RetoVLA: Reusing Register Tokens for Spatial Reasoning in Vision-Language-Action Models

Jiyeon Koo[1,†], Taewan Cho[1,†], Hyunjoon Kang[1], Eunseom Pyo[1], Tae Gyun Oh[1], Taeryang Kim[1],
and Andrew Jaeyong Choi[1,*]

Abstract— Recent Vision-Language-Action (VLA) models demonstrate remarkable generalization in robotics but are restricted by their substantial size and computational cost, limiting real-world deployment. However, conventional lightweighting methods often sacrifice critical capabilities, particularly spatial reasoning. This creates a trade-off between efficiency and performance. To address this challenge, our work reuses Register Tokens, which were introduced for artifact removal in Vision Transformers but subsequently discarded. We suppose that these tokens contain essential spatial information and propose RetoVLA, a novel architecture that reuses them directly by injecting them into the Action Expert.

RetoVLA maintains a lightweight structure while leveraging this repurposed spatial context to enhance reasoning. We demonstrate RetoVLA's effectiveness through a series of comprehensive experiments. On our custom-built 7-DOF robot arm, the model achieves a 17.1%p absolute improvement in success rates for complex manipulation tasks. Our results confirm that reusing Register Tokens directly enhances spatial reasoning, demonstrating that what was previously discarded as an artifact is in fact a valuable, unexplored resource for robotic intelligence. A video demonstration is available at:
https://youtu.be/2CseBR-snZg

I. INTRODUCTION

By integrating vast web-scale knowledge into robotic control, Vision-Language-Action (VLA) models such as RT-2 [1] and OpenVLA [2] have demonstrated remarkable generalization in understanding complex language instructions. However, this success relies on massive, billion-parameter models, which require substantial computational costs. This fundamental issue presents a significant challenge to their practical deployment on real-world robotic platforms with limited on-board computing resources.

Previous efforts to address this efficiency problem have primarily concentrated on physically reducing the model size, as seen in approaches like SmolVLA [4]. However, this inevitably leads to a trade-off between performance and efficiency. This requires a compromise on the extensive representational power of VLMs, particularly their ability to comprehend complex spatial relationships and long-term contexts. Simply discarding information risks compromising fundamental aspects of robotic intelligence in exchange for computational gains.

[†]These authors contributed equally to this work.

[1]All authors are with the School of Computing, Gachon University, 1342 Seongnam-daero, Sujeong-gu, Seongnam 13120, Republic of Korea.

[*]Andrew Jaeyong Choi is the corresponding author. Email: andrewjchoi@gachon.ac.kr

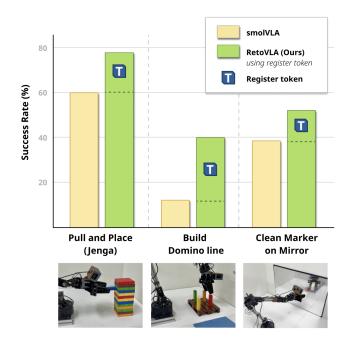Student emails: {halo1225, taewan2002, kanghj1537, vydms123, luili0307, ktr1110}@gachon.ac.kr

Fig. 1. Comparison of RetoVLA and the SmolVLA baseline on challenging real-world tasks. (Top) Our model, RetoVLA (green), significantly outperforms the baseline (yellow). (Bottom) This performance gain comes from reusing the Register Token [3] (indicated by the 'T' icon) to inject global spatial context into the Action Expert.

This paper examines the underlying issue, seeking a solution not in the 'deletion' of information, but in its 'active reuse'. Our starting point is the work of [3], which revealed that modern large-scale Vision Transformers (ViT) [5] like DINOv2 [6] inherently produce artifacts known as high-norm outlier tokens during training. These tokens typically emerge from image patches relatively low informational content, such as blank walls or the sky. The model repurposes the representational space of these patches to serve as a temporary scratchpad for storing and processing internal global information. While this is a natural learning mechanism, it has been shown to deteriorate the local information of the corresponding patch tokens, significantly impairing performance on dense prediction tasks.

