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

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,*]
[1]*School of Computing, Gachon University*

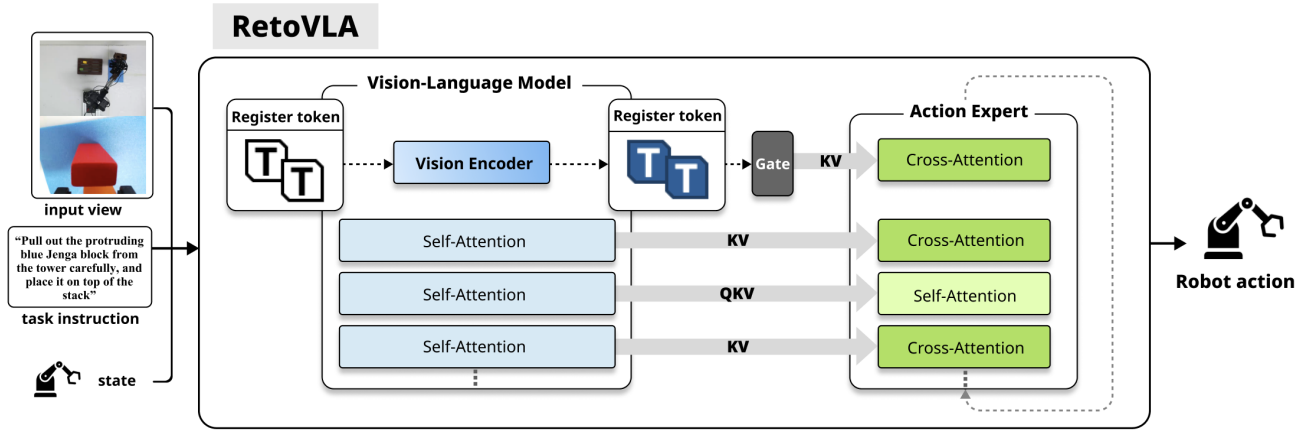

Fig. 1: Overview of the RetoVLA architecture. A dedicated spatial pathway (dashed arrow) injects global scene context by routing Register Tokens directly into the Action Expert. Unlike standard encoders that discard these tokens post-processing, RetoVLA repurposes them to seamlessly integrate spatial and semantic features, requiring no additional parameters.

***Abstract*— Vision-Language-Action (VLA) models have demonstrated robust performance across diverse robotic tasks. However, their high memory and computational demands often limit real-time deployment. While existing model compression techniques reduce the parameter footprint, they often drop in 3D spatial reasoning and scene layout understanding. This work introduces RetoVLA, an architecture designed to maintain spatial awareness in lightweight models by repurposing Register Tokens—learnable parameters originally introduced to mitigate attention artifacts in Vision Transformers. While these tokens are generally discarded once used, we repurpose them for their dense representation of global spatial context. RetoVLA integrates these recycled tokens directly into the action-planning module through a dedicated spatial context injection path. Our proposed design enables the recovery of global context without increasing the total parameter count. Real-world experiments using a 7-DOF manipulator show a 17.1%p improvement in average success rates over the baseline. Our results demonstrate that leveraging internal register tokens provides a highly effective mechanism for developing efficient, spatially-aware robotic agents. A video demonstration is available at: https://youtu.be/2CseBR-snZg**

I. INTRODUCTION

Vision-Language-Action (VLA) models such as RT-2 [1] and OpenVLA [2] map natural language instructions to robotic motor commands. Through web-scale pre-training, they enable strong zero-shot generalization in unseen environments. However, their scale and computational cost remain a major bottleneck for real-time deployment on physical hardware.

To address this efficiency issue, prior work has focused on smaller models such as SmolVLA [3]. However, reducing model size comes at a cost. Lightweight models often lose the capacity to represent 3D layouts and spatial relationships.

We address this limitation by reusing information that is typically discarded. Darcet et al. [4] observed that large Vision Transformers (ViTs) [5], such as DINOv2 [6], temporarily store global scene information in background image patches during training. Although this behavior supports global understanding, it degrades the visual details of those patches.

To mitigate these artifacts, researchers introduced Register Tokens [4]. These tokens act as dedicated scratchpads that absorb global information while preserving the visual fidelity of image patches. Although they are usually discarded after use, we examine whether they retain meaningful spatial information.

We hypothesize that these tokens encode a highly compressed summary of workspace layouts and 3D relationships, and that preserving them can improve robotic scene understanding.

