Topic	Papers / References	Link
Introduction LLMs, GPT3, emergent abilities, in-context learning, scaling laws, text2image models, robotic applications, say-can	Jang, To Understand Language is to Understand Generalization Liang, Stanford CS324 lecture notes Brown et al., Language models are few-shot learners Wei et al., Emergent Abilities of Large Language Models Ahn et al., Do As I Can, Not As I Say: Grounding Language in Robotic Affordances Kaplan et al., Scaling Laws for Neural Language Models	
Deep RL, robotics		
Transformers - architecture, pretraining, fine tuning, application to text and images	Attention Is All You Need (Vaswani et al., 2017)	https://arxiv.org/abs/1706.03762
	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Devlin et al., 2018)	https://arxiv.org/abs/1810.04805
	Improving Language Understanding by Generative Pre-Training (Radford et al., 2018)	https://cdn.openai.com/research-covers/language-unsupervi
	An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (Dosovitskiy et al., 2020)	https://arxiv.org/abs/2010.11929
	LoRA: Low-Rank Adaptation of Large Language Models (Hu et al., 2021)	https://arxiv.org/abs/2106.09685
Diffusion models	Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020)	https://arxiv.org/abs/2006.11239
	High-Resolution Image Synthesis with Latent Diffusion Models (Rombach et al., 2021)	https://arxiv.org/abs/2112.10752
	Denoising Diffusion Implicit Models (DDIM) (Song et al., 2020)	https://arxiv.org/abs/2010.02502
	Flow Matching for Generative Modeling (Lipman et al., 2022)	https://arxiv.org/abs/2210.02747
LLMs	Language Models are Unsupervised Multitask Learners (Radford et al., 2019)	https://cdn.openai.com/better-language-models/language_m
	Language Models are Few-Shot Learners (Brown et al., 2020)	https://arxiv.org/abs/2005.14165
	Scaling Laws for Neural Language Models (Kaplan et al., 2020)	https://arxiv.org/abs/2001.08361
	Chain-of-thought prompting elicits reasoning in large language models (Wei et al., 2023)	https://arxiv.org/abs/2201.11903
	Large Language Models are Zero-Shot Reasoners (Kojima et al., 2023)	https://arxiv.org/abs/2205.11916
RLHF	Deep Reinforcement Learning from Human Preferences (Christiano et al., 2017)	https://arxiv.org/abs/1706.03741
	Learning to Summarize with Human Feedback (Stiennon et al., 2019)	https://arxiv.org/abs/2009.01325
	Training Language Models to Follow Instructions with Human Feedback (Ouyang et al., 2022)	https://arxiv.org/abs/2203.02155
	Direct Preference Optimization: Your Language Model is Secretly a Reward Model (Rafailov et al., 2023)	https://arxiv.org/abs/2305.18290
LLMs for planning	Planning with Large Language Models for Code Generation (Zhang et al., 2023)	https://arxiv.org/abs/2303.05510
	Faster sorting algorithms discovered using deep reinforcement learning (Mankowitz et al., 2024)	https://www.nature.com/articles/s41586-023-06004-9

Topic	Papers / References	Link
	Large Language Models Still Can't Plan (A Benchmark for LLMs on Planning and Reasoning about Change) (Valmeekam et al., 2022)	https://openreview.net/forum?id=wUU-7XTL5XO
	On the Planning Abilities of Large Language Models - A Critical Investigation (Valmeekam et al., 2024)	https://openreview.net/forum?id=X6dEqXIsEW
Reasoning models	Self-Consistency Improves Chain of Thought Reasoning in Language Models (Wang et al., 2022)	https://arxiv.org/abs/2203.11171
	Tree of Thoughts: Deliberate Problem Solving with Large Language Models (Yao et al., 2022)	https://arxiv.org/abs/2305.10601
	DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning (Deepseek-AI et al., 2025)	https://arxiv.org/abs/2501.12948
Imitation learning in robotics	Diffusion policy: Visuomotor policy learning via action diffusion (Chi et al., 2024)	https://journals.sagepub.com/doi/full/10.1177/027836492412
