barroso, B. Joffe, and A. Garg, "Adapt3r: Adaptive 3d scene representation for domain transfer in imitation learning," 2025. [Online]. Available: https://arxiv.org/abs/2503.04877

[28] Y. Fu, Q. Feng, N. Chen, Z. Zhou, M. Liu, M. Wu, T. Chen, S. Rong, J. Liu, H. Dong, and S. Zhang, "Cordvip: Correspondence-based visuomotor policy for dexterous manipulation in real-world," 2025. [Online]. Available: https://arxiv.org/abs/2502.08449

[29] S. Chen, R. Garcia, C. Schmid, and I. Laptev, "Polarnet: 3d point clouds for language-guided robotic manipulation," 2023. [Online]. Available: https://arxiv.org/abs/2309.15596

[30] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," 2017. [Online]. Available: https://arxiv.org/abs/1711.06396

[31] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," 2020. [Online]. Available: https://arxiv.org/abs/2003.08934

[32] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," 2023. [Online]. Available: https://arxiv.org/abs/2308.04079

[33] S. Wei, Y. Luo, Y. Wang, and C. Luo, "Robust multimodal learning via representation decoupling," 2024. [Online]. Available: https://arxiv.org/abs/2407.04458

[34] M. K. Reza, A. Prater-Bennette, and M. S. Asif, "Robust multimodal learning with missing modalities via parameter-efficient adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 2, p. 742–754, Feb. 2025. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2024.3476487

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[36] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[37] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[38] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone, "Libero: Benchmarking knowledge transfer for lifelong robot learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 44 776–44 791, 2023.

[39] T. Mu, Z. Ling, F. Xiang, D. Yang, X. Li, S. Tao, Z. Huang, Z. Jia, and H. Su, "Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations," 2021. [Online]. Available: https://arxiv.org/abs/2107.14483

[40] J. Gu, F. Xiang, X. Li, Z. Ling, X. Liu, T. Mu, Y. Tang, S. Tao, X. Wei, Y. Yao, X. Yuan, P. Xie, Z. Huang, R. Chen, and H. Su, "Maniskill2: A unified benchmark for generalizable manipulation skills," 2023. [Online]. Available: https://arxiv.org/abs/2302.04659

[41] S. Tao, F. Xiang, A. Shukla, Y. Qin, X. Hinrichsen, X. Yuan, C. Bao, X. Lin, Y. Liu, T. kai Chan, Y. Gao, X. Li, T. Mu, N. Xiao, A. Gurha, V. N. Rajesh, Y. W. Choi, Y.-R. Chen, Z. Huang, R. Calandra, R. Chen, S. Luo, and H. Su, "Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai," 2025. [Online]. Available: https://arxiv.org/abs/2410.00425

[42] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "Rlbench: The robot learning benchmark & learning environment," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3019–3026, 2020.

[43] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel, "Gello: A general, low-

cost, and intuitive teleoperation framework for robot manipulators," 2024. [Online]. Available: https://arxiv.org/abs/2309.13037

[44] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "Bc-z: Zero-shot task generalization with robotic imitation learning," 2022. [Online]. Available: https://arxiv.org/abs/2202.02005

[45] T. Yu, D. Quillen, Z. He, R. Julian, A. Narayan, H. Shively, A. Bellathur, K. Hausman, C. Finn, and S. Levine, "Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning," 2021. [Online]. Available: https://arxiv.org/abs/1910.10897