The proposed solution to this issue, the Register Token [3], provides an explicit 'scratchpad' for the model, thereby preventing the misuse of image patch tokens and refining the attention maps. However, after absorbing these artifacts, the Register Tokens [3] have been systematically discarded in downstream tasks. This leads us to a critical question that
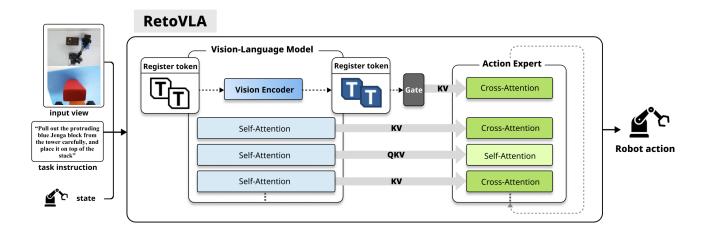
Fig. 2. The RetoVLA architecture. Our key innovation is the Spatial Context Injection path (dashed arrow), which enhances a standard VLM-based policy. Register Tokens [3], summarizing global scene information from the Vision Encoder, are passed through a learnable gate. They are then injected as Key-Value pairs into the final cross-attention layer of the Action Expert, allowing it to reason over both high-level semantic features and global spatial context simultaneously.

diverges from previous work: What is the true nature of this 'global information,' and is it merely noise to be discarded?

In the context of robotic manipulation, we propose that this information constitutes the essential Spatial Context of a scene—encompassing the overall layout, the 3D relationships between objects, and the structure of the workspace. For a robot, this information is not a disposable asset; it is critical.

Building on this hypothesis, we propose RetoVLA (Reusing Register Tokens [3] VLA), which achieves both efficiency and high performance through the active reuse of information. As summarized in Figure 1, our core contributions are as follows: (1) Redefining the Role of Register Tokens [3] for Spatial Context Injection: A novel VLA architecture, detailed in Figure 2, that redefines the register token [3] from a passive 'purifier' used for artifact removal to an active 'spatial context provider'. We have designed a novel module to directly inject these tokens as Key-Value pairs into the Action Expert's attention mechanism, allowing it to leverage the global spatial context of the entire scene until the final phase of action generation. (2) Analyzing the Synergy with Efficient VLAs: We analyze how register token [3] injection compensates for the information loss that occurs in efficient models like SmolVLA [4], which reduce the depth of the VLM. Our approach proves to be an effective pathway for maintaining a high level of spatial reasoning with significantly lower computational overhead. (3) Experimental Validation: Through rigorous experiments on the LIBERO simulation benchmark and with the custom-built 7-DOF robot arm we developed for this research, we demonstrate that RetoVLA significantly outperforms the baseline model, particularly on long-horizon tasks that require a complex understanding of 3D spatial structures or multi-step sequential manipulation. In our real-robot experiments, we achieved an increase in the average success rate from 50.3% to 67.4%, a 17.1%p absolute improvement.

This research reconstructs conventional design principle by re-evaluating internal information flow, enabling a new class of low-cost, high-performance robotic intelligence.

II. RELATED WORK

A. Vision-Language-Action (VLA) Models

The application of large-scale transformer architectures to robotics, pioneered by models like RT-1 [7], marked a significant advancement in generalist policies. This paradigm was further advanced by VLA models, which extend pre-trained Vision-Language Models (VLMs) to end-to-end control. A significant advance was demonstrated in RT-2 [1], which established this concept by showing that web-scale knowledge could be directly transferred to manipulation by treating robot actions as text tokens during co-fine-tuning. Building on these foundations, a series of powerful models including the open-source OpenVLA [2] and $\pi 0$ [8] have emerged, showcasing remarkable generalization capabilities. However, this success relies on massive, often billion-parameter architectures. This reliance on scale results in substantial computational demands and slow inference, posing a critical bottleneck for their practical deployment. While recent research continues to push the boundaries of VLA capabilities—such as improving spatial precision [9], enhancing visual robustness [10], and enabling open-world manipulation with minimal robot data [11], [12]—the fundamental challenge of computational efficiency remains a primary concern.

