Novelty Assessment Report

Paper: H^{3}DP: Triply-Hierarchical Diffusion Policy for Visuomotor Learning

PDF URL: https://openreview.net/pdf?id=Q1CP0iAmOb

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-29

Abstract

Visuomotor policy learning has witnessed substantial progress in robotic manipulation, with recent approaches predominantly relying on generative models to model the action distribution. However, these methods often overlook the critical coupling between visual perception and action prediction. In this work, we introduce Triply-Hierarchical Diffusion Policy (H^{3}DP), a novel visuomotor learning framework that explicitly incorporates hierarchical structures to strengthen the integration between visual features and action generation. H^{3}DP contains $\mathbf{3}$ levels of hierarchy: (1) depth-aware input layering that organizes RGB-D observations based on depth information; (2) multi-scale visual representations that encode semantic features at varying levels of granularity; and (3) a hierarchically conditioned diffusion process that aligns the generation of coarse-to-fine actions with corresponding visual features. Extensive experiments demonstrate that H^{3}DP yields a $+ \mathbf{27.5}$% average relative improvement over baselines across $ \mathbf{44}$ simulation tasks and achieves superior performance in $\mathbf{4}$ challenging bimanual real-world manipulation tasks. Project Page: https://h3-dp.github.io/.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Visuomotor Policy Learning with Hierarchical Visual-Action Coupling**

A total of **50 papers** were analyzed and organized into a taxonomy with **20 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Hierarchical Policy Architectures**
- **Multi-Scale Visual Representation Learning**
- **Generative Action Models**
- **Multi-Modal Sensory Integration**
- **Specialized Task Domains**
- **Learning Paradigms and Training Strategies**
- **Neuroscience-Inspired and Cognitive Models**

Complete Taxonomy Tree

- Visuomotor Policy Learning with Hierarchical Visual-Action Coupling Survey Taxonomy
- Hierarchical Policy Architectures
  - Vision-Language-Action Hierarchies
  - VLM-Based Task Decomposition (5 papers)
    - [5] HAMSTER: Hierarchical Action Models For Open-World Robot Manipulation (Li Yi, 2025) View paper
    - [9] Roboos: A hierarchical embodied framework for cross-embodiment and multi-agent collaboration (Tan Huajie, 2025) View paper
    - [35] A Hierarchical Vision-Language and Reinforcement Learning Framework for Robotic Task and Motion Planning in Collaborative Manipulation (Junnan Zhang, 2026) View paper
    - [45] From Code to Action: Hierarchical Learning of Diffusion-VLM Policies (Peschl, 2025) View paper
    - [46] RDD: Retrieval-Based Demonstration Decomposer for Planner Alignment in Long-Horizon Tasks (Yan Mingxuan, 2025) View paper
  - Direct VLA Prediction (5 papers)
    - [4] LLaDA-VLA: Vision Language Diffusion Action Models (Wen Yu-qing, 2025) View paper
    - [14] Semanticvla: Semantic-aligned sparsification and enhancement for efficient robotic manipulation (Wei Li, 2025) View paper
    - [15] VDRive: Leveraging Reinforced VLA and Diffusion Policy for End-to-end Autonomous Driving (Guo Ziang, 2025) View paper
    - [16] Large VLM-based Vision-Language-Action Models for Robotic Manipulation: A Survey (Shao Rui, 2025) View paper
    - [38] HiMoE-VLA: Hierarchical Mixture-of-Experts for Generalist Vision-Language-Action Policies (Zhiying Du, 2025) View paper
  - Fast-Slow Dual-System VLA (3 papers)
    - [18] A Self-Correcting Vision-Language-Action Model for Fast and Slow System Manipulation (Li Chen-Xuan, 2024) View paper
    - [22] MoTVLA: A Vision-Language-Action Model with Unified Fast-Slow Reasoning (Huang Wenhui, 2025) View paper
    - [41] Hierarchical Vision Language Action Model Using Success and Failure Demonstrations (Jeongeun Park, 2025) View paper
  - Non-Linguistic Hierarchical Control
  - Skill-Based Hierarchical Learning (4 papers)
    - [8] Hierarchical visuomotor control of humanoids (Josh Merel, 2018) View paper
    - [27] Skill-Based Hierarchical Reinforcement Learning for Target Visual Navigation (Shuo Wang, 2023) View paper
    - [40] Learning transferable motor skills with hierarchical latent mixture policies (Rao, 2021) View paper
    - [43] Generalizable Hierarchical Skill Learning via Object-Centric Representation (Zhao Haibo, 2025) View paper

