Survey of π_0, π_0-FAST, and $\pi_{0.5}$: Vision-Language-Action Models in the Physical AI Framework

Seonghyun Kim
Digital Convergence Research Laboratory
Electronics and Telecommunications Research Institute
Daejeon, Korea
kim-sh@etri.re.kr

Samyeul Noh
Digital Convergence Research Laboratory
Electronics and Telecommunications Research Institute
Daejeon, Korea
samuel@etri.re.kr

Ingook Jang
Digital Convergence Research Laboratory
Electronics and Telecommunications Research Institute
Daejeon, Korea
ingook@etri.re.kr

***Abstract*—This paper provides a survey and comparative analysis of three key VLA-based robot policies from Physical Intelligence: π_0, π_0-FAST, and $\pi_{0.5}$. π_0 introduced a foundation model for robot manipulation that couples a pre-trained vision-language backbone with a diffusion-based action policy. Building on π_0, the π_0-FAST model incorporated a novel Frequency-space Action Sequence Tokenization (FAST) scheme, enabling up to five times faster training while matching the performance of diffusion models. The latest model, $\pi_{0.5}$, extends π_0's architecture with co-training on diverse data sources and a hierarchical policy structure to achieve real-world generalization demonstrating, for the first time, successful robotic manipulation in entirely new unseen environments.**

Keywords—Vision-language-action model, Physical AI, Robot foundation model.

I. Introduction

Recent advances in large language models and vision–language models have strengthened instruction following, long-horizon reasoning, and grounded visual perception [1]-[3]. Building on this progress, artificial intelligence models for general-purpose robotic control have emerged as a promising approach [4][5]. By leveraging large-scale pre-trained vision–language representations and coupling them with robot-specific action policies, such systems aim to enable robots to understand natural instructions and perform complex physical tasks in diverse environments.

A representative example of this paradigm is the π_0 model, which demonstrated that a unified policy could perform a wide variety of dexterous tasks from folding laundry to organizing drawers across different robot embodiments using a flow-matching diffusion-based action generator [6]. Building on π_0, π_0-FAST model introduced a discrete action representation using Frequency-space Action Sequence Tokenization (FAST), enabling approximately five times faster training without compromising performance [7]. More recently, $\pi_{0.5}$ extended the architecture through heterogeneous multi-modal co-training and a hierarchical control strategy, achieving real-world generalization to novel environments unseen during training [8].

II. Vision-Language-Action Policies

VLA models define a unified policy framework that jointly maps high-dimensional visual observations and natural language goals to executable low-level actions. This formulation enables end-to-end learning of instruction-conditioned control across diverse robotic tasks.

In general, a VLA model is formulated to learn a policy as

$$\pi(a_{1:t}|o_{1:t}, x) \tag{1}$$

that generates a sequence of actions $a_{1:t}$ conditioned on a stream of observations $o_{1:t}$ and an input instruction x, where $a_{1:t}$ are low-level robot actions, $o_{1:t}$ are raw observations, and x is a high-level task specification typically given as natural language.

In practice, VLA policies are implemented using causal sequence models as transformers that encode the instruction x and the visual context $o_{1:T}$ into a shared latent representation. The action sequence $a_{1:T}$ is then predicted either:
• Autoregressively, as in tokenized models like π_0-FAST:

$$\pi(a_t|a_{1:t-1}o_{1:t}, x) \tag{2}$$

• Via flow-matching or diffusion**,** as in continuous policies like π_0 and $\pi_{0.5}$:

$$\pi(a_{1:t}|z_{1:t}) \tag{3}$$

where $z_t = f(o_t, x)$ is a latent representation.

This formulation allows VLA policies to combine temporal perception, semantic understanding, and physical control within a single architecture. Depending on the model, the action head may operate in discrete token space or continuous control space.

III. Training Dataset

The training data underlying the Pi model series significantly affects the scalability, generalization, and task versatility of each policy. All three models, π_0, π_0-FAST, and $\pi_{0.5}$, were trained on large-scale robot demonstration datasets collected across multiple embodiments, environments, and task categories [6]-[8].