Based on this hypothesis, we propose RetoVLA (Reusing Register Tokens [4] VLA). RetoVLA improves efficiency by

†These authors contributed equally to this work.
†First authors: {halo1225, taewan2002}@gachon.ac.kr.
*Corresponding author: andrewjchoi@gachon.ac.kr.

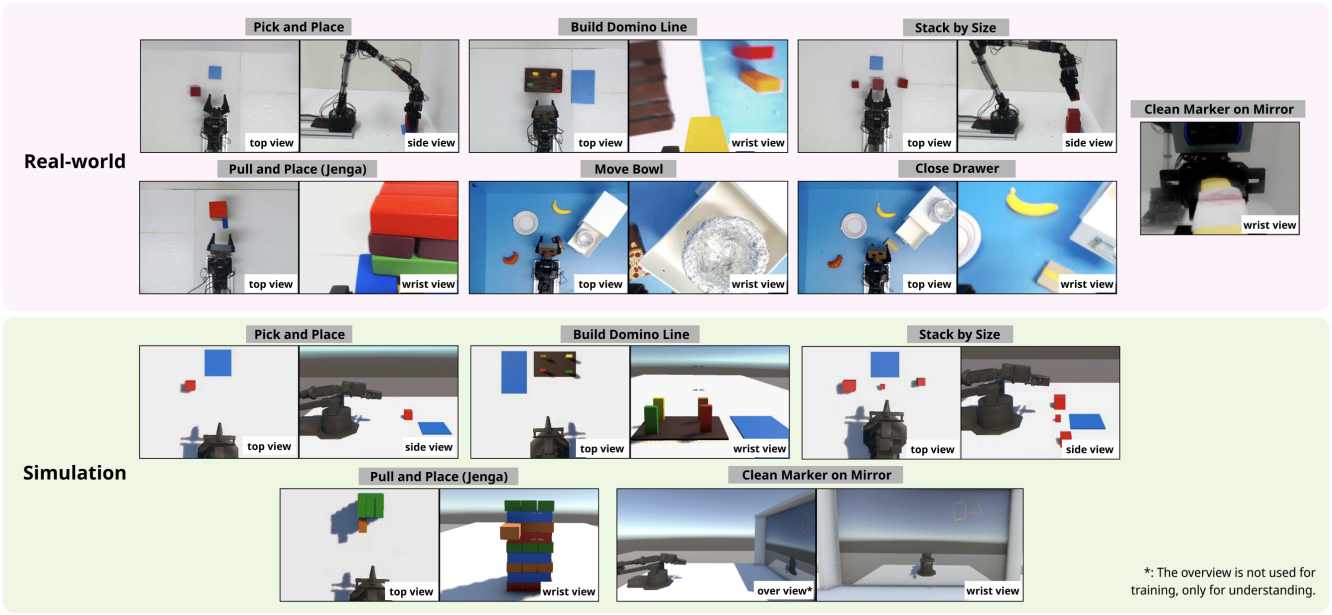

Fig. 2: Experimental setup overview. (Top) We use a custom robot arm for seven real-world manipulation tasks. Two tasks, 'Move Bowl' and 'Close Drawer', come directly from the LIBERO benchmark. (Bottom) We also implemented five additional tasks in the simulation to match our real-world setup, alongside the two LIBERO tasks.

recycling this latent information. Figure 3 summarizes the main result. Our contributions are:

1) A spatial context injection method: As shown in Figure 1, we repurpose Register Tokens [4] from artifact absorbers into providers of spatial context and feed them directly into the Action Expert.

2) An efficient design: We show that these tokens recover spatial awareness lost in lightweight models such as SmolVLA [3] without adding computational overhead.

3) Evaluation in simulation and on hardware: On the LIBERO benchmark and on a real 7-DOF robot, RetoVLA significantly outperforms the baseline, improving the real-world average success rate from 50.3% to 67.4% (+17.1%p).