    - Temporal Hierarchies (3 papers)
        - [20] Hierarchical Visual Policy Learning for Long-Horizon Robot Manipulation in Densely Cluttered Scenes (Wang Hecheng, 2023) View paper
        - [28] Hierarchical Learning for Closed-Loop Robotic Manipulation in Cluttered Scenes via Depth Vision, Reinforcement Learning, and Behaviour Cloning (Helmut C Y Yu, 2025) View paper
        - [33] Learning from Interventions using Hierarchical Policies for Safe Learning (Jing Bi, 2022) View paper
- Multi-Scale Visual Representation Learning
    - Hierarchical Visual Feature Encoding ★ (4 papers)
    - [0] H^{3}DP: Triply-Hierarchical Diffusion Policy for Visuomotor Learning (Anon et al., 2026) View paper
    - [1] Spatial Policy: Guiding Visuomotor Robotic Manipulation with Spatial-Aware Modeling and Reasoning (Liu Yi-jun, 2025) View paper
    - [6] HDP: Triply-Hierarchical Diffusion Policy for Visuomotor Learning (Y Lu, 2025) View paper
    - [48] VAT: Vision Action Transformer by Unlocking Full Representation of ViT (Wenhao Li, 2025) View paper
    - Spatial Attention and Graph-Based Reasoning (3 papers)
    - [2] InterACT: Inter-dependency Aware Action Chunking with Hierarchical Attention Transformers for Bimanual Manipulation (Lee Andrew, 2024) View paper
    - [12] Information-Theoretic Graph Fusion with Vision-Language-Action Model for Policy Reasoning and Dual Robotic Control (Shunlei Li, 2025) View paper
    - [34] Robust Urban Navigation for Autonomous Vehicles using Vision based Segmentation Graph Attention and Hierarchical Learning (Mohammed I. Habelalmateen, 2025) View paper
    - Multi-View and Depth-Aware Perception (2 papers)
    - [31] Multi-Modal Fusion in Contact-Rich Precise Tasks via Hierarchical Policy Learning (Jin, 2022) View paper
    - [44] Uni-Sight: An E2E Vision-Language-Action System Unifying Multi-View Alignment and Multi-Modal Fusion (Daixun Li, 2025) View paper
- Generative Action Models
    - Diffusion-Based Action Generation (2 papers)
    - [17] EquiContact: A Hierarchical SE(3) Vision-to-Force Equivariant Policy for Spatially Generalizable Contact-rich Tasks (Seo, 2025) View paper
    - [50] Compositional Foundation Models for Hierarchical Planning (Ajay, 2023) View paper
    - Intermediate Representation Generation (2 papers)
    - [10] NVSPolicy: Adaptive Novel-View Synthesis for Generalizable Language-Conditioned Policy Learning (Shi Le, 2025) View paper
    - [13] Pixel motion as universal representation for robot control (Ranasinghe, 2025) View paper
    - Frequency and Continuous Token Representations (1 papers)
    - [49] FreqPolicy: Frequency Autoregressive Visuomotor Policy with Continuous Tokens (Zhong Yiming, 2025) View paper
- Multi-Modal Sensory Integration (1 papers)
    - [19] See, feel, act: Hierarchical learning for complex manipulation skills with multisensory fusion (M. Oller, 2019) View paper
- Specialized Task Domains
    - Bimanual and Dual-Arm Manipulation (1 papers)
    - [39] RGMP: Recurrent Geometric-prior Multimodal Policy for Generalizable Humanoid Robot Manipulation (Xuetao Li, 2025) View paper
    - Autonomous Driving and Navigation (1 papers)
    - [3] Hierarchical generative adversarial imitation learning with mid-level input generation for autonomous driving on urban environments (Gustavo Claudio Karl Couto, 2024) View paper
    - Navigation and Exploration (3 papers)
    - [11] CL-CoTNav: Closed-Loop Hierarchical Chain-of-Thought for Zero-Shot Object-Goal Navigation with Vision-Language Models (Cai Yu-xin, 2025) View paper
    - [24] Towards socially aware visual navigation with hierarchical learning (Johnson, 2025) View paper
    - [25] Hierarchical Cross-Modal Agent for Robotics Vision-and-Language Navigation (Muhammad Zubair Irshad, 2021) View paper
    - Non-Robotic Control Applications (1 papers)
    - [29] Towards General Computer Control with Hierarchical Agents and Multi-Level Action Spaces (Dong Zihan, 2025) View paper
- Learning Paradigms and Training Strategies
    - Reinforcement Learning Approaches (2 papers)
    - [21] Hierarchical reinforcement learning, sequential behavior, and the dorsal frontostriatal system (Miriam Janssen, 2022) View paper
    - [26] Sample-Efficient Robot Skill Learning for Construction Tasks: Benchmarking Hierarchical Reinforcement Learning and Vision-Language-Action VLA Model (Zhaofeng Hu, 2025) View paper
    - Cross-Embodiment and Heterogeneous Data (1 papers)
    - [42] VAMOS: A Hierarchical Vision-Language-Action Model for Capability-Modulated and Steerable Navigation (Castro, 2025) View paper
- Neuroscience-Inspired and Cognitive Models
    - Active Inference and Predictive Coding (3 papers)
    - [23] Active Predictive Coding: A Unifying Neural Model for Active Perception, Compositional Learning, and Hierarchical Planning (Rajesh P. N. Rao, 2023) View paper
    - [30] A hierarchical active inference framework for stable robotic control (D Liu, 2025) View paper
    - [47] An active inference model of hierarchical action understanding, learning and imitation. (Riccardo Proietti, 2023) View paper
    - Neural Mechanisms and Brain-Inspired Architectures (4 papers)
    - [7] Dynamic causal model application on hierarchical human motor control estimation in visuomotor tasks (Ningjia Yang, 2024) View paper
    - [32] Neural and computational mechanisms of action processing: interaction between visual and motor representations (Martin A. Giese, 2015) View paper
    - [36] Learning to perceive the world as articulated: an approach for hierarchical learning in sensory-motor systems (J Tani, 1999) View paper
    - [37] Neuronal correlates of continuous manual tracking under varying visual movement feedback in a virtual reality environment (Jakub Limanowski, 2017) View paper