The π_0 model was trained on approximately 10,000 hours of robot data, drawn from eight different robot embodiments and over 60 distinct tasks, including manipulation, cleaning, and assembly activities [6]. These demonstrations came from a mix of scripted behaviors and human teleoperation, enabling the model to generalize across physical forms and skill types.

The π_0-FAST model retained a similar dataset scope, but transformed the demonstration trajectories using the FAST tokenization scheme, which compresses continuous action sequences into discrete representations [7]. This formulation enabled efficient next-token prediction training using standard autoregressive transformers.

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. [25ZR1100, A Study of Hyper-Connected Thinking Internet Technology by autonomous connecting, controlling, and evolving ways].

Table I. Comparison of π_0, π_0-FAST, and $\pi_{0.5}$ models.

Architecture	Training Strategy	Key Results	Key Contributions
π_0: Pre-trained VLM backbone with flow-matching action policy [6]	Multi-task training on ~10k hours from 8 robots (various embodiments), web data [6]	Generalist policy performs diverse dexterous tasks (e.g., folding laundry, assembling boxes) via prompting or fine-tuning [6]	Introduced VLA foundation model concept; demonstrated cross-embodiment skill transfer [6]
π_0-FAST: Autoregressive Transformer policy with FAST action tokenization [7]	Next-token prediction on DCT-compressed action sequences; scaled to 10k hour dataset [7]	Matched performance of diffusion-based policy while training is 5× faster on dexterous tasks [7]	Proposed FAST tokenization enabled efficient large-scale VLA training [7]
$\pi_{0.5}$: Hybrid hierarchical policy (high-level subtask prediction, continuous low-level controller) [8]	Co-training on heterogeneous data (multi-robot, simulation, web, human instructions); fine-tuned on ~400h real home robot data [8]	A VLA to generalize to unseen environments (cleaning new homes) with ~94% success; approaches in-domain model performance [8]	Demonstrated open-world generalization via multi-modal co-training; integrated planning and control in one model [8]

The $\pi_{0.5}$ model extended this foundation by introducing a co-training strategy across a heterogeneous mix of robot platforms and real-world environments [8]. Although the paper does not enumerate individual data types, it emphasizes that $\pi_{0.5}$ was trained on a broader distribution including long-horizon tasks in unseen home environments and that its learning process integrated diverse sources of behavior supervision. Fine-tuning was performed on real-world data collected in domestic scenes, enabling $\pi_{0.5}$ to generalize more effectively to natural household contexts.

IV. Comparative Overview of the PI Model Series

The Pi model series represents a structured progression in the development of VLA policies, targeting improvements in efficiency and generalization. Table I summarizes the architectural differences, training methods, empirical performance, and core innovations introduced in each model: π_0, π_0-FAST, and $\pi_{0.5}$.

The π_0 model established a foundation for generalist robot control by integrating a pre-trained vision-language backbone with a continuous action policy trained via flow matching. It demonstrated that large-scale multi-task behavior cloning across various robot embodiments could yield a unified policy capable of handling manipulation tasks via language prompts or fine-tuned commands.

To address the computational demands of diffusion-based learning, π_0-FAST introduced a discrete action representation using frequency-space action sequence tokenization. This allowed the model to shift from continuous action regression to next-token prediction, enabling significantly faster training while preserving the high success rates of the π_0 baseline.

Expanding beyond in-distribution performance, $\pi_{0.5}$ employed a hierarchical architecture with co-training on diverse data sources—ranging from robot demonstrations and web-scale vision-language. By decoupling high-level reasoning from low-level control, and insulating pre-trained knowledge during motor learning, $\pi_{0.5}$ achieved strong generalization to novel environments such as unseen homes, marking a new milestone in the Pi series.

Overall, this series reflects a systematic enhancement of both model capability and learning efficiency, evolving from strong in-domain competence to robust real-world applicability.