B. Lightweight Vision-Language-Action Models

To address the scalability challenges of large-scale VLAs, a significant research effort has focused on developing lightweight models. These approaches, such as SmolVLA [4], typically achieve efficiency by combining smaller VLM backbones with techniques like layer skipping, visual token reduction, and parameter-efficient adapters (e.g., LoRA [13]). Other works like TinyVLA [14] address slow inference by replacing autoregressive action generation with faster,
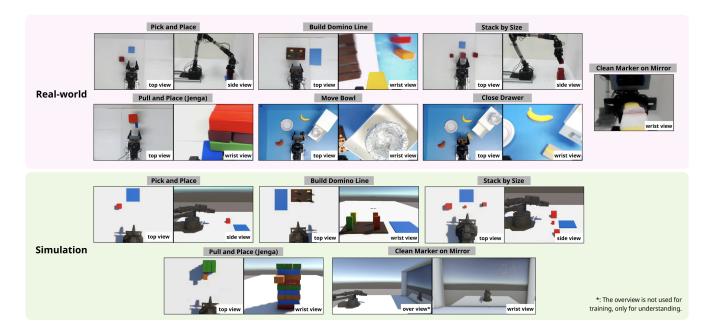
Fig. 3. Overview of the experimental setups for real-world and simulation tasks. (Top) The real-world setup employs our robot arm to perform seven diverse manipulation tasks. Two of these tasks, 'Move Bowl' and 'Close Drawer', are directly adapted from the LIBERO benchmark. (Bottom) In addition to the two tasks adapted from LIBERO, the remaining five tasks are implemented in the simulation to be consistent with the real-world setup and perform manipulation tasks. The simulation is also implemented using Unity and the MuJoCo plugin.

diffusion-based decoders, building on the success of diffusion policies for visuomotor control [15].

While these architectural innovations greatly improve computational efficiency, they often introduce a trade-off between efficiency and performance. A common limitation noted in these works is that the process of lightweighting can lead to a degradation in performance on complex tasks, particularly by sacrificing the understanding of complex spatial relationships and global context awareness inherent in larger models. Diverging from this paradigm of information reduction, our work introduces a new direction. We propose reusing Register Tokens [3] —previously discarded artifacts from Vision Transformers [5] —to restore this lost spatial reasoning capability. By actively reusing this latent information, RetoVLA enhances the spatial intelligence of a lightweight VLA framework without compromising its efficiency, thereby directly addressing a key limitation of prior lightweighting efforts.

C. Artifacts and Register Tokens in Vision Transformers

Recent studies on large-scale ViTs [5] identified an artifact where the model repurposes tokens from low-information patches as an internal 'scratchpad' to process global scene information, which corrupts local features and degrades performance on dense prediction tasks [16]–[18]. The established solution is the introduction of learnable Register Tokens [3], which provide a dedicated scratchpad to prevent this misuse. However, this solution has been narrowly framed, viewing Register Tokens merely as a passive 'purifier' whose contents are systematically discarded after absorbing artifacts. Our work challenges this convention, supposing that this discarded information is not mere noise but a valuable 'compressed spatial summary' crucial for robotic manipulation. Therefore, RetoVLA is the first approach to rediscover the value of this information and leverage it directly for robotic action generation.

III. METHOD

This section details the architecture of RetoVLA (Fig. 2), the first model purpose-built to operationalize our core hypothesis: that the Register Tokens, previously discarded as passive 'purifiers' [3], can be transformed into active 'spatial context providers'. Our entire methodology is designed around this central principle of information reuse. This principle stands in distinct contrast to conventional lightweighting strategies, which achieve efficiency primarily through information reduction. We achieve our goal by injecting this repurposed spatial context directly into the Action Expert, providing it with a richer, dual-stream of information for generating spatially intelligent actions.

A. Architecture Overview and Information Flow

RetoVLA is built upon a standard VLM backbone and Action Expert structure, a common paradigm in recent VLA research. Our key modification is a redesigned information flow. While conventional models pass a single stream of semantic features from the VLM to the Action Expert (typically a Transformer decoder), RetoVLA establishes a dual-stream pathway: (1) high-level semantic features and (2) Register Token [3] features, which provide a compressed summary of the scene's global spatial information. These two streams are then fused within the Action Expert's cross-attention layers to determine the final, spatially-aware action.

TABLE I

TASK DESCRIPTIONS AND CAMERA CONFIGURATIONS FOR REAL-WORLD EXPERIMENTS.

