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

DSPv2: Improved Dense Policy for Effective and Generalizable Whole-body Mobile Manipulation

Yue Su[1,2,3] Chubin Zhang[2,4] Sijin Chen[1] Liufan Tan[2] Yansong Tang[4] Jianan Wang[2‡] Xihui Liu[1†]
[1]The University of Hong Kong [2]Astribot [3]Xidian University [4]Tsinghua University

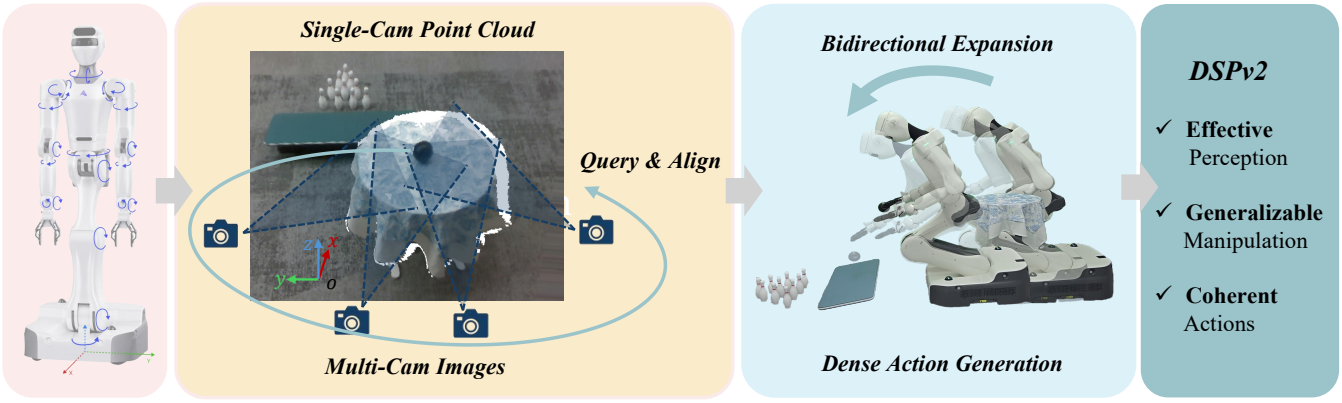

Fig. 1: **DSPv2 is a whole-body mobile manipulation policy** that achieves generalizable performance by fusing multi-view 2D semantic perception with 3D spatial awareness, and generates coherent whole-body actions via dense action head.

***Abstract*— Learning whole-body mobile manipulation via imitation is essential for generalizing robotic skills to diverse environments and complex tasks. However, this goal is hindered by significant challenges, particularly in effectively processing complex observation, achieving robust generalization, and generating coherent actions. To address these issues, we propose DSPv2, a novel policy architecture. DSPv2 introduces an effective encoding scheme that aligns 3D spatial features with multi-view 2D semantic features. This fusion enables the policy to achieve broad generalization while retaining the fine-grained perception necessary for precise control. Furthermore, we extend the Dense Policy paradigm to the whole-body mobile manipulation domain, demonstrating its effectiveness in generating coherent and precise actions for the whole-body robotic platform. Extensive experiments show that our method significantly outperforms existing approaches in both task performance and generalization ability. Project page is available at: https://selen-suyue.github.io/DSPv2Net/.**

I. Introduction

Recent advances in imitation learning have enabled robust performance on manipulation tasks using single or dual-arm configurations [51, 7, 47], culminating in the development of generalizable [41, 12] and general [5, 30, 26] policies. In stark contrast, applying these techniques to whole-body household robots, which possess immense market potential and real-world applicability, still faces considerable difficulties, even in achieving reliable task-specific downstream policies [25].

This difficulty stems primarily from two intertwined aspects: the complexity of the robot's morphology and the expanded scope of its task environment. From the robot's perspective, a whole-body household robot possesses significantly higher DoF and a more expansive workspace compared to fixed-base robot arms [16]. Tasks executed by such a robot often necessitate the coordination of dual-arm manipulation, whole-body motion, and chassis navigation. This coupling results in a complex and high-dimensional action space and imposes stringent requirements on the coordination among its joints. From the scene's perspective, a majority of existing policies [51, 39] are developed for constrained, tabletop manipulation tasks. In contrast, whole-body robots are expected to manipulate in a broad spectrum of domestic and outdoor environments. Thus, they often require multiple cameras to achieve comprehensive awareness. These complicated scene and observation lead to a diversely distributed observation space, thereby substantially compounding the difficulty of policy learning [36].

Research addressing this problem is scarce. Current solutions [25] obtain observations by projecting and merging multi-view point clouds into the base frame, followed by a sampling step. Actions are then generated with diffusion head. However, the sparse sampling required by the 3D vision backbone [33] in this method limits the utilization of multi-view observations, which degrades task execution accuracy. Furthermore, this approach fails to address the problem of generalization. This is, however, crucial for the policy's practical applicability. Beyond the aforementioned issues, the coupled dynamics of whole-body systems present another major hurdle. Minor misalignment between components, such as the manipulator and the mobile base, can rapidly amplify, thereby undermining the coherence of the entire motion [25, 6].

To address these challenges, we propose **DSPv2**, an

‡Project Lead
†Corresponding Author

improved Dense Policy [36] for whole-body mobile manipulation. As shown in Figure 1, our model is designed to effectively leverage multi-view observations to achieve robust whole-body manipulation performance, generate coherent actions by bidirectional autoregressive way, while also generalizing to unseen objects and environments.