Narrative

Core task: visuomotor policy learning with hierarchical visual-action coupling. The field addresses how agents can learn to map visual observations to motor actions by exploiting structure at multiple levels of abstraction. The taxonomy reveals several complementary perspectives: Hierarchical Policy Architectures decompose decision-making into high-level planning and low-level control (e.g., Hierarchical Imitation Driving[3], HAMSTER[5]); Multi-Scale Visual Representation Learning focuses on extracting features at different spatial or temporal resolutions (e.g., HDP[6], VAT[48]); Generative Action Models treat action sequences as distributions to be sampled or refined; Multi-Modal Sensory Integration combines vision with tactile or proprioceptive signals; Specialized Task Domains target navigation, manipulation, or driving; Learning Paradigms span imitation, reinforcement, and self-supervised methods; and Neuroscience-Inspired models draw on predictive coding or active inference principles. Together, these branches reflect a shared goal of bridging the gap between raw sensory input and coordinated motor output through layered representations.

A particularly active line of work explores how to encode visual information hierarchically so that coarse scene understanding guides fine-grained action selection. H3DP[0] sits within the Multi-Scale Visual Representation Learning branch, specifically under Hierarchical Visual Feature Encoding, where it emphasizes coupling visual features at different scales directly to corresponding action granularities. This contrasts with approaches like Spatial Policy[1], which may prioritize spatial attention mechanisms, or HDP[6], which structures policies around explicit hierarchical decompositions of the action space. Meanwhile, VAT[48] leverages transformer architectures for multi-scale encoding, highlighting a trend toward attention-based feature aggregation. The central trade-off across these methods involves balancing representational expressiveness—capturing rich visual detail—with computational efficiency and sample complexity during training. H3DP[0] addresses this by tightly integrating visual and action hierarchies, aiming to improve generalization across tasks that demand both global scene context and precise local control.

Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

1. Spatial Policy: Guiding Visuomotor Robotic Manipulation with Spatial-Aware Modeling and Reasoning

Authors: Liu Yi-jun, Liu Yuwei, Yijun Liu, Meng Yuan, Yuwei Liu, et al. (21 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Vision-centric hierarchical embodied models have demonstrated strong potential. However, existing methods lack spatial awareness capabilities, limiting their effectiveness in bridging visual plans to actionable control in complex environments. To address this problem, we propose Spatial Policy (SP), a unified spatial-aware visuomotor robotic manipulation framework via explicit spatial modeling and reasoning. Specifically, we first design a spatial-conditioned embodied video generation module to ...

Relationship Analysis

Both papers belong to the Hierarchical Visual Feature Encoding category, focusing on multi-scale visual representations for visuomotor policy learning. They overlap in using hierarchical visual features to guide action generation, with both employing depth information and multi-scale architectures. However, H^{3}DP uses a triply-hierarchical diffusion-based approach with depth-aware layering and coarse-to-fine action generation aligned with diffusion denoising stages, while Spatial Policy focuses on spatial-conditioned video generation with flow-based action prediction and spatial reasoning feedback for replanning.

2. HDP: Triply-Hierarchical Diffusion Policy for Visuomotor Learning

Authors: Y Lu, Y Tian, Z Yuan, X Wang, P Hua, et al. (6 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

â¦ However, these methods often overlook the critical coupling between visual perception and â¦ policy learning framework grounded in three levels of hierarchy: input, representation, and â¦

⚠ Similarity Notice

These papers appear to be the same work or very closely related variants. Both introduce H^{3}DP (Triply-Hierarchical Diffusion Policy) with identical core contributions: depth-aware input layering, multi-scale visual representations, and hierarchically conditioned diffusion for visuomotor learning. The titles, abstracts, methodology, and experimental results are nearly identical, suggesting they are likely the same paper at different submission stages or venues.

3. VAT: Vision Action Transformer by Unlocking Full Representation of ViT

Authors: Wenhao Li, Chengwei Ma, Weixin Mao | **Year/Venue**: 2025 | **URL**: View paper

Abstract

In robot learning, Vision Transformers (ViTs) are standard for visual perception, yet most methods discard valuable information by using only the final layer's features. We argue this provides an insufficient representation and propose the Vision Action Transformer (VAT), a novel architecture that is extended from ViT and unlocks the full feature hierarchy of ViT. VAT processes specialized action tokens with visual features across all transformer layers, enabling a deep and progressive fusion of...