Task Name	Task Instruction	Camera View
Pick and Place	*Pick up the red cube and place it on the pallet*	Top, Side
Stack by Size	*Pick up the red cubes and stack them on the fixed blue platform in order from largest to smallest, placing the biggest one first and the smallest one last*	Top, Side
Pull and Place (Jenga)	*Pull out the protruding blue Jenga block from the tower carefully, and place it on top of the stack*	Top, Wrist
Build Domino Line*	*Pick up the red, orange, yellow, and green blocks in that order, and place them upright in a straight line like dominoes*	Top, Wrist
Close Drawer	*Close the top drawer of the cabinet*	Top, Wrist
Move Bowl	*Pick up the silver bowl on the box and place it on the plate*	Top, Wrist
Clean Marker on Mirror**	*Pick up the eraser using mirror reflection, erase drawing from mirror*	Wrist

* We design a long-horizon manipulation task comprising an average of 900 frames per episode, which is 2–3 times longer than typical real-world tasks.
** To minimize the use of visual inputs, we restrict the agent to a single wrist-mounted camera and introduce a mirror as an auxiliary visual modality to compensate for the limited viewpoint.

B. Depth-Adaptive VLM Backbone

To establish a computationally efficient yet powerful foundation, RetoVLA strategically employs a "shallower" VLM backbone, utilizing only the initial N=L/2 layers of a pre-trained VLM. This is a deliberate design choice informed by the findings of SmolVLA [4], which demonstrated this configuration provides an optimal trade-off between semantic feature extraction and inference speed. By adopting this efficient backbone, we can clearly isolate and analyze the performance gains derived directly from our novel spatial context injection mechanism.

C. Spatial Context Injection via Register Tokens

The core technical contribution of RetoVLA is the mechanism for injecting previously discarded Register Tokens [3] into the Action Expert. This process consists of three main steps.

a) Register Token Generation: The image patch embeddings from the VLM's Vision Encoder, denoted as $\mathbf{P} \in \mathbb{R}^{B \times N \times D_{\text{vlm}}}$, are fed into a Spatial Context Aggregator module. We implement this with a standard Multi-head Attention block [19], where a set of learnable initial register tokens [3], $\mathbf{R}_{\text{init}} \in \mathbb{R}^{K \times D_{\text{vlm}}}$, act as the Query, and the image patches serve as Keys and Values. This allows the register tokens [3] to effectively "query" the entire visual scene and summarize the most salient global information into a set of scene-dependent Register Tokens [3], $\mathbf{R}_{\text{scene}}$:

$$\mathbf{R}_{\text{scene}} = \text{Attention}(\mathbf{Q} = \mathbf{R}_{\text{init}}, \mathbf{K} = \mathbf{P}, \mathbf{V} = \mathbf{P}) \quad (1)$$

where K is the number of register tokens [3].

b) Injection into the Action Expert: The generated $\mathbf{R}_{\text{scene}}$ is projected to match the Action Expert's dimension and then transformed into Key ($\mathbf{K}_{\text{reg}}$) and Value ($\mathbf{V}_{\text{reg}}$) pairs. We inject this information by concatenating it with the standard semantic Key-Value pairs ($\mathbf{K}_{\text{vlm}}, \mathbf{V}_{\text{vlm}}$) from the VLM in the final cross-attention layer. This non-intrusive approach allows the Action Expert's queries to flexibly attend to either high-level semantic features or the global spatial summary, depending on the task's requirements.

$$\mathbf{K}_{\text{final}} = \text{Concat}(\mathbf{K}_{\text{vlm}}, \sigma(g) \cdot \mathbf{K}_{\text{reg}}) \quad (2)$$
$$\mathbf{V}_{\text{final}} = \text{Concat}(\mathbf{V}_{\text{vlm}}, \sigma(g) \cdot \mathbf{V}_{\text{reg}}) \quad (3)$$

c) Gating Mechanism: We observed that not all tasks benefit equally from global context; for tasks demanding extreme local precision, this information could act as a distraction. To address this, we introduce a learnable scalar parameter, g, as a dynamic gate. Its value is passed through a sigmoid function, $\sigma(\cdot)$, to scale the impact of the Register Tokens [3]. This gating mechanism allows the model to adaptively modulate the influence of global spatial context, learning to amplify it for tasks requiring a holistic scene understanding and suppress it for those demanding local precision.

D. Training Objective: Conditional Flow Matching

We train RetoVLA using a conditional flow matching objective, a powerful paradigm for generative modeling [20]. The goal is to learn a vector field $\mathbf{v}_t$ that transforms a simple noise distribution into a distribution of desired robot actions, conditioned on the visual and language context.