Specifically, to address the challenges of leveraging complex observations for generalizable perception, we design an effective 2D-3D feature fusion method. First, DSPv2 projects the uncolored point cloud from the head camera into the base frame as a global observation, which is then processed by a Sparse 3D Encoder [8] to extract voxel-level spatial features. Concurrently, RGB images from cameras on the head, wrist, and torso are encoded into patch-level semantic features using a fine-tuned lightweight vision foundation model [31]. To fuse these modalities, we design a Q-former [27] that uses the positional information of the spatial voxels to query for and align with corresponding semantic features within the multi-view 2D feature maps. This decoupled approach allows the spatial features to be processed independently, with the foundation model learning color features to achieve scene-level generalization, while the fusion process facilitates the utilization of geometry-aware multi-view features. To tackle the problem of error amplification and motion incoordination in action generation, the fused observation features are fed into a Dense Head [36]. It performs coarse-to-fine autoregressive generation and applies bidirectional attention across the temporal dimension. This mechanism allows us to mitigate error amplification and incoordination among different robot components during trajectory generation, thereby achieving precise and robust action prediction.

We conduct extensive real-world experiments, including five different types of tasks and five distinct generalization tests. We compare our method against widely-used 2D and 3D policies [7, 47], as well as other whole-body mobile manipulation policies [25]. We also apply DSPv2 to various action heads and compare the results. Furthermore, we evaluate the effectiveness and generalization of different configurations of our visual encoder. Our key contributions are as follows:

- We propose an effective whole-body mobile manipulation policy that fully utilizes multi-view observations by aligning 3D spatial features and 2D semantic features, surpassing existing whole-body policies.
- We propose the first generalization approach for whole-body downstream policies and present its strong generalization even with limited task-specific data.
- We extend the Dense Policy to whole-body mobile manipulation and demonstrate its effectiveness in generating coherent and precise whole-body actions.

II. Related Work

A. Imitation Learning for Manipulation

Advances in imitation learning [22, 13] have led to progress in robotic manipulation. For tasks set in complex scenes, 2D policies [7, 51] often learn by encoding multi-view observations to gain a comprehensive understanding. Early 3D policies [47, 46] encoded downsampled sparse point clouds to obtain robust feature representations after pooling [33], a step necessary for stable, denoising-based action generation. However, this method sacrifices comprehensive spatial information, thereby limiting performance [39]. Some 3D foundation policies [43, 24] attempt to improve performance through pre-training on a large number of heterogeneous robot datasets, while other approaches use RGB-D representations to mitigate this deficit [29]. Currently, high-performing 3D policies [39, 12, 37] employ sparse convolutions on voxels to extract voxel-level features without losing scene information. This representation balances information preservation with the robustness of feature representation, resulting in better action generation. We believe this encoding method is well-suited for household scene understanding, especially under complex action and observation space.

Furthermore, in the action generation module, some policies employ generative models to map observations to action distributions [7, 4]. An advantage of these models is their ability to fit diverse action patterns without deterministic sampling [51]. More recently, many studies have explored using autoregressive models [10, 34] for action generation [50, 14, 18, 29, 36]. These models leverage temporal dependencies between actions to produce more coherent trajectories and can mitigate the error accumulation that diffusion models may suffer from due to environmental disturbances and domain shift [2, 36].

B. Policy-level Generalization

Aside from Vision-Language-Action (VLA) models [4, 21, 26], which possess generalization capabilities from pre-training on large, heterogeneous data [9], the generalization of policies [41, 38] trained on task-specific data often relies on incorporating lightweight vision foundation models [31, 48]. This allows them to adapt to variations in scenes and manipulated objects. Current methods [12] extending this approach to 3D involve processing single-view point cloud and corresponding image independently. The features from the vision foundation model are then projected to their corresponding locations in the point cloud to enable generalizable dual-arm tabletop manipulation.

C. Whole-body Mobile Manipulation

Although many studies have focused on mobile manipulation [49, 6, 45] and whole-body manipulation [44, 40], work extending to whole-body mobile manipulation remains scarce. Current solutions [25] encode the merged point cloud using PointNet [33] and then generate actions for different robot components, from the chassis to the arms, with diffusion models [19, 35]. The actions for later-generated components are autoregressively conditioned on the former to avoid the error amplification highlighted in this research. Furthermore, this approach has not yet achieved policy-level generalization. Our research just aims to suppress the aforementioned errors and achieve effective manipulation by