Relationship Analysis

Both papers belong to the Hierarchical Visual Feature Encoding category, focusing on extracting and fusing multi-scale visual representations for visuomotor policy learning. They overlap in their use of hierarchical visual features from Vision Transformers to improve action prediction, with both emphasizing the importance of multi-level feature extraction rather than single-layer representations. However, H^{3}DP introduces a triply-hierarchical framework with depth-aware layering, multi-scale visual representations, and hierarchically conditioned diffusion for action generation, while VAT focuses specifically on unlocking the full feature hierarchy of ViT by processing action tokens across all transformer layers without the depth-based input structuring or diffusion-based action generation approach.

Contributions Analysis

Overall novelty summary. The paper proposes a triply-hierarchical diffusion policy framework that integrates depth-aware input layering, multi-scale visual representations, and hierarchically conditioned action generation. It resides in the Hierarchical Visual Feature Encoding leaf, which contains four papers total (including this one). This leaf sits within Multi-Scale Visual Representation Learning, a moderately populated branch addressing how visual features at different granularities inform action prediction. The taxonomy reveals this is an active but not overcrowded research direction, with sibling leaves exploring spatial attention mechanisms and multi-view perception.

The broader Multi-Scale Visual Representation Learning branch neighbors Hierarchical Policy Architectures (which decomposes tasks into high-level planning and low-level execution) and Generative Action Models (which treats actions as distributions). The Hierarchical Visual Feature Encoding leaf explicitly excludes methods using only final-layer features or single-scale encodings, positioning it as a middle ground between flat visual processing and full task decomposition. Nearby leaves like Spatial Attention and Graph-Based

Reasoning focus on relational modeling rather than scale-based feature hierarchies, while Multi-View and Depth-Aware Perception emphasizes viewpoint integration over hierarchical conditioning.

Among thirty candidates examined, the triply-hierarchical framework itself shows no clear refutation (ten candidates, zero refutable). The hierarchically conditioned diffusion process similarly appears novel (ten candidates, zero refutable). However, the depth-aware layering strategy encounters one refutable candidate among ten examined, suggesting some prior work addresses depth-based input organization. The limited search scope means these statistics reflect top-K semantic matches and citation expansion, not exhaustive coverage. The framework's novelty appears strongest in its integrated coupling of visual and action hierarchies rather than individual components.

Given the search examined thirty candidates across three contributions, the analysis captures immediate semantic neighbors but cannot rule out relevant work outside this scope. The taxonomy structure suggests the paper occupies a moderately explored niche where hierarchical visual encoding meets diffusion-based action generation. The depth-aware layering component shows the most overlap with prior art, while the end-to-end integration of three hierarchical levels appears less directly anticipated by examined literature.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Triply-Hierarchical Diffusion Policy (H³DP) framework

Description: The authors propose H³DP, a visuomotor policy learning framework that integrates three levels of hierarchy: depth-aware input layering of RGB-D observations, multi-scale visual representations encoding features at varying granularity, and hierarchically conditioned diffusion process aligning coarse-to-fine action generation with corresponding visual features.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. LLaDA-VLA: Vision Language Diffusion Action Models

URL: View paper

Brief Assessment

LLaDA-VLA[4] focuses on vision-language-action models for robotic manipulation using masked diffusion models adapted from pretrained vision-language models, while the original paper presents a visuomotor policy learning framework with depth-aware layering, multi-scale visual representations, and hierarchically conditioned diffusion. These are fundamentally different architectural approaches and application domains.

2. Hierarchical Visual Policy Learning for Long-Horizon Robot Manipulation in Densely Cluttered Scenes

URL: View paper

Brief Assessment

Hierarchical Visual Policy[20] focuses on hierarchical reinforcement learning with high-level policy and action primitives (push, pick, place) for cluttered manipulation, not on diffusion-based visuomotor policy learning with multi-scale visual representations and hierarchically conditioned diffusion processes.

3. Vision-Language-Action Model and Diffusion Policy Switching Enables Dexterous Control of an Anthropomorphic Hand

URL: View paper

Brief Assessment

VLA Diffusion Switching[74] focuses on combining VLA models with diffusion policies for task switching in dexterous manipulation, not on hierarchical visuomotor policy learning with depth-aware layering, multi-scale representations, and hierarchically conditioned diffusion processes as proposed in H³DP.

4. MinD: Unified Visual Imagination and Control via Hierarchical World Models

URL: View paper

Brief Assessment

MinD[71] focuses on a dual-system world model combining video generation with action policies for risk-aware planning, not on hierarchical integration of RGB-D observations, multi-scale visual representations, and hierarchically conditioned diffusion for visuomotor learning as in H³DP.

5. Any2Policy: Learning Visuomotor Policy with Any-Modality

URL: View paper

Brief Assessment

Any2Policy[73] focuses on multi-modal learning (text, audio, image, video, point cloud) for robotic policy learning, not on hierarchical diffusion-based visuomotor frameworks. The candidate addresses modality fusion and alignment rather than hierarchical visual-action coupling in diffusion processes.

6. HDP: Triply-Hierarchical Diffusion Policy for Visuomotor Learning

URL: View paper

Brief Assessment

HDP[6] is the same paper as the original submission (both describe H³DP with identical three-level hierarchy). This is not a prior work that could refute novelty.