Let $\mathbf{a}_0$ be the ground-truth action sequence and $\mathbf{a}_1 \sim \mathcal{N}(0, \mathbf{I})$ be a sequence of pure noise. We define a probability flow that interpolates between them as $\mathbf{a}_t = (1-t)\mathbf{a}_0 + t\mathbf{a}_1$ for a time variable $t \in [0, 1]$. The corresponding vector field is $\mathbf{u}_t = \mathbf{a}_1 - \mathbf{a}_0$.

The model, parameterized by θ, learns to predict this vector field, $\mathbf{v}_\theta(\mathbf{a}_t, t, c)$, where c represents the conditioning information (visual observations, language instruction, and robot state). The training objective is to minimize the Mean Squared Error (MSE) between the predicted and the true vector fields. The loss function is defined as:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t, \mathbf{a}_1, \mathbf{a}_0, c} \left[\|\mathbf{v}_\theta(\mathbf{a}_t, t, c) - (\mathbf{a}_1 - \mathbf{a}_0)\|^2 \right] \quad (4)$$

This objective trains the Action Expert to effectively denoise any noisy action $\mathbf{a}_t$ at any time t and steer it towards

"pick up the black bowl in the top drawer of the wooden cabinet and place it on the plate"
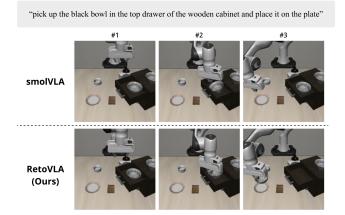
Fig. 4. Reusing Register Tokens [3] enables complex 3D spatial reasoning. While the baseline SmolVLA [4] fails by grasping a visually similar but incorrect object, RetoVLA correctly interprets the instruction "in the top drawer" by utilizing the injected spatial context. This highlights our model's superior ability to understand complex, multi-step manipulation commands.
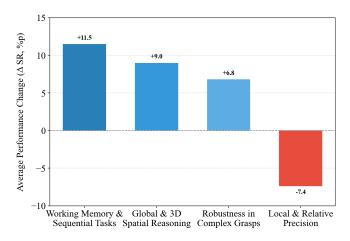

Fig. 5. Performance analysis of RetoVLA grouped by core capabilities. The chart shows the average success rate change compared to the baseline across 40 LIBERO tasks. RetoVLA shows significant improvements in tasks requiring high-level reasoning such as Working Memory and Global & 3D Spatial Reasoning, while a trade-off is observed in tasks that demand extreme local precision.

the correct ground-truth action $\mathbf{a}_0$, guided by the rich context provided by the VLM and our injected Register Tokens [3].

IV. EXPERIMENTS

We validate RetoVLA through three core sets of experiments: a standardized evaluation on the LIBERO benchmark, a demonstration of practical effectiveness on our robot arm, and a deeper analysis in a custom sim-to-real environment. An overview of our task environments is provided in Figure 3.

A. Experimental Setup

a) Standardized Benchmark: Our primary simulation analysis was conducted on the LIBERO benchmark [21], [22], a standardized suite composed of four main categories—'Spatial', 'Object', 'Goal', and '10 (Long)'—which comprehensively measure a model's diverse manipulation capabilities.

b) Real-World Environment and Task Design: Real-world performance was evaluated on our robot arm designed and built for this study. As detailed in Table I, we selected a suite of seven manipulation tasks to evaluate a

wide spectrum of skills, ranging from foundational pick-and-place to complex challenges requiring long-horizon planning ('Build Domino Line'), delicate interaction ('Jenga'), and precise 3D spatial understanding ('Close Drawer'). Since the baseline model was pre-trained only on the SO-100 robot, we performed fine-tuning on our robot arm using a collected dataset of 1,804 episodes.

c) Custom Simulation Environment: To bridge the gap between simulation and the real world, we built a custom simulation environment using the MuJoCo plugin within Unity. This physics-based simulation closely resembles our real-world setting, allowing for controlled yet realistic analysis of a subset of our real-world tasks.

d) Model and Training Setup: Unless otherwise specified, all models leveraged only the initial 16 layers of the pre-trained SmolVLM2-500M [23] backbone for computational efficiency. To isolate the impact of our contribution, we trained both models for 100k steps with a batch size of 64 under identical hyperparameters; the only difference was that RetoVLA injects two register tokens [3] into its Action Expert, a mechanism absent in the baseline.