V. Perfromance Comparisons

This section presents a performance comparison of the π_0, π_0-FAST, and $\pi_{0.5}$ models, focusing on their training efficiency and final task success rates across household manipulation benchmarks.

A. Training Efficiency

Figure 1 represents the relative training time required for each model to reach a remarkable success rate on manipulation tasks [7]. This comparison highlights the impact of different policy architectures and training strategies on sample efficiency. As shown in Fig. 1, π_0-FAST reaches a remarkable success rate in roughly 20% of the training time required by π_0, owing to the FAST tokenization scheme. By representing actions as discrete tokens, π_0-FAST can leverage efficient sequence learning. $\pi_{0.5}$ also benefits from FAST during its pre-training phase achieving comparable sample efficiency to π_0-FAST although its post-training phase employs the slower flow-matching controller. The $\pi_{0.5}$ leverages the efficiency of FAST during its pre-training phase, which enables it to achieve sample efficiency comparable to π_0-FAST. However, a key distinction lies in the post-training stage, where $\pi_{0.5}$ employs a slower flow-matching controller [8]. Consequently, it is inferred that the overall training time for $\pi_{0.5}$ would be longer than that for π_0-FAST, which utilizes the FAST tokenization method throughout its entire training pipeline.

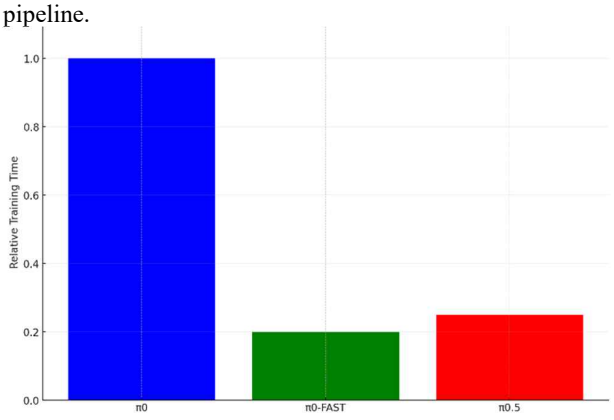

Fig. 1. Relative training time to reach 0.8 success rate for π_0, π_0-FAST, and $\pi_{0.5}$. π_0-FAST trains about 5× faster than π_0 to achieve the same performance level, while $\pi_{0.5}$ attains similar efficiency.

B. Task Performance

Figure 2 represents a comparison of the task success rates for different models, evaluating their ability to successfully execute language-guided tasks within two distinct kitchen environments [8]. The task success rate metric evaluates a robot's ability to successfully place the target object in the specified location. As shown in Fig. 2, the $\pi_{0.5}$ achieved a slightly higher success rate than the π_0-FAST and a significantly higher rate than the π_0. This result highlights the critical importance of discrete token training for robust language following and task completion. Furthermore, the experiments with out-of-distribution objects, which were not included in the training set, provide a deeper insight into the models' generalization capabilities. The ability to successfully handle novel household items demonstrates that the approach can generalize to previously unseen objects, a key requirement for real-world robotic applications. Overall, the evolution from π_0 to π_0-FAST to $\pi_{0.5}$ shows a trend of maintaining or improving task success while greatly reducing training time and expanding generalization.

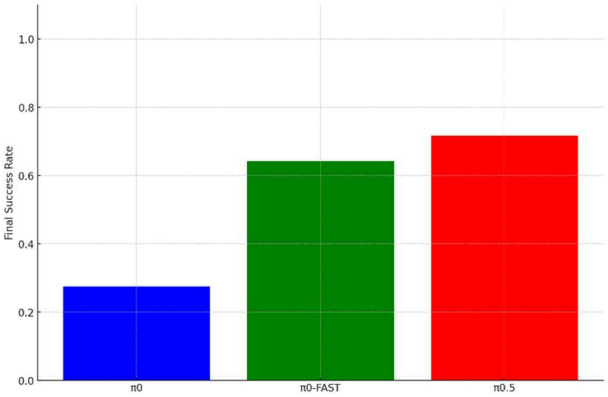

Fig. 2. Comparisons of task success rates for π_0, π_0-FAST, and $\pi_{0.5}$. $\pi_{0.5}$ achieves the highest task success rate, with π_0-FAST outperforming π_0.