	BC-Z: Zero-Shot Task Generalization with Robotic Imitation Learning (Jang et al., 2022)	https://arxiv.org/abs/2202.02005
	Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware (Zhao et al., 2023)	https://arxiv.org/abs/2304.13705
	Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low-Cost Whole-Body Teleoperation (Fu et al., 2024)	https://arxiv.org/abs/2401.02117
VLMs	CLIP: Learning Transferable Visual Models From Natural Language Supervision (Radford et al., 2021)	https://arxiv.org/abs/2103.00020
	Flamingo: a Visual Language Model for Few-Shot Learning (Alayrac et al., 2022)	https://arxiv.org/abs/2204.14198
	Prismatic VLMs: Investigating the Design Space of Visually-Conditioned Language Models (Karamcheti et al., 2024)	https://openreview.net/forum?id=6FXtu8clyp
VLAs	Open X-Embodiment: Robotic Learning Datasets and RT-X Models (Open X-Embodiment Collaboration, 2023)	https://robotics-transformer-x.github.io/
	Octo: An Open-Source Generalist Robot Policy (Octo Model Team et al., 2024)	https://arxiv.org/abs/2405.12213
	OpenVLA: An Open-Source Vision-Language-Action Model (Kim et al., 2024)	https://arxiv.org/abs/2406.09246
	π0 : A Vision-Language-Action Flow Model for General Robot Control (Black et al., 2024)	https://arxiv.org/abs/2410.24164
LLM agents	Toolformer: Language Models Can Teach Themselves to Use Tools (Schick et al., 2023)	https://arxiv.org/abs/2302.04761
	Do As I Can, Not As I Say: Grounding Language in Robotic Affordances (Ahn et al., 2022)	https://arxiv.org/abs/2204.01691
	HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face (Shen et al., 2023)	https://arxiv.org/abs/2303.17580
	Generative Agents: Interactive Simulacra of Human Behavior (Park et al., 2023)	https://arxiv.org/abs/2304.03442
Video generation	Deep Visual Foresight for Planning Robot Motion (Finn et al., 2017)	https://arxiv.org/abs/1610.00696
	VideoGPT: Video Generation using VQ-VAE and Transformers (Yan et al., 2021)	https://arxiv.org/abs/2104.10157
	Video Diffusion Models (Ho et al., 2022)	https://arxiv.org/abs/2204.03458

Topic	Papers / References	Link
	Imagen Video: High Definition Video Generation with Diffusion Models (Ho et al., 2022)	https://arxiv.org/abs/2210.02303
World models	World Models (Ha et al., 2018)	https://arxiv.org/abs/1803.10122
	Learning Plannable Representations with Causal InfoGAN (Kurutach et al., 2018)	https://arxiv.org/abs/1807.09341
	Mastering Atari with Discrete World Models (Hafner et al., 2022)	https://arxiv.org/abs/2010.02193
	Learning Universal Policies via Text-Guided Video Generation (Du et al., 2023)	https://arxiv.org/abs/2302.00111
	Genie: Generative Interactive Environments (Bruce et al., 2024)	https://arxiv.org/abs/2402.15391
	ReWiND: Language-Guided Rewards Teach Robot Policies without New Demonstrations	
	A Real-to-Sim-to-Real Approach to Robotic Manipulation with VLM-Generated Iterative Keypoint Rewards	
	DINO-WM: World Models on Pre-trained Visual Features enable Zero-shot Planning	
	Reflective Planning: Vision-Language Models for Multi-Stage Long-Horizon Robotic Manipulation (Feng, Yunhai, et al. 2025)	https://reflect-vlm.github.io/
	Robotic Control via Embodied Chain-of-Thought Reasoning	
	Policy Adaptation via Language Optimization: Decomposing Task	https://arxiv.org/abs/2408.16228
	HAMSTER: Hierarchical Action Models For Open-World Robot Manipulation	https://arxiv.org/abs/2502.05485
	REGENT: A Retrieval-Augmented Generalist Agent That Can Act In-Context in New Environments	https://arxiv.org/abs/2412.04759
	Modeling the Real World with High-Density Visual Particle Dynamics	https://arxiv.org/pdf/2406.19800
	GR00T N1: An Open Foundation Model for Generalist Humanoid	https://arxiv.org/abs/2503.14734
	UniVLA: Learning to Act Anywhere with Task-centric Latent Actions	https://www.arxiv.org/abs/2505.06111
	Are transformers truly foundational for robotics?	