7. HIQL: Offline Goal-Conditioned RL with Latent States as Actions

URL: View paper

Brief Assessment

HIQL[72] focuses on offline goal-conditioned RL using hierarchical policies with latent states as actions, not visuomotor policy learning with diffusion models. The hierarchical structures serve fundamentally different purposes: HIQL uses high-level policies to predict subgoal states and low-level policies for primitive actions in goal-reaching tasks, while H³DP integrates depth-aware visual layering with multi-scale representations for diffusion-based action generation.

8. HieroAction: Hierarchically Guided VLM for Fine-Grained Action Analysis

URL: View paper

Brief Assessment

HieroAction[70] focuses on evaluating human actions through vision-language models with stepwise reasoning and hierarchical policy learning for action assessment. This is fundamentally different from H³DP's visuomotor policy learning framework for robotic manipulation that integrates depth-aware input layering, multi-scale visual representations, and hierarchically conditioned diffusion processes.

9. Spatial Policy: Guiding Visuomotor Robotic Manipulation with Spatial-Aware Modeling and Reasoning

URL: View paper

Brief Assessment

Spatial Policy[1] focuses on spatial-aware modeling through spatial plan tables and flow-based action prediction for embodied video generation, which is fundamentally different from H^{3}DP's depth-aware layering, multi-scale visual representations, and hierarchically conditioned diffusion process for visuomotor policy learning.

10. : A Vision-Language-Action Flow Model for General Robot Control

URL: View paper

Brief Assessment

VLA Flow Model[69] focuses on a vision-language-action model for general robot control using flow matching, not on hierarchical visuomotor policy learning with depth-aware layering and multi-scale representations as proposed in H^{3}DP.

Contribution 2: Depth-aware layering strategy for RGB-D input

Description: The authors introduce a method that decomposes RGB-D images into multiple non-overlapping layers based on depth values, enabling the policy to explicitly distinguish foreground from background and suppress distractors and occlusions, thereby enhancing spatial structure understanding in cluttered visual scenarios.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Depth Helps: Improving Pre-trained RGB-based Policy with Depth Information Injection

URL: View paper

Brief Assessment

Depth Information Injection[65] focuses on predicting depth from RGB for deployment scenarios, not on decomposing RGB-D into depth-based layers for spatial structure understanding. The candidate uses depth completion and codebooks for missing modality learning, while the original contribution explicitly partitions RGB-D into non-overlapping layers based on depth values to distinguish foreground/background.

2. Enhancing spatial awareness via multi-modal fusion of cnn-based visual and depth features

URL: View paper

Brief Assessment

Multi-Modal Spatial Awareness[66] focuses on intermediate fusion of RGB and depth features for semantic segmentation in indoor scenes, without employing a depth-based layering strategy that decomposes images into non-overlapping layers. The candidate's approach fuses modalities at the feature level rather than organizing input data by depth boundaries to suppress distractors.

3. Learning depth-aware deep representations for robotic perception

URL: View paper

Prior Art Analysis

Depth-Aware Representations[67] demonstrates prior work that uses depth information to decompose RGB-D inputs into multiple layers for robotic perception tasks. The candidate paper introduces 'daconv' (depth-aware convolution), a CNN block that explicitly exploits depth to learn scale-aware feature representations by performing convolutions at multiple scales guided by depth-dependent functions. This approach partitions RGB-D data based on depth values to enable networks to distinguish foreground from background and handle spatial structure understanding. The candidate's method of using depth to create layered representations that help suppress distractors and improve spatial reasoning predates the original paper's claimed novelty of depth-aware layering for manipulation tasks.

Evidence

Evidence 1 - **Rationale**: Both papers propose using depth information to create structured representations that improve spatial understanding in robotic perception tasks. The candidate explicitly introduces a depth-aware mechanism for CNNs. - **Original**: we introducedepth-aware layeringstrategy that partitions the rgb-d input into distinct layers based on depth cues. this approach not only enables the policy to explicitly distinguish between foreground and background, but also suppresses distractors and occlusions [40; 1], thereby enhancing the unde... - **Candidate**: we show that the performance of deep architectures can be boosted by introducing daconv, a novel, general-purpose cnn block which exploits depth to learn scale-aware feature representations. we demonstrate the benefits of daconv on a variety of robotics oriented tasks, involving affordance detection,...

Evidence 2 - **Rationale**: Both papers describe mechanisms for partitioning or processing RGB-D data based on depth values to create layered or scale-aware representations that distinguish spatial regions. - **Original**: define{d 0 =d min, d1, . . . , dn =d max}as the depth boundaries for each layer. image layeri m is formed by selecting pixels with depth in[d m-1, dm), i.e., m(i,j) m =i [dm-1≤d(i,j) <dm], i m =i⊙m m,(1) whereianddare the rgb-d image and depth map, respectively,iis the indicator function. this repre... - **Candidate**: we introduce an additional network (depthnet) fed with depth information, working in parallel with the main network (predictionnet). the role of depthnet is to provide the daconv blocks in predictionnet with depth-related features that will trigger the decision about which scale to choose within e...