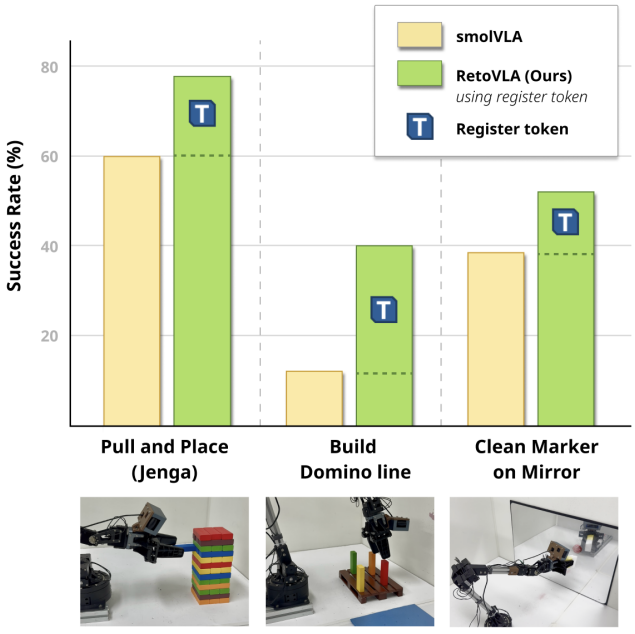

Fig. 3: Comparison of RetoVLA and the SmolVLA baseline on challenging real-world tasks. (Top) RetoVLA (green) significantly outperforms the baseline (yellow). (Bottom) This gain comes from reusing Register Tokens [4] (indicated by the 'T' icon) to inject global spatial context into the Action Expert.

II. RELATED WORK

A. Vision-Language-Action (VLA) Models

Transformers are now the standard backbone for robot learning. Early systems such as RT-1 [7] supported multi-task manipulation, while subsequent models connected large VLMs directly to robot actions. RT-2 [1], for example, represented robot actions as discrete text tokens, and more recent open-weight models such as OpenVLA [2] and $\pi 0$ [8] show strong zero-shot generalization.

However, their billions of parameters lead to slow inference and create major bottlenecks for reactive physical robots. Recent work improves spatial awareness [9], visual robustness [10], and low-data learning [11], [12], but the core problem of computational efficiency remains largely unresolved.

B. Lightweight Vision-Language-Action Models

To deploy VLAs on robots, recent work has developed smaller models. SmolVLA [3] uses smaller backbones, layer skipping, and LoRA [13] to reduce memory cost. Other