	Autonomous Improvement of Instruction Following Skills via Foundation Models	https://arxiv.org/abs/2407.20635
	Discrete Diffusion Modeling by Estimating the Ratios of the Data Distribution	https://arxiv.org/abs/2310.16834
	Interpreting Emergent Planning in Model-Free Reinforcement Learning	
	PhysTwin: Physics-Informed Reconstruction and Simulation of Deformable Objects from Videos	https://arxiv.org/abs/2503.17973
VLMs, Reasoning models	Commonsense Reasoning for Legged Robot Adaptation with Vision-Language Models	https://arxiv.org/abs/2407.02666
	LLARVA: Vision-Action Instruction Tuning Enhances Robot Learning	

Topic	Papers / References	Link
	Scaling up Test-Time Compute with Latent Reasoning: A Recurrent Depth Approach	https://arxiv.org/abs/2502.05171
In-Context Learning Enables Robot Action Prediction in LLMs	In-Context Learning Enables Robot Action Prediction in LLMs	https://arxiv.org/pdf/2410.12782
DREAMGEN: Unlocking Generalization in Robot Learning through Neural Trajectories	https://arxiv.org/pdf/2505.12705	https://arxiv.org/pdf/2505.12705
FMs in TAMP	Meta-Optimization and Program Search using Language Models for Task and Motion Planning	
https://2024.corl.org/program/papers		

Paper Title	Summary	Keywords	Link
In-Context Learning Enables Robot Action Prediction in LLMs	Introduces RoboPrompt, a framework enabling large language models to predict robot actions through in-context learning without additional training.	LLMs, LLMs for planning, Reasoning models	https://arxiv.org/abs/2410.12782
Commonsense Reasoning for Legged Robot Adaptation with Vision-Language Models	Proposes VLM-Predictive Control, combining in-context adaptation and multi-skill planning to enhance legged robots' adaptability using vision-language models.	VLMs, Reasoning models	https://arxiv.org/abs/2407.02666
ReWiND: Language-Guided Rewards Teach Robot Policies without New Demonstrations	Presents ReWiND, a framework that learns robot manipulation tasks from language instructions without per-task demonstrations by leveraging a language-conditioned reward model.	LLMs, RLHF, Imitation learning	https://arxiv.org/abs/2505.10911
The Pitfalls of Imitation Learning when Actions are Continuous	Analyzes the limitations of imitation learning in continuous action spaces, highlighting the necessity for complex policy parameterizations to avoid performance degradation.	Imitation learning, Theory	https://arxiv.org/abs/2503.09722
Are transformers truly foundational for robotics?	Examines the role of transformers in robotics, questioning their foundational status and exploring alternative architectures for robotic applications.	Transformers, Theory	https://rdcu.be/emLxY
UniVLA: Learning to Act Anywhere with Task-centric Latent Actions	Introduces UniVLA, a framework that learns cross-embodiment vision-language-action policies using task-centric latent actions derived from videos.	VLAs, VLMs, LLM agents	https://www.arxiv.org/abs/2505.06111
Meta-Optimization and Program Search using Language Models for Task and Motion Planning	Explores the use of language models for meta-optimization and program search to enhance task and motion planning in robotics.	LLMs, LLMs for planning, Reasoning models	https://www.arxiv.org/abs/2505.03725
GR00T N1: An Open Foundation Model for Generalist Humanoid Robots	Presents GR00T N1, an open-source foundation model designed to accelerate the development of generalist humanoid robots through a dual-system architecture.	LLMs, World models, LLM agents	https://arxiv.org/abs/2503.14734
Interpreting Emergent Planning in Model-Free Reinforcement Learning	Investigates how planning behaviors can emerge in model-free reinforcement learning agents without explicit planning modules.	Reasoning models, Theory	https://arxiv.org/abs/2504.01871