4. Hierarchical, Dense and Dynamic 3D Reconstruction Based on VDB Data Structure for Robotic Manipulation Tasks

URL: View paper

Brief Assessment

Hierarchical VDB Reconstruction[63] focuses on 3D scene reconstruction using TSDF integration of depth images for robotic manipulation, not on depth-aware layering strategies for policy learning or visual perception in visuomotor control.

5. RoboFlamingo-Plus: Fusion of Depth and RGB Perception with Vision-Language Models for Enhanced Robotic Manipulation

URL: View paper

Brief Assessment

RoboFlamingo-Plus[61] focuses on integrating depth data into vision-language models for language-guided manipulation tasks, using a pre-trained resampler and cross-attention mechanisms. This differs fundamentally from the original paper's depth-aware layering strategy, which decomposes RGB-D images into non-overlapping layers based on depth values to suppress distractors and enhance spatial structure understanding in cluttered scenarios.

6. Integrating visual foundation models for enhanced robot manipulation and motion planning: A layered approach

URL: View paper

Brief Assessment

Visual Foundation Integration[64] focuses on a high-level layered framework for integrating visual foundation models across perception, cognition, planning, execution, and learning layers. It does not describe a depth-aware layering strategy that decomposes RGB-D images into multiple non-overlapping layers based on depth values to suppress distractors and enhance spatial structure understanding.

7. HDP: Triply-Hierarchical Diffusion Policy for Visuomotor Learning

URL: View paper

Brief Assessment

HDP[6] is the same paper as the original submission. Both describe the identical depth-aware layering mechanism using the same mathematical formulation and implementation.

8. Act3D: 3D Feature Field Transformers for Multi-Task Robotic Manipulation

URL: View paper

Brief Assessment

Act3D[62] uses 3D feature fields with adaptive resolution sampling in a coarse-to-fine manner, but does not decompose RGB-D images into depth-based layers. The original paper's depth-aware layering explicitly partitions images into non-overlapping depth layers to distinguish foreground/background, which is a different approach from Act3D's 3D point sampling strategy.

9. Disentangled Object-Centric Image Representation for Robotic Manipulation

URL: View paper

Brief Assessment

Disentangled Object-Centric[68] focuses on semantic segmentation-based disentanglement (robot, objects of interest, obstacles) rather than depth-based layering. The candidate does not demonstrate prior work on depth-aware layering strategies.

10. Structered deep visual models for robot manipulation

URL: View paper

Brief Assessment

Structured Visual Models[60] focuses on a five-layer framework (perception, cognition, planning, execution, learning) for robot manipulation using visual foundation models, without any mention of depth-aware layering or RGB-D decomposition strategies. The candidate does not address the specific technical contribution of partitioning RGB-D images into depth-based layers.

Contribution 3: Hierarchically conditioned diffusion process for action generation

Description: The authors design a diffusion-based action generation mechanism where coarse visual features guide initial denoising steps to shape global action structure (low-frequency components), while fine-grained features inform later steps to refine precise details (high-frequency components), establishing tighter coupling between action generation and visual encoding.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Generate the Forest before the Trees -- A Hierarchical Diffusion model for Climate Downscaling

URL: View paper

Brief Assessment

Forest before Trees[59] focuses on climate downscaling with spatial hierarchies for weather data generation, not robotic action generation or visuomotor policy learning. The hierarchical diffusion process operates on completely different data modalities and tasks.

2. Leveraging the Spatial Hierarchy: Coarse-to-fine Trajectory Generation via Cascaded Hybrid Diffusion

URL: View paper

Brief Assessment

Cascaded Trajectory Diffusion[53] focuses on trajectory generation for urban mobility data using a two-stage cascaded diffusion process (road segment-level and GPS-level), not visuomotor action generation conditioned on multi-scale visual features. The hierarchical structure serves different purposes in different domains.

3. HieraSurg: Hierarchy-Aware Diffusion Model for Surgical Video Generation

URL: View paper

Brief Assessment

HieraSurg[52] focuses on surgical video generation with semantic segmentation guidance, not robotic action generation. The hierarchical conditioning operates on visual features for video synthesis rather than coupling visual perception with robotic action prediction.

4. Generate subgoal images before act: Unlocking the chain-of-thought reasoning in diffusion model for robot manipulation with multimodal prompts

URL: View paper

Brief Assessment

Generate Subgoal Images[56] focuses on using diffusion models for high-level visual planning (generating subgoal images from multimodal prompts), not for hierarchically conditioning action generation with coarse-to-fine visual features as in the original paper.

5. Coarse-to-Fine: a Hierarchical Diffusion Model for Molecule Generation in 3D

URL: View paper

Brief Assessment

Coarse-to-Fine Molecule[51] applies hierarchical diffusion to molecular generation in 3D chemistry, not robotic action generation. The candidate focuses on generating molecular structures with coarse fragment representations followed by fine-grained atomic assembly, which is fundamentally different from visuomotor policy learning for robotic manipulation.