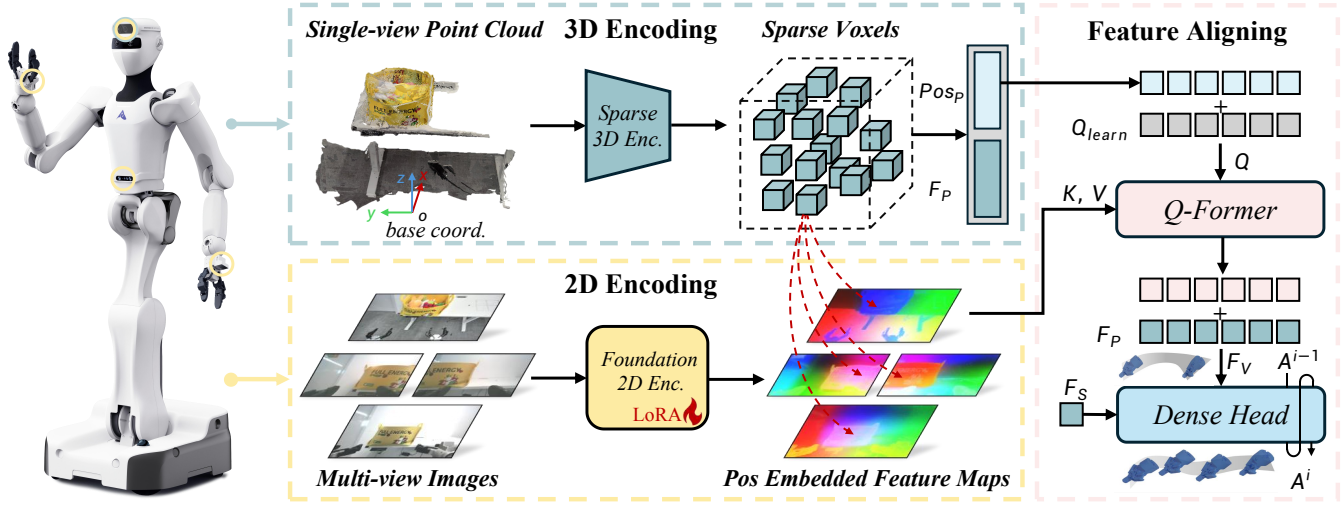

Fig. 2: **Overview of DSPv2**. First, a sparse 3D encoder processes the point cloud, which is projected from the head-mounted camera to the base frame, to obtain voxel-level feature tokens. A 2D vision foundation model is also used to acquire patch-level feature maps. Subsequently, in the feature alignment process, a Q-former is designed to query multi-view semantic features from feature maps for the voxels and fuse them with spatial features, based on the positional information of the 3D voxels and 2D patches. Finally, the resulting features are fed into a dense head to generate the future action sequence in a bidirectional autoregressive paradigm.

developing a more efficient visual encoder and an autoregressive policy that leverages temporal dependencies, with an effective method to realize policy-level generalization.

III. Method

A. Problem Formulation

Our platform is a 25-DOF robot consisting of a 3-DOF chassis, a 4-DOF torso, dual 7-DOF arms, a 2-DOF head, and 2 grippers. The robot is controlled via pose commands for each component relative to the base [16]. Specifically, the action $\boldsymbol{a}_t$ at each timestep t includes the above components' poses, where the chassis's action is an offset from its previous state. The task of our policy is therefore to predict the future action sequence $\boldsymbol{A}_{t:t+T} = \{\boldsymbol{a}_t, \boldsymbol{a}_{t+1}, \ldots, \boldsymbol{a}_{t+T-1}\}$ over a horizon of T time steps.

Given the observation at the current timestep: $\boldsymbol{O}_t = \{\boldsymbol{I}_t, \boldsymbol{P}_t, \boldsymbol{S}_t\}$, where $\boldsymbol{I}_t$ comprises the RGB images from the robot's head, torso, and dual wrist cameras; $\boldsymbol{P}_t$ is the uncolored point cloud from the head camera, projected into the base frame to mitigate the effects of varying head poses on policy learning [25]; and $\boldsymbol{S}_t$ denotes the current robot state. In line with the standard behavior cloning framework, our objective is to learn a policy that maximizes the conditional probability $\boldsymbol{P}(\boldsymbol{A}|\boldsymbol{O})$ over a dataset of expert demonstrations. The overall process of DSPv2 is shown in the Figure 2

B. Observation Encoding

For 3D encoding, we use a sparse convolutional network [8] to extract voxel features, $\boldsymbol{F}_P$, from the uncolored geometric point cloud. This process yields a highly sparse representation with fewer than 300 tokens, with the upper bound determined by the observation space and voxel resolution of our tasks. For each of these feature tokens, we generate a corresponding positional embedding, $\boldsymbol{Pos}_P$, using a sine-cosine function on its spatial location [39].

To encode 2D features, we utilize a DINOv2-base backbone [31] fine-tuned with LoRA [20]. This allows the model to adapt to our specific task while retaining its powerful, task-agnostic semantic representations, which is crucial for generalization. This encoder generates patch-level feature maps $\boldsymbol{F}_I^{v_1-v_n}$ from n views, to which we add learnable positional embeddings, $\boldsymbol{Pos}_I$, to encode the spatial location of each patch.

The robot state is defined by the current poses of its components, with one important exception: we exclude the chassis's pose from the input. Our experiments indicate this forces the policy to rely on environmental context for navigation, rather than its own trajectory history, which improves learning. To further prevent the policy from overfitting, we randomly mask the remaining pose inputs with a 30% probability during training [36]. The state representation is then encoded into the feature space as $\boldsymbol{F}_S$ by an MLP.

C. Feature Aligning

The visual features subsequently undergo a fusion operation. For this, We chose the Q-former [27] architecture, inspired by its success in vision-language modeling, for its effectiveness in distilling salient information from large feature maps into a fixed set of queries, which is ideal for aligning sparse 3D tokens with dense 2D features.

Thus, we design a Q-former that maintains a set of 300 learnable query tokens, $\boldsymbol{Q}_{learn}$. The function of this module is to use the positional information of the sparse spatial tokens to query for corresponding features in the multi-view images, thereby aligning the spatial and semantic features.

Specifically, the learnable tokens are first added to the

spatial positional embeddings $\boldsymbol{Pos}_P$ to create spatially-aware queries, $\boldsymbol{Q}$, as follows:

$$\boldsymbol{Q} = \boldsymbol{Q}_{learn} + \boldsymbol{Pos}_P. \tag{1}$$

These queries then attend to the 2D feature maps from all views ($\boldsymbol{F}_I^{v_1-v_n}$), which, augmented with their respective positional embeddings ($\boldsymbol{Pos}_I$), serve as the keys ($\boldsymbol{K}$) and values ($\boldsymbol{V}$) in a cross-attention mechanism. In this attention process, positional information effectively flows from the 3D voxel grid to the 2D patch grids. This flow serves as an efficient indexing mechanism, enabling the policy to align features across multiple views. The output of this operation, representing the retrieved semantic features, is then added to the original spatial features $\boldsymbol{F}_P$. This sum forms the final fused visual representation, $\boldsymbol{F}_V$, formulated as:

$$\boldsymbol{F}_V = Attn(\boldsymbol{Q}, \boldsymbol{F}_I^{v_1-v_n}+\boldsymbol{Pos}_I, \boldsymbol{F}_I^{v_1-v_n}+\boldsymbol{Pos}_I)+\boldsymbol{F}_P. \tag{2}$$

D. Dense Generation

Beyond the challenges of effective information utilization and generalization, a significant focus in recent research has been the problem of error amplification at the action generation stage [25]. Specifically, traditional diffusion models are prone to compounding errors, particularly when facing distributional shifts [2]. In the context of diffusion-based policies for whole-body robots, this issue is manifested during the denoising process: an action dimension for one component can only attend to the states of other components at the same noise level. For example, during the early denoising steps, the torso's action cannot be conditioned on a fully-denoised state of the mobile base. Consequently, coordination between components is compromised, leading to a hierarchical accumulation of errors from the base up to the dual arms [6, 25].