Is a Good Foundation Necessary for Efficient Reinforcement Learning? The Computational Role of the Base Model in Exploration	Analyzes the impact of foundational models on the efficiency of reinforcement learning, particularly in exploration strategies.	World models, Theory	https://arxiv.org/abs/2503.07453
PhysTwin: Physics-Informed Reconstruction and Simulation of Deformable Objects from Videos	Introduces PhysTwin, a method for reconstructing and simulating deformable objects from videos using physics-informed models.	Video generation, World models	https://arxiv.org/abs/2503.17973
Reflective Planning: Vision-Language Models for Multi-Stage Long-Horizon Robotic Manipulation	Proposes a reflective planning approach using vision-language models to handle multi-stage, long-horizon robotic manipulation tasks.	VLMs, LLMs for planning	https://arxiv.org/abs/2502.16707
HAMSTER: Hierarchical Action Models For Open-World Robot Manipulation	Introduces HAMSTER, a hierarchical action model framework designed for open-world robot manipulation tasks.	Reasoning models, LLM agents	https://arxiv.org/abs/2502.05485
A Real-to-Sim-to-Real Approach to Robotic Manipulation with VLM-Generated Iterative Keypoint Rewards	Presents a real-to-sim-to-real methodology for robotic manipulation using vision-language model-generated iterative keypoint rewards.	VLMs, Imitation learning	https://arxiv.org/abs/2502.08643
DINO-WM: World Models on Pre-trained Visual Features enable Zero-shot Planning	Introduces DINO-WM, leveraging pre-trained visual features to enable zero-shot planning through world models.	World models, LLMs for planning	https://arxiv.org/abs/2411.04983
Scaling up Test-Time Compute with Latent Reasoning: A Recurrent Depth Approach	Explores scaling test-time computation using latent reasoning through a recurrent depth approach.	Reasoning models, Theory	https://arxiv.org/abs/2502.05171

REGENT: A Retrieval-Augmented Generalist Agent That Can Act In-Context in New Environments	Presents REGENT, a retrieval-augmented agent capable of in-context action in novel environments.	LLM agents, LLMs for planning	https://arxiv.org/abs/2412.04759
Discrete Diffusion Modeling by Estimating the Ratios of the Data Distribution	Proposes a discrete diffusion modeling technique by estimating data distribution ratios.	Diffusion models, Theory	https://arxiv.org/abs/2310.16834
DreamGen: Unlocking Generalization in Robot Learning through Neural Trajectories	Introduces DreamGen, enhancing robot learning generalization via neural trajectory generation.	Imitation learning, World models	https://arxiv.org/abs/2505.12705
Modeling the Real World with High-Density Visual Particle Dynamics	Explores modeling real-world environments using high-density visual particle dynamics.	World models, Video generation	https://arxiv.org/abs/2406.19800
LLARVA: Vision-Action Instruction Tuning Enhances Robot Learning	Presents LLARVA, a method for enhancing robot learning through vision-action instruction tuning.	VLMs, Imitation learning	https://arxiv.org/abs/2406.11815
Mobility VLA: Multimodal Instruction Navigation with Long-Context VLMs and Topological Graphs	Introduces Mobility VLA, combining long-context vision-language models with topological graphs for multimodal instruction navigation.	VLAs, VLMs	https://arxiv.org/abs/2407.07775
Robotic Control via Embodied Chain-of-Thought Reasoning	Explores robotic control through embodied chain-of-thought reasoning processes.	Reasoning models, LLM agents	https://arxiv.org/abs/2407.08693
Policy Adaptation via Language Optimization: Decomposing Tasks for Few-Shot Imitation	Proposes a method for policy adaptation by decomposing tasks using language optimization for few-shot imitation learning.	LLMs, Imitation learning	https://arxiv.org/abs/2408.16228
Autonomous Improvement of Instruction Following Skills via Foundation Models	Proposes a method where robots autonomously refine their instruction-following abilities using feedback generated via large foundation models.	LLMs, LLM agents, Imitation learning	https://arxiv.org/abs/2407.20635