6. Act As You Wish: Fine-Grained Control of Motion Diffusion Model with Hierarchical Semantic Graphs

URL: View paper

Brief Assessment

Act As Wish[58] focuses on text-driven human motion generation using hierarchical semantic graphs, not visuomotor policy learning with visual features guiding robotic action generation.

7. Equivariant Blurring Diffusion for Hierarchical Molecular Conformer Generation

URL: View paper

Brief Assessment

Equivariant Blurring Diffusion[54] addresses molecular conformer generation with a two-stage hierarchical approach (coarse fragment-level then fine atomic details), which is fundamentally different from the original paper's hierarchically conditioned diffusion for robotic action generation guided by multi-scale visual features.

8. CoCoDiff: Diversifying Skeleton Action Features via Coarse-Fine Text-Co-Guided Latent Diffusion

URL: View paper

Brief Assessment

CoCoDiff[57] focuses on skeleton-based action recognition using diffusion models to generate diverse action features in latent space, not on visuomotor policy learning with hierarchical visual-action coupling as in the original paper.

9. Hierarchical Diffusion Policy for Kinematics-Aware Multi-Task Robotic Manipulation

URL: View paper

Brief Assessment

Hierarchical Diffusion Policy[55] uses a hierarchical structure where a high-level agent predicts next-best poses and a low-level diffusion policy generates trajectories, but does not employ the specific coarse-to-fine visual feature conditioning mechanism described in the original paper where different visual scales guide different denoising stages.

10. HDP: Triply-Hierarchical Diffusion Policy for Visuomotor Learning

URL: View paper

Brief Assessment

HDP[6] is the same paper as the original submission. Both describe the same hierarchically conditioned diffusion process where coarse features guide early denoising and fine features guide later steps.

Appendix: Text Similarity Detection

Textual similarity detection checked 29 papers and found 3 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. HDP: Triply-Hierarchical Diffusion Policy for Visuomotor Learning

Detected in: Core Task (sibling), Contribution: contribution_1, Contribution: contribution_2, Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] H^{3}DP: Triply-Hierarchical Diffusion Policy for Visuomotor Learning View paper
- [1] Spatial Policy: Guiding Visuomotor Robotic Manipulation with Spatial-Aware Modeling and Reasoning View paper
- [2] InterACT: Inter-dependency Aware Action Chunking with Hierarchical Attention Transformers for Bimanual Manipulation View paper
- [3] Hierarchical generative adversarial imitation learning with mid-level input generation for autonomous driving on urban environments View paper
- [4] LLaDA-VLA: Vision Language Diffusion Action Models View paper
- [5] HAMSTER: Hierarchical Action Models For Open-World Robot Manipulation View paper
- [6] HDP: Triply-Hierarchical Diffusion Policy for Visuomotor Learning View paper
- [7] Dynamic causal model application on hierarchical human motor control estimation in visuomotor tasks View paper
- [8] Hierarchical visuomotor control of humanoids View paper
- [9] Roboos: A hierarchical embodied framework for cross-embodiment and multi-agent collaboration View paper
- [10] NVSPolicy: Adaptive Novel-View Synthesis for Generalizable Language-Conditioned Policy Learning View paper
- [11] CL-CoTNav: Closed-Loop Hierarchical Chain-of-Thought for Zero-Shot Object-Goal Navigation with Vision-Language Models View paper
- [12] Information-Theoretic Graph Fusion with Vision-Language-Action Model for Policy Reasoning and Dual Robotic Control View paper
- [13] Pixel motion as universal representation for robot control View paper
- [14] Semanticvla: Semantic-aligned sparsification and enhancement for efficient robotic manipulation View paper
- [15] VDRive: Leveraging Reinforced VLA and Diffusion Policy for End-to-end Autonomous Driving View paper
- [16] Large VLM-based Vision-Language-Action Models for Robotic Manipulation: A Survey View paper
- [17] EquiContact: A Hierarchical SE(3) Vision-to-Force Equivariant Policy for Spatially Generalizable Contact-rich Tasks View paper
- [18] A Self-Correcting Vision-Language-Action Model for Fast and Slow System Manipulation View paper
- [19] See, feel, act: Hierarchical learning for complex manipulation skills with multisensory fusion View paper
- [20] Hierarchical Visual Policy Learning for Long-Horizon Robot Manipulation in Densely Cluttered Scenes View paper
- [21] Hierarchical reinforcement learning, sequential behavior, and the dorsal frontostriatal system View paper
- [22] MoTVLA: A Vision-Language-Action Model with Unified Fast-Slow Reasoning View paper
- [23] Active Predictive Coding: A Unifying Neural Model for Active Perception, Compositional Learning, and Hierarchical Planning View paper
- [24] Towards socially aware visual navigation with hierarchical learning View paper
- [25] Hierarchical Cross-Modal Agent for Robotics Vision-and-Language Navigation View paper
- [26] Sample-Efficient Robot Skill Learning for Construction Tasks: Benchmarking Hierarchical Reinforcement Learning and Vision-Language-Action VLA Model View paper