Although method mentioned in Sec. II-C addresses the problem semi-autoregressively, this comes at the cost of substantially prolonged inference latency due to its hierarchical diffusion structure. Such a delay is a considerable drawback in dynamic household environments [32, 21, 3, 11].

Thus, we adopt Dense Policy [36] as our action head, which has been demonstrated to be an inference-efficient autoregressive policy. It achieves coarse-to-fine autoregressive generation by predicting the action sequence in a bidirectional, expanding fashion from the observation. This characteristic allows the action for one robot component to be conditioned on the already-determined actions of other components at critical timesteps within the trajectory. As a result, inter-component coordination is improved, which in turn mitigates the error amplification phenomenon [1]. The logarithmic inference complexity of Dense Policy also reduces the number of inference iterations while increasing the number of causal inference steps in time, thereby curbing the amplification of errors in time series [23].

As a result, the action generation can be formulated as:

$$\boldsymbol{P}(\boldsymbol{A}|\boldsymbol{F}_V, \boldsymbol{F}_S) = \prod_{i=1}^{n} \boldsymbol{P}(\boldsymbol{A}^i|\boldsymbol{A}^{i-1}, \boldsymbol{A}^{i-2}, \ldots, \boldsymbol{A}^0, \boldsymbol{F}_V, \boldsymbol{F}_S), \tag{3}$$

where

$$\boldsymbol{A}^n = \{\boldsymbol{a}_{t+i}^n |\ i\ mod\ \frac{T}{2^n} = 0, \quad i \in \mathbb{N}_{<T}\}. \tag{4}$$

IV. Experiments

A. Tasks

1) Evaluation Tasks: We evaluate our policy on five distinct tasks:

Pick and Place: The robot navigates to a cloth basket on a table, picks and then places it in an open area on the floor.

Deliver: The robot grasps a water bottle from a low table and delivers it to the counter.

Sort: The robot stacks two popcorn buckets scattered on the ground into a largest popcorn tube. We denote the steps of these two stacks as *Stack-I* and *Stack-II*

Bowling: The robot grasps a bowling ball from a table, navigates to a suitable position, and throws the ball to knock down the target pins. The whole process is divided into *Grasp* and *Hit*.

Cart Pushing: The robot must grasp the handle of a shopping cart and push it to navigate to a 1.5 meters distance through an environment while avoiding obstacles.

2) Generalization Tests: Additionally, we designed five challenging generalization tests to evaluate the policy's robustness under conditions not present in the expert demonstrations.

Light: The *Pick and Place* task is performed under low-light conditions, with the main room lights turned off.

Spatial Arrangement: For the *Pick and Place* task, the height of the table is raised by 5cm.

Object Color: In the *Deliver* task, the color of the target water bottle is changed to one unseen during training.

Object Shape: The *Sort* task is performed with popcorn buckets of a novel shape and smaller relative size.

Scene: The *Cart Pushing* task is executed in a completely new environment with a different layout and background.

The above experiments are shown in the Figure 3.

B. Setup

Platform. Our experiments are conducted on the Astribot-S1 robot [16], a platform with 25 degrees of freedom (DOFs). For onboard computation, it is equipped with an Intel i7-1370PE CPU. The robot's perception system includes an Orbbec Femto Bolt on the head, an Orbbec Gemini 335 on the torso, and an Intel Realsense D401 on each of the two wrists. All of these are RGB-D cameras. Additionally, the mobile base is mounted with two Livox MID-360 LiDARs, which were not used in our experiments. All expert demonstrations were collected via teleoperation using a Meta Quest 3S VR headset.

Demonstrations. Unless otherwise specified, we collected 100 demonstrations for each task.

Baseline. In addition to WB-WIMA [25], which also targets whole-body mobile manipulation, we select classical visuomotor policies DP [7] and DP3 [47] as baselines for comparison. Furthermore, to evaluate the effectiveness of Dense Head for whole-body tasks, we also include a variant