B. LIBERO Benchmark Results

Our tests on the LIBERO simulation show that RetoVLA is a specialist, not a general booster. While overall scores in Table II show modest gains, a deeper analysis in Figure 5 reveals that RetoVLA excels at tasks requiring Working Memory (+11.5%p) and Global & 3D Spatial Reasoning (+9.0%p). This strongly suggests that reusing Register Tokens [3] directly contributes to long-term planning and understanding 3D scenes, as exemplified in Figure 4 where RetoVLA succeeds in a complex drawer task where the baseline fails. Conversely, we observed a trade-off: performance dropped on tasks requiring very precise local movements, suggesting that global information can be distracting for fine-grained local control.

TABLE II

OVERALL SUCCESS RATES ON THE FOUR MAIN CATEGORIES OF THE LIBERO BENCHMARK

Category	Task Characteristics	SmolVLA [4] (SR)	RetoVLA (SR)
Spatial	Single spatial relations	75.8%	**76.2%**
Object	Object-centric, local manipulation	70.8%	**71.8%**
Goal	Goal-directed, global placement	80.4%	80.4%
10 (Long)	Long-horizon, complex scenes	50.4%	50.4%

SUCCESS RATES ON SIMULATION ENVIRONMENT MANIPULATION TASKS

Task Name	SmolVLA [4] (SR)	RetoVLA (SR)	Performance Change (△)
Pick and Place	88%	**96%**	**+6.0%p**
Stack by Size	86%	**88%**	**+2.0%p**
Pull and Place (Jenga)	66%	**82%**	**+16.0%p**
Build Domino Line	28%	**52%**	**+24.0%p**
Clean Marker on Mirror	46%	**56%**	**+10.0%p**
MSR	62.8% ± 11.56%	**74.8%** ± **8.8%**	**+12.0%p**

TABLE IV

SUCCESS RATES ON REAL-WORLD MANIPULATION TASKS

Task Name	SmolVLA [4] (SR)	RetoVLA (SR)	Performance Change (△)
Pick and Place	86%	**92%**	**+6.0%p**
Stack by Size	**80%**	76%	-4.0%p
Pull and Place (Jenga)	60%	**78%**	**+18.0%p**
Build Domino Line	12%	**40%**	**+28.0%p**
Clean Marker on Mirror	38%	**52%**	**+14.0%p**
Close Drawer	60%	**96%**	**+36.0%p**
Move Bowl	16%	**38%**	**+14.0%p**
MSR	50.28% ± 11.06%	**67.42%** ± **9.07%**	**+17.14%p**

C. Real-World Experiments

To validate these findings in a more challenging setting, we tested RetoVLA on our robot arm. The results, summarized in Table IV, are even more compelling. RetoVLA significantly outperforms the baseline, boosting the Mean Success Rate (MSR) from 50.28% to 67.42%—a +17.14%p absolute improvement. This success was particularly pronounced in tasks requiring deep spatial understanding, such as 'Build Domino Line' (+28%p) and 'Close Drawer' (+36%p). Furthermore, its success in the 'Jenga' task (+18%p) shows that the injected context helps the model understand not just 'where' an object is, but 'how' to interact with it gently. These real-robot results confirm that our approach is a powerful and effective method for enhancing the spatial intelligence of lightweight robotic agents.

D. Custom Simulation Experiments

To establish a clearer causal link between our architectural changes and the observed performance gains, we conducted experiments in a custom simulation environment that mirrors our real-world setup. This controlled setting allows us to isolate the impact of our method from the unpredictability of real-world physics. The results, detailed in Table III, show a distinct MSR improvement of +12.0%p (62.8% to 74.8%).

Crucially, the performance trends were remarkably consistent with our other experiments. The most significant gains were again observed in tasks requiring long-term planning and sophisticated spatial interaction, such as 'Build Domino Line' (+24.0%p) and 'Pull and Place (Jenga)' (+16.0%p). This strong correlation across three different experimental domains (LIBERO, real-world, and custom simulation) provides compelling evidence that the performance improvements are not coincidental or environment-specific. Instead, they are a direct result of RetoVLA's enhanced spatial reasoning capabilities, validating that the injection of Register Tokens [3] is the core driver of its success.