VI. Conclusion and Open Challenges

This paper presented a comparative survey of the π_0, π_0-FAST, and $\pi_{0.5}$ models within the Physical Intelligence framework. Through architectural and empirical analysis, we demonstrated how these models collectively advance the field of general-purpose robot control.

The π_0 model introduced the concept of a vision-language-action foundation policy, capable of performing a wide range of dexterous tasks across different embodiments via prompting or fine-tuning. π_0-FAST improved upon this baseline with a novel discrete tokenization of continuous actions, enabling up to fivefold training efficiency without compromising success rates. $\pi_{0.5}$ further extended the framework by introducing hierarchical planning and co-training across heterogeneous robot and environment data, achieving strong performance even in previously unseen environments.

Despite these achievements, several open challenges remain. First, $\pi_{0.5}$—while exhibiting robust generalization—still struggles with certain physical and perceptual limitations, such as unfamiliar hardware (e.g., novel drawer handles), occlusions in visual input, and inconsistent high-level subtask inference. Second, current co-training approaches, though effective, leave room for improvement in balancing diverse data modalities and increasing the complexity of learned behaviors.

Future work may focus on:

- Improving performance in partially observable environments through better memory and long-term context modeling.
- Enhancing prompt comprehension by incorporating more sophisticated and richly annotated instruction data.
- Expanding the training regime with larger-scale and more varied datasets, including real-world household interactions.
- Exploring new supervision modalities such as verbal instruction, enabling more natural human-robot interaction.
- Developing more adaptive planning architectures capable of operating across highly dynamic and ambiguous real-world scenarios.

In summary, the π-series models have made significant progress toward generalist robot intelligence, offering scalable architectures, efficient learning methods, and real-world applicability. Addressing the remaining limitations will be essential in achieving the next milestone—robots that can robustly learn, reason, and act across truly open-ended environments.

References

[1] S. Yeo, Y.-S. Ma, S. C. Kim, H. Jun, and T. Kim, "Framework for evaluating code generation ability of large language models," *ETRI Journal*, vol. 46, no. 1, pp. 106–117, Feb. 2024.

[2] J. Roh, M. Kim, and K. Bae, "Towards a small language model powered chain-of-reasoning for open-domain question answering," *ETRI Journal*, vol. 46, no. 1, pp. 11–21, Feb. 2024.

[3] S. Jeon, J. Lee, D. Yeo, Y.-J. Lee, and S. J. Kim, "Multimodal audiovisual speech recognition architecture using a three-feature multi-fusion method for noise-robust systems," *ETRI Journal*, vol. 46, no. 1, pp. 22–34, Feb. 2024.

[4] K.-I. Na and B. Park, "Real-time 3D multi-pedestrian detection and tracking using 3D LiDAR point cloud for mobile robot," *ETRI Journal*, vol. 45, no. 5, pp. 836–846, Oct. 2023.

[5] T. S. Nguyen, H. N. Cao, and M. T. Pham, "Semantic potential field for mobile robot navigation using grid maps," *ETRI Journal*, vol. 47, no. 3, pp. 422–432, Jun. 2025.

[6] K. Black et al., "π_0: A Vision-Language-Action Flow Model for General Robot Control," arXiv preprint arXiv:2410.24164, 2024.

[7] K. Pertsch et al., "FAST: Efficient Action Tokenization for Vision-Language-Action Models," arXiv preprint arXiv:2501.09747, 2025.

[8] K. Black et al., "$\pi_{0.5}$: A Vision-Language-Action Model with Open-World Generalization," arXiv preprint arXiv:2504.16054, 2025.