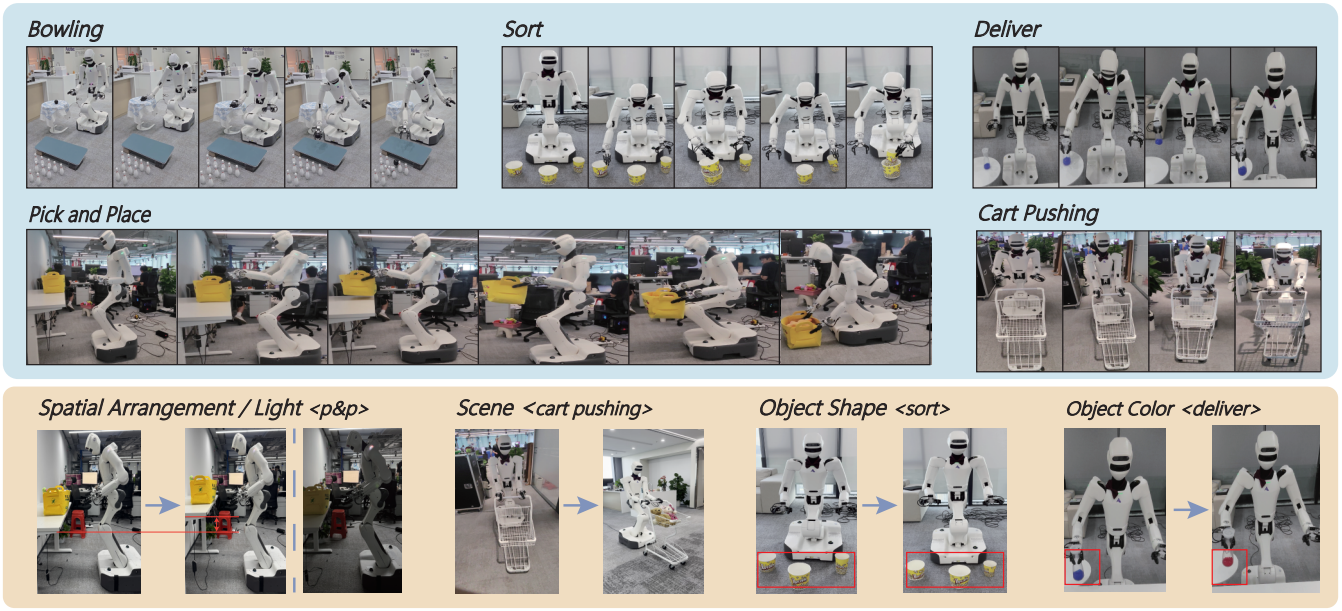

Fig. 3: **Overview of the procedure of our experiments.** The blue area above shows the execution process of the five tasks, and the yellow area below shows the setup of our generalization tests.

Method	*Pick and Place*		*Deliver*	*Sort*		*Bowling*		*Cart Pushing*
	Pick (%)	*Place (%)*	*Succ. (%)*	*Stack-I (%)*	*Stack-II (%)*	*Grasp (%)*	*Hit (%)*	*Succ. (%)*
WB-WIMA [25]	50	25	45	45	0	65	40	**90**
DP3 [47]	40	10	10	45	10	55	35	80
DP [7]	30	20	55	50	15	50	40	80
DSPv2	**80**	**60**	**100**	**90**	**25**	**80**	**50**	**90**

TABLE I: **Detailed performance of DSPv2 and baselines in real-world tasks under the original setup.**

that pairs our visual encoder with a Diffusion Head for comparison.

Protocols. For the real-world evaluation, 20 trials are performed per method for each task, unless otherwise specified. All methods are evaluated under closely matched randomized initial scene configurations for each trial.

C. Effective Manipulation

As shown in Table I, DSPv2 achieves the best performance on the majority of tasks. We attribute this result primarily to the following aspects:

Effective Environmental Perception. Compared to baseline policies, DSPv2 utilizes and effectively fuses both multi-view semantic features and 3D spatial features. This contributes to a more comprehensive perception of the environment. In the ***Pick and Place*** task, the robot must first navigate to the target location to pick the basket, and the accuracy of navigation depends on observation and localization of the basket. DSPv2 is consistently able to move to a position directly facing the target, an aspect where other policies often deviate. In the ***Cart Pushing*** task, after grasping the cart, the robot must move 1.5 meters while avoiding obstacles. This requires coordinated actions from the chassis and the arm to adjust its direction, which relies on obstacle detection. The superior spatial perception of DSPv2 enables it to achieve the best obstacle avoidance performance.

Precise Action Generation. The comprehensive encoding and fusion of observational information yield finer-grained features for action generation, resulting in higher action precision for DSPv2. This is evident in the ***Pick and Place*** task, where DSPv2 consistently grasps the basket's edge from above. In contrast, other 3D baselines often grasp the side of the basket horizontally, which causes the object to be dropped during the place phase, leading to task failure even after a successful pick. In the ***Sort*** task, high-precision alignment is required to stack the grasped popcorn bucket into the target bucket. For this, the sparse sampling of 3D baselines is insufficient for accurate depth perception when the two buckets are close, causing errors in the generated actions. Meanwhile, 2D policies have greater difficulty estimating depth. In contrast, DSPv2 leverages its efficient feature representation to accomplish this fine-grained action.

Coherent Whole-Body Motion. As previously discussed, a key advantage of Dense Head is that any given robot component can refine its action generation by observing the planned actions of other components within the keyframes in the trajectory. Unlike diffusion models, which are conditioned on noised values, our policy observes raw actions.

Method	*Pick and Place*				*Deliver*	*Sort*	*Cart Pushing*
	Spatial Arrangement		*Light*		*Object Color*	*Object Shape*	*Scene*
	Pick (%)	*Place (%)*	*Pick (%)*	*Place (%)*	*Succ. (%)*	*Stack-I (%)*	*Succ. (%)*
WB-WIMA [25]	50→30	25→0	50→30	25→10	45→35	45→20	90→60
DP3 [47]	40→0	10→0	40→30	10→0	10→10	45→40	80→60
DP [7]	30→30	20→0	30→0	20→0	55→35	50→50	80→60
DSPv2	80→**50**	60→**20**	80→**60**	60→**40**	100→**85**	90→**80**	90→**80**

TABLE II: **Detailed performance of DSPv2 and baselines in Generalization Tests.** The left side of the arrow is the performance of the original setup, and the right side is the performance of the generalized setup.

This allows it to better capture temporal dependencies, thereby mitigating error propagation and enhancing motion coherence. For instance, the ***Deliver*** task requires grasping a bottle from a low table, a maneuver that demands tight coordination among the chassis, torso, and arm. This task admits multiple solutions through different combinations of movements (e.g., the chassis can move further forward, allowing the torso to lean further back), making seamless coordination paramount. The ***Bowling*** task presents a similar challenge, not only in grasping the ball but also in the release phase. The arm's final release posture is conditioned on both the torso's height and the distance moved by the chassis. Such scenarios place high demands on motion coherence. By leveraging its effective feature processing and bidirectional autoregressive action generation, DSPv2 consistently outperforms the baselines in these complex, coordinated tasks.