V. CONCLUSIONS

In this paper, we introduced RetoVLA, a novel VLA architecture that reuses discarded Register Tokens [3] to inject spatial context into the Action Expert. Our approach enhances a model's spatial reasoning and working memory while maintaining high computational efficiency. Through extensive experiments, we demonstrated that RetoVLA excels at complex, multi-step manipulation tasks. Our findings prove that redefining Register Tokens [3] from a passive 'purifier' to an active 'spatial context provider' is an effective strategy for creating efficient yet intelligent robotic policies.

However, our work also reveals key limitations and opens avenues for future research. The performance trade-off on precision tasks warrants a deeper investigation into more sophisticated gating or fusion mechanisms. While our focus was on lightweight models, future work should validate this approach on larger backbones like OpenVLA [2]. Furthermore, evaluating RetoVLA's performance in dynamic environments and for other robotic domains, such as mobile navigation, remains a critical next step.

To accelerate future research and ensure full reproducibility, we will release our code, final model weights, experimental data, and the detailed hardware specifications for our robot arm to the community.

APPENDIX

This appendix provides supplementary materials, including a detailed ablation study on the number of register tokens [3] and a comprehensive task-by-task performance breakdown on the LIBERO benchmark suites.

A. Ablation Study

Table V presents our ablation study on the number of Register Tokens [3], conducted on the LIBERO Spatial benchmark to validate our design choices.

TABLE V

ABLATION STUDY ON THE NUMBER OF REGISTER TOKENS [3] USING THE LIBERO SPATIAL BENCHMARK.

Model Architecture	# Register Tokens [3]	Peak (SR)
SmolVLA [4] (Baseline)	0	75.8%
RetoVLA (Ours)	**2**	**76.2%**
RetoVLA (Ours)	4	75.2%
RetoVLA (Ours)	12	74.0%
RetoVLA (Ours)	16	74.6%

As shown in the table, optimal performance was achieved with 2 Register Tokens [3] (76.2%), outperforming the baseline (75.8%). However, increasing the number of tokens beyond two degraded performance. This suggests a trade-off where an excessive number may introduce distracting noise or redundancy instead of useful global information.

B. Detailed LIBERO Benchmark Results

Tables VI-IX provide a detailed, task-by-task comparison of our final model against the SmolVLA [4] baseline, with all results reported from each model's peak performance

TABLE VI

DETAILED COMPARISON ON THE LIBERO SPATIAL BENCHMARK.

Index	SmolVLA [4] (SR)	RetoVLA (SR)	Performance Change (Δ)
0	**88.0%**	80.0%	-8.0%p
1	**88.0%**	80.0%	-8.0%p
2	92.0%	**98.0%**	+6.0%p
3	70.0%	**86.0%**	+16.0%p
4	50.0%	**62.0%**	+12.0%p
5	**84.0%**	66.0%	-18.0%p
6	92.0%	**96.0%**	+4.0%p
7	62.0%	**72.0%**	+10.0%p
8	**80.0%**	68.0%	-12.0%p
9	52.0%	**54.0%**	+2.0%p
MSR	75.8%	**76.2%**	**+0.4%p**

TABLE VII

DETAILED COMPARISON ON THE LIBERO OBJECT BENCHMARK.

Index	SmolVLA [4] (SR)	RetoVLA (SR)	Performance Change (Δ)
0	56.0%	**58.0%**	+2.0%p
1	**82.0%**	74.0%	-8.0%p
2	70.0%	**76.0%**	+6.0%p
3	50.0%	50.0%	0.0%p
4	**94.0%**	82.0%	-12.0%p
5	54.0%	54.0%	0.0%p
6	82.0%	**84.0%**	+2.0%p
7	54.0%	**64.0%**	+10.0%p
8	78.0%	**92.0%**	+14.0%p
9	**88.0%**	84.0%	-4.0%p
MSR	70.8%	**71.8%**	**+1.0%p**

TABLE VIII

DETAILED COMPARISON ON THE LIBERO GOAL BENCHMARK.