- [27] Skill-Based Hierarchical Reinforcement Learning for Target Visual Navigation View paper
- [28] Hierarchical Learning for Closed-Loop Robotic Manipulation in Cluttered Scenes via Depth Vision, Reinforcement Learning, and Behaviour Cloning View paper
- [29] Towards General Computer Control with Hierarchical Agents and Multi-Level Action Spaces View paper
- [30] A hierarchical active inference framework for stable robotic control View paper
- [31] Multi-Modal Fusion in Contact-Rich Precise Tasks via Hierarchical Policy Learning View paper
- [32] Neural and computational mechanisms of action processing: interaction between visual and motor representations View paper
- [33] Learning from Interventions using Hierarchical Policies for Safe Learning View paper
- [34] Robust Urban Navigation for Autonomous Vehicles using Vision based Segmentation Graph Attention and Hierarchical Learning View paper
- [35] A Hierarchical Vision-Language and Reinforcement Learning Framework for Robotic Task and Motion Planning in Collaborative Manipulation View paper
- [36] Learning to perceive the world as articulated: an approach for hierarchical learning in sensory-motor systems View paper
- [37] Neuronal correlates of continuous manual tracking under varying visual movement feedback in a virtual reality environment View paper
- [38] HiMoE-VLA: Hierarchical Mixture-of-Experts for Generalist Vision-Language-Action Policies View paper
- [39] RGMP: Recurrent Geometric-prior Multimodal Policy for Generalizable Humanoid Robot Manipulation View paper
- [40] Learning transferable motor skills with hierarchical latent mixture policies View paper
- [41] Hierarchical Vision Language Action Model Using Success and Failure Demonstrations View paper
- [42] VAMOS: A Hierarchical Vision-Language-Action Model for Capability-Modulated and Steerable Navigation View paper
- [43] Generalizable Hierarchical Skill Learning via Object-Centric Representation View paper
- [44] Uni-Sight: An E2E Vision-Language-Action System Unifying Multi-View Alignment and Multi-Modal Fusion View paper
- [45] From Code to Action: Hierarchical Learning of Diffusion-VLM Policies View paper
- [46] RDD: Retrieval-Based Demonstration Decomposer for Planner Alignment in Long-Horizon Tasks View paper
- [47] An active inference model of hierarchical action understanding, learning and imitation. View paper
- [48] VAT: Vision Action Transformer by Unlocking Full Representation of ViT View paper
- [49] FreqPolicy: Frequency Autoregressive Visuomotor Policy with Continuous Tokens View paper
- [50] Compositional Foundation Models for Hierarchical Planning View paper
- [51] Coarse-to-Fine: a Hierarchical Diffusion Model for Molecule Generation in 3D View paper
- [52] HieraSurg: Hierarchy-Aware Diffusion Model for Surgical Video Generation View paper
- [53] Leveraging the Spatial Hierarchy: Coarse-to-fine Trajectory Generation via Cascaded Hybrid Diffusion View paper
- [54] Equivariant Blurring Diffusion for Hierarchical Molecular Conformer Generation View paper
- [55] Hierarchical Diffusion Policy for Kinematics-Aware Multi-Task Robotic Manipulation View paper
- [56] Generate subgoal images before act: Unlocking the chain-of-thought reasoning in diffusion model for robot manipulation with multimodal prompts View paper
- [57] CoCoDiff: Diversifying Skeleton Action Features via Coarse-Fine Text-Co-Guided Latent Diffusion View paper
- [58] Act As You Wish: Fine-Grained Control of Motion Diffusion Model with Hierarchical Semantic Graphs View paper
- [59] Generate the Forest before the Trees -- A Hierarchical Diffusion model for Climate Downscaling View paper
- [60] Structered deep visual models for robot manipulation View paper
- [61] RoboFlamingo-Plus: Fusion of Depth and RGB Perception with Vision-Language Models for Enhanced Robotic Manipulation View paper
- [62] Act3D: 3D Feature Field Transformers for Multi-Task Robotic Manipulation View paper
- [63] Hierarchical, Dense and Dynamic 3D Reconstruction Based on VDB Data Structure for Robotic Manipulation Tasks View paper
- [64] Integrating visual foundation models for enhanced robot manipulation and motion planning: A layered approach View paper
- [65] Depth Helps: Improving Pre-trained RGB-based Policy with Depth Information Injection View paper
- [66] Enhancing spatial awareness via multi-modal fusion of cnn-based visual and depth features View paper
- [67] Learning depth-aware deep representations for robotic perception View paper
- [68] Disentangled Object-Centric Image Representation for Robotic Manipulation View paper
- [69] : A Vision-Language-Action Flow Model for General Robot Control View paper
- [70] HieroAction: Hierarchically Guided VLM for Fine-Grained Action Analysis View paper
- [71] MinD: Unified Visual Imagination and Control via Hierarchical World Models View paper
- [72] HIQL: Offline Goal-Conditioned RL with Latent States as Actions View paper
- [73] Any2Policy: Learning Visuomotor Policy with Any-Modality View paper
- [74] Vision-Language-Action Model and Diffusion Policy Switching Enables Dexterous Control of an Anthropomorphic Hand View paper