D. Generalization Ability

The performance of the policies in the five generalization tests is presented in Table II. DSPv2 maintains its leading performance over the baselines on the vast majority of tasks, and its relative performance drop compared to the original task setting is minimal. Specifically, its generalization performance is demonstrated in two main aspects:

Spatial Understanding. In the ***Spatial Arrangement*** experiment, the table height was increased, while in the ***Object Shape*** experiment, the sizes of the popcorn buckets were altered. Both modifications significantly change the point cloud occupancy of the scene. Consequently, 3D baselines that rely on sparse sampling for encoding are severely impacted, as these changes lead to substantial shifts in their features after max-pooling, degrading policy performance.

In contrast, DSPv2's Sparse Encoder processes the complete point cloud at the voxel level without information loss, yielding a more robust representation. Furthermore, even when the point cloud shifts, the Q-former's role in binding voxels to their corresponding multi-view semantic features makes the resulting representation more resilient to such disturbances. This leads to more stable observation features and enhances the policy's generalization capability.

It is worth noting that in the second stage of the ***Sort*** task, where the target bucket's height was substantially reduced, all methods, including DSPv2, failed to perform the pick. This is because the expert demonstrations contained virtually no examples of the right arm picking from such a low height. As established in prior works, current imitation learning policies struggle to generalize to actions that lie significantly outside the distribution of the expert demonstrations.

Semantic Understanding. In the ***Light***, ***Object Color***, and ***Scene*** experiments, we altered the ambient lighting, object colors, and scene background, respectively. These modifications introduce significant visual changes, to which the 2D and 3D baselines are particularly sensitive as they lack visual pre-training. In contrast, DSPv2 isolates the impact of these visual shifts. Its 3D branch focuses exclusively on spatial information by processing the color-agnostic xyz point cloud. Meanwhile, its 2D branch uses a Foundation Encoder pre-trained on upstream tasks, yielding semantic information with stronger generalization capabilities. By aligning these two feature sources, the model minimizes the performance degradation caused by visual domain shifts.

E. Ablation

We divide our ablation study into two parts: the Vision Encoder and the Action Head:

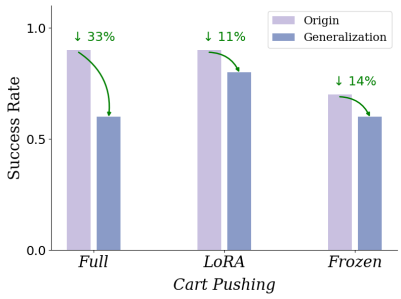

Fig. 4: Performance of DSPv2 under different fine-tuning strategies for DINOv2 in the ***Cart Pushing*** task.

Vision Encoder. In this section, we investigate the optimal learning method for the 2D vision foundation model to achieve policy-level generalization. We conduct our comparison on the ***Cart Pushing*** task and its corresponding scene generalization experiment. Specifically, we evaluate three approaches for DINOv2 [31]: full fine-tuning of all parameters, LoRA fine-tuning [20], and freezing all parameters.

The results in Figure 4 show that in the original scene, both the fully fine-tuned and LoRA-tuned models achieve a 90% success rate, whereas freezing DINOv2 yields only a

Method	*Pick and Place*		*Deliver*	*Sort*		*Bowling*		*Cart Pushing*
	Pick (%)	*Place (%)*	*Succ. (%)*	*Stack-I (%)*	*Stack-II (%)*	*Grasp (%)*	*Hit (%)*	*Succ. (%)*
DSPv2	**80**	**60**	**100**	**90**	**25**	80	**50**	**90**
DSPv2 w. Diff Head	50	30	**100**	60	15	**90**	35	80

TABLE III: **Detailed performance of DSPv2 and DSPv2 with diffusion head in tasks under the original setup.**

Method	*Pick and Place*				*Deliver*	*Sort*	*Cart Pushing*
	Spatial Arrangement		*Light*		*Object Color*	*Object Shape*	*Scene*
	Pick (%)	*Place (%)*	*Pick (%)*	*Place (%)*	*Succ. (%)*	*Stack-I (%)*	*Succ. (%)*
DSPv2	80→**50**	60→**20**	80→**60**	60→**40**	100→85	90→**80**	90→**80**
DSPv2 w. Diff Head	50→40	30→0	50→40	30→20	100→**90**	60→40	80→60

TABLE IV: **Detailed performance of DSPv2 and DSPv2 with diffusion head in Generalization Tests.** The left side of the arrow is the performance of the original setup, and the right side is the performance of the generalized setup.

70% success rate. This indicates that LoRA fine-tuning does not sacrifice the policy's precision in extracting semantic features compared to full fine-tuning. In contrast, freezing the vision model prevents it from capturing fine-grained information [28].

When the scene is altered, the success rates change to 60%, 80%, and 60%, respectively. The fully fine-tuned approach suffers a substantial performance drop, as the policy loses the inherent semantic priors of the foundation model. The frozen DINOv2 policy also experiences a larger performance degradation than LoRA. We posit that while freezing the backbone preserves all semantic priors, it prevents the model from extracting sufficient task-level information. This causes the model to rely more heavily on the 3D encoder, which in turn destabilizes the 3D-2D feature binding performed by the Q-former in the generalization setting, reducing the robustness of the visual encoding. Therefore, DSPv2 adopts DINOv2 fine-tuned with LoRA as its final 2D encoding method.

Action Head. In Table III and IV, we also present the performance of DSPv2 when its action head is replaced with a diffusion head. The experimental results show that DSPv2's encoder, when paired with a diffusion head, still outperforms the baselines and exhibits strong generalization. However, the Dense Head demonstrates superior performance.