Index	SmolVLA [4] (SR)	RetoVLA (SR)	Performance Change (Δ)
0	60.0%	**74.0%**	+14.0%p
1	94.0%	**100.0%**	+6.0%p
2	**86.0%**	84.0%	-2.0%p
3	**70.0%**	68.0%	-2.0%p
4	**90.0%**	86.0%	-4.0%p
5	**90.0%**	82.0%	-8.0%p
6	54.0%	**62.0%**	+8.0%p
7	92.0%	92.0%	0.0%p
8	**94.0%**	84.0%	-10.0%p
9	**74.0%**	72.0%	-2.0%p
MSR	80.4%	80.4%	0.0%p

TABLE IX

DETAILED COMPARISON ON THE LIBERO 10 BENCHMARK.

Index	SmolVLA [4] (SR)	RetoVLA (SR)	Performance Change (Δ)
0	**28.0%**	24.0%	-4.0%p
1	38.0%	**56.0%**	+18.0%p
2	64.0%	**74.0%**	+10.0%p
3	76.0%	**82.0%**	+6.0%p
4	**32.0%**	24.0%	-8.0%p
5	82.0%	**94.0%**	+12.0%p
6	**36.0%**	26.0%	-10.0%p
7	10.0%	**22.0%**	+12.0%p
8	**62.0%**	52.0%	-10.0%p
9	**76.0%**	50.0%	-26.0%p
MSR	50.4%	50.4%	0.0%p

checkpoint (typically 40k-50k steps for SmolVLA [4] and 60k-70k for RetoVLA) for a fair comparison. These results reinforce the central findings discussed in the main paper: RetoVLA excels in tasks requiring global spatial reasoning and working memory, at the cost of a performance trade-off in those demanding high local precision.

ACKNOWLEDGMENT

REFERENCES

[1] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning*. PMLR, 2023, pp. 2165–2183.

[2] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[3] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski, "Vision transformers need registers," *arXiv preprint arXiv:2309.16588*, 2023.

[4] M. Shukor, D. Aubakirova, F. Capuano, P. Kooijmans, S. Palma, A. Zouitine, M. Aractingi, C. Pascal, M. Russi, A. Marafioti *et al.*, "Smolvla: A vision-language-action model for affordable and efficient robotics," *arXiv preprint arXiv:2506.01844*, 2025.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[6] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[7] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[8] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, "π_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[9] W. Cai, I. Ponomarenko, J. Yuan, X. Li, W. Yang, H. Dong, and B. Zhao, "Spatialbot: Precise spatial understanding with vision language models," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 9490–9498.

[10] A. J. Hancock, A. Z. Ren, and A. Majumdar, "Run-time observation interventions make vision-language-action models more visually robust," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 9499–9506.

[11] G. Tang, S. Rajkumar, Y. Zhou, H. R. Walke, S. Levine, and K. Fang, "Kalie: Fine-tuning vision-language models for open-world manipulation without robot data," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 9507–9515.

[12] X. Tong, P. Ding, Y. Fan, D. Wang, W. Zhang, C. Cui, M. Sun, H. Zhao, H. Zhang, Y. Dang *et al.*, "Quart-online: Latency-free large multimodal language model for quadruped robot learning," *arXiv preprint arXiv:2412.15576*, 2024.

[13] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[14] J. Wen, Y. Zhu, J. Li, M. Zhu, Z. Tang, K. Wu, Z. Xu, N. Liu, R. Cheng, C. Shen *et al.*, "Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation," *IEEE Robotics and Automation Letters*, 2025.

[15] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, p. 02783649241273668, 2023.

[16] O. Siméoni, G. Puy, H. V. Vo, S. Roburin, S. Gidaris, A. Bursuc, P. Pérez, R. Marlet, and J. Ponce, "Localizing objects with self-supervised transformers and no labels," *arXiv preprint arXiv:2109.14279*, 2021.

[17] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.

[18] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 179–12 188.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[20] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," *arXiv preprint arXiv:2210.02747*, 2022.

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

[22] A. Polin, "libero_spatial_no_noops_lerobot_v21: A processed version of the libero spatial benchmark for the lerobot framework," https://huggingface.co/datasets/aopolin-lv/libero_spatial_no_noops_lerobot_v21, 2024.

[23] A. Marafioti, O. Zohar, M. Farré, M. Noyan, E. Bakouch, P. Cuenca, C. Zakka, L. B. Allal, A. Lozhkov, N. Tazi *et al.*, "Smolvlm: Redefining small and efficient multimodal models," *arXiv preprint arXiv:2504.05299*, 2025.