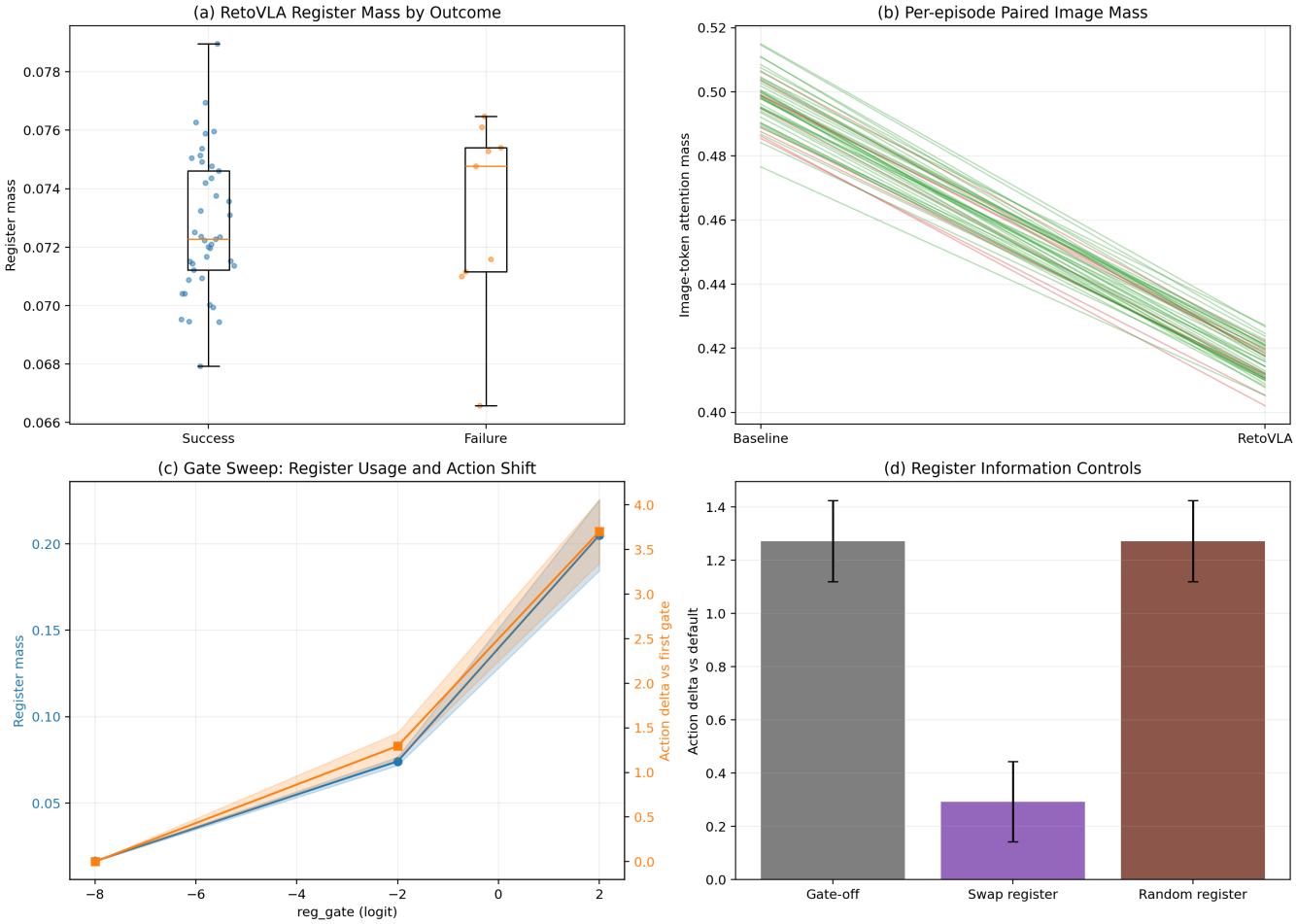

Fig. 4: Causal Analysis of Register Tokens. (a) Consistent attention mass across outcomes indicates active utilization. (b) RetoVLA reduces attention on image patches, moving its attention to global context. (c) Gate value (g) directly changes the action output, establishing causality. (d) Randomized tokens degrade performance, confirming specific information encoding.

methods explore model merging [14] or diffusion-based decoders [15], [16] for faster action generation.

Although these models are faster, they often lose critical 3D spatial reasoning and global context awareness. Prior work has tried to recover spatial information with external depth encoders [17], but those methods add computational overhead. RetoVLA takes a different approach: instead of adding new encoders, we reuse standard ViT Register Tokens [4]. This recycled latent information restores spatial awareness without sacrificing efficiency.

C. Artifacts and Register Tokens in Vision Transformers

Large ViTs [5] use background image patches as scratchpads for global scene information, which supports learning but hurts local feature predictions [18]–[20]. Learnable Register Tokens [4] address this issue by absorbing those artifacts. Although they are typically discarded after processing, we argue that they contain a useful spatial summary. RetoVLA uses these discarded tokens directly to guide robot motion.

III. METHOD

This section describes the RetoVLA architecture (Fig. 1). The core idea is to repurpose Register Tokens [4] as providers of spatial context. Instead of removing information, we recycle latent representations and route them directly into the Action Expert to provide geometric cues for motion planning.

A. Architecture Overview and Information Flow

RetoVLA retains the baseline structure but changes the internal data flow. Instead of sending only local patch features, we pass two streams. The Action Expert receives standard image patches together with Register Tokens [4] that carry a global scene summary. A cross-attention layer combines these streams.

B. Depth-Adaptive VLM Backbone

To balance speed and capability, we use the first $N = L/2$ layers of the pre-trained VLM. SmolVLA [3] showed that this truncation accelerates inference while preserving semantic capability.

TABLE I: Details of the seven real-world tasks and where we placed the cameras.

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...*	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...*	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.*

C. Spatial Context Injection via Register Tokens

We inject Register Tokens [4] directly into the Action Expert in three steps.

a) Register Token Generation: First, VLM image patch features ($\mathbf{P} \in \mathbb{R}^{B\times N\times D_{\text{vlm}}}$) enter a Spatial Context Aggregator, implemented as a standard multi-head attention block [21]. Initial Register Tokens [4] ($\mathbf{R}_{\text{init}} \in \mathbb{R}^{K\times D_{\text{vlm}}}$) act as the query, while image patches act as keys and values. This produces a global scene summary, $\mathbf{R}_{\text{scene}}$:

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

where K is the number of Register Tokens.

b) Injection into the Action Expert: We project $\mathbf{R}_{\text{scene}}$ to match the Action Expert and form key ($