We attribute this to the mechanism previously discussed: the Dense Policy effectively utilizes keyframe actions from preceding levels within its autoregressive process. This renders the robot's different components mutually "visible" during generation, which enhances policy coherence and mitigates error propagation [17]. This advantage becomes particularly critical in the generalization tests. When the observation distribution shifts, generative models such as diffusion models are prone to larger initial errors in the observation-to-action mapping [2, 15]. This error is then significantly amplified during the generation process. In contrast, DSPv2 circumvents this problem, leading to more robust performance.

V. LIMITATIONS

We acknowledge that DSPv2 has several limitations. First, for whole-body mobile manipulation, DSPv2 cannot yet solve tasks that require highly constrained or high-frequency actions, which are common in real-world scenarios [42, 3]; future work could explore incorporating high-frequency modalities like tactile sensing into the policy or addressing this through suitable whole-body action integration. Second, regarding generalization, DSPv2, like many imitation learning algorithms, fails to generalize to action modes not covered in the expert demonstrations; this might require the design of an action head with stronger reasoning capabilities to provide support. Finally, a promising direction for household robotics is the development of a general VLA model capable of executing diverse, long-horizon, and multi-stage tasks from language instructions. This remains a challenge for future work in whole-body mobile manipulation, and we look forward to the emergence of large-scale whole-body mobile manipulation datasets to support this research, along with VLA designs adapted for them.

VI. CONCLUSION

We propose DSPv2, an efficient and generalizable dense policy for whole-body mobile manipulation. By effectively combining sparse 3D spatial features with multi-view 2D semantic features, DSPv2 achieves robust perception of complex environments. Furthermore, we extend the Dense Policy to whole-body tasks, and its bidirectional autoregressive generation mechanism produces coherent and precise actions, significantly outperforming existing methods. Extensive real-world experiments demonstrate that DSPv2 achieves significant advancements in both task performance and generalization, offering a promising solution for deploying whole-body robotic systems in complex environments.

VII. ACKNOWLEDGMENTS

The research work described in this paper was conducted in the JC STEM Lab of Autonomous Intelligent Systems funded by The Hong Kong Jockey Club Charities Trust.

REFERENCES

[1] Anurag Ajay et al. *Is Conditional Generative Modeling all you need for Decision-Making?* 2023. arXiv: 2211.15657 [cs.LG]. URL: https://arxiv.org/abs/2211.15657.

[2] Jacob Austin et al. *Structured Denoising Diffusion Models in Discrete State-Spaces*. 2023. arXiv: 2107.03006 [cs.LG]. URL: https://arxiv.org/abs/2107.03006.

[3] Kevin Black, Manuel Y. Galliker, and Sergey Levine. *Real-Time Execution of Action Chunking Flow Policies*. 2025. arXiv: 2506.07339 [cs.RO]. URL: https://arxiv.org/abs/2506.07339.

[4] Kevin Black et al. π_0*: A Vision-Language-Action Flow Model for General Robot Control*. 2024.

[5] Chilam Cheang and Sijin Chen et al. *GR-3 Technical Report*. 2025. arXiv: 2507.15493 [cs.RO]. URL: https://arxiv.org/abs/2507.15493.

[6] Sixiang Chen et al. *AC-DiT: Adaptive Coordination Diffusion Transformer for Mobile Manipulation*. 2025. arXiv: 2507.01961 [cs.RO]. URL: https://arxiv.org/abs/2507.01961.

[7] Chi, C. et al. "Diffusion Policy: Visuomotor Policy Learning via Action Diffusion". In: *RSS*. 2023.

[8] Christopher Choy, JunYoung Gwak, and Silvio Savarese. "4d spatio-temporal convnets: Minkowski convolutional neural networks". In: *CVPR*. 2019.

[9] Open X-Embodiment Collaboration et al. "Open X-Embodiment: Robotic Learning Datasets and RT-X Models". In: *ICRA*. 2024.

[10] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019.

[11] Danny Driess et al. *Knowledge Insulating Vision-Language-Action Models: Train Fast, Run Fast, Generalize Better*. 2025. arXiv: 2505.23705 [cs.LG]. URL: https://arxiv.org/abs/2505.23705.

[12] Hongjie Fang and Chenxi Wang et al. "AirExo-2: Scaling up Generalizable Robotic Imitation Learning with Low-Cost Exoskeletons". In: *arXiv preprint arXiv:03081* (2025).

[13] P. Florence et al. "Implicit behavioral cloning". In: *CoRL*. 2022.

[14] Letian Fu et al. "In-Context Imitation Learning via Next-Token Prediction". In: *arXiv preprint arXiv:2408.15980* (2024).

[15] Dechen Gao et al. *VITA: Vision-to-Action Flow Matching Policy*. 2025. arXiv: 2507.13231 [cs.CV]. URL: https://arxiv.org/abs/2507.13231.

[16] Guang Gao and Jianan Wang et al. *Towards Human-level Intelligence via Human-like Whole-Body Manipulation*. 2025. arXiv: 2507.17141 [cs.RO]. URL: https://arxiv.org/abs/2507.17141.

[17] Ziteng Gao and Mike Zheng Shou. *D-AR: Diffusion via Autoregressive Models*. 2025. arXiv: 2505.23660 [cs.CV]. URL: https://arxiv.org/abs/2505.23660.

[18] Zhefei Gong et al. "CARP: Visuomotor Policy Learning via Coarse-to-Fine Autoregressive Prediction". In: *arXiv preprint arXiv:2412.06782* (2024).

[19] J. Ho, A. Jain, and P. Abbeel. "Denoising diffusion probabilistic models". In: *NeurIPS*. 2020.

[20] Edward J. Hu et al. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. arXiv: 2106.09685 [cs.CL]. URL: https://arxiv.org/abs/2106.09685.

[21] Physical Intelligence and Kevin Black et al. $\pi_{0.5}$*: a Vision-Language-Action Model with Open-World Generalization*. 2025. arXiv: 2504.16054 [cs.LG]. URL: https://arxiv.org/abs/2504.16054.

[22] Jang, E. et al. "BC-Z: Zero-Shot Task Generalization with Robotic Imitation Learning". In: *CoRL*. 2021.

[23] Michael Janner et al. *Planning with Diffusion for Flexible Behavior Synthesis*. 2022. arXiv: 2205.09991 [cs.LG]. URL: https://arxiv.org/abs/2205.09991.

[24] Yueru Jia et al. *Lift3D Foundation Policy: Lifting 2D Large-Scale Pretrained Models for Robust 3D Robotic Manipulation*. 2024. arXiv: 2411.18623 [cs.CV]. URL: https://arxiv.org/abs/2411.18623.

[25] Yunfan Jiang et al. "BEHAVIOR Robot Suite: Streamlining Real-World Whole-Body Manipulation for Everyday Household Activities". In: *9th Annual Conference on Robot Learning*. 2025. URL: https://openreview.net/forum?id=v2KevjWScT.

[26] Moo Jin Kim, Karl Pertsch, and Karamcheti et al. "OpenVLA: An Open-Source Vision-Language-Action Model". In: *arXiv preprint arXiv:2406.09246* (2024).

[27] Junnan Li et al. *BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models*. 2023. arXiv: 2301.12597 [cs.CV]. URL: https://arxiv.org/abs/2301.12597.

[28] Haotian Liu et al. *Visual Instruction Tuning*. 2023. arXiv: 2304.08485 [cs.CV]. URL: https://arxiv.org/abs/2304.08485.

[29] Yiyang Lu et al. *H^3DP: Triply-Hierarchical Diffusion Policy for Visuomotor Learning*. 2025. arXiv: 2505.07819 [cs.RO]. URL: https://arxiv.org/abs/2505.07819.

[30] Octo Model Team. "Octo: An Open-Source Generalist Robot Policy". In: *RSS*. 2024.

[31] Maxime Oquab, Timothée Darcet, and Moutakanni et al. *DINOv2: Learning Robust Visual Features without Supervision*. 2023.

[32] Karl Pertsch et al. *FAST: Efficient Action Tokenization for Vision-Language-Action Models*. 2025. arXiv: 2501.09747 [cs.RO].

[33] Charles R Qi et al. "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation". In: *arXiv preprint arXiv:1612.00593* (2016).

[34] Alec Radford et al. "Improving language understanding by generative pre-training". In: (2018).

[35] J. Song, C. Meng, and S. Ermon. "Denoising diffusion implicit models". In: *ICLR*. 2021.

[36] Yue Su et al. "Dense Policy: Bidirectional Autoregressive Learning of Actions". In: *arXiv preprint arXiv:2503.13217* (2025).

[37] Yue Su et al. "Motion Before Action: Diffusing Object Motion as Manipulation Condition". In: *IEEE Robotics and Automation Letters* 10.7 (2025), pp. 7428–7435. DOI: 10.1109/LRA.2025.3577424.

[38] Yihe Tang et al. *UAD: Unsupervised Affordance Distillation for Generalization in Robotic Manipulation*. 2025. arXiv: 2506.09284 [cs.RO]. URL: https://arxiv.org/abs/2506.09284.

[39] Chenxi Wang et al. "RISE: 3D Perception Makes Real-World Robot Imitation Simple and Effective". In: *IROS*. 2024.

[40] Congcong Wen et al. *Humanoid Agent via Embodied Chain-of-Action Reasoning with Multimodal Foundation Models for Zero-Shot Loco-Manipulation*. 2025. arXiv: 2504.09532 [cs.RO]. URL: https://arxiv.org/abs/2504.09532.

[41] Shangning Xia et al. "CAGE: Causal Attention Enables Data-Efficient Generalizable Robotic Manipulation". In: *arXiv preprint arXiv:2410.14974* (2024).

[42] Han Xue et al. *Reactive Diffusion Policy: Slow-Fast Visual-Tactile Policy Learning for Contact-Rich Manipulation*. 2025. arXiv: 2503.02881 [cs.RO].

[43] Rujia Yang et al. *FP3: A 3D Foundation Policy for Robotic Manipulation*. 2025. arXiv: 2503.08950 [cs.RO]. URL: https://arxiv.org/abs/2503.08950.

[44] Shaofeng Yin et al. *VisualMimic: Visual Humanoid Loco-Manipulation via Motion Tracking and Generation*. 2025. arXiv: 2509.20322 [cs.RO]. URL: https://arxiv.org/abs/2509.20322.

[45] Zhecheng Yuan et al. *HERMES: Human-to-Robot Embodied Learning from Multi-Source Motion Data for Mobile Dexterous Manipulation*. 2025. arXiv: 2508.20085 [cs.RO]. URL: https://arxiv.org/abs/2508.20085.

[46] Yanjie Ze et al. "Generalizable Humanoid Manipulation with 3D Diffusion Policies". In: *arXiv preprint arXiv:2410.10803* (2024).

[47] Ze, Y. et al. "3D Diffusion Policy: Generalizable Visuomotor Policy Learning via Simple 3D Representations". In: *RSS*. 2024.

[48] Xiaohua Zhai et al. *Sigmoid Loss for Language Image Pre-Training*. 2023. arXiv: 2303.15343 [cs.CV]. URL: https://arxiv.org/abs/2303.15343.

[49] Jiazhao Zhang et al. *GAMMA: Graspability-Aware Mobile MAnipulation Policy Learning based on Online Grasping Pose Fusion*. 2024. arXiv: 2309.15459 [cs.RO]. URL: https://arxiv.org/abs/2309.15459.

[50] Xinyu Zhang et al. *Autoregressive Action Sequence Learning for Robotic Manipulation*. 2024.

[51] Zhao, T. Z. et al. "Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware". In: *RSS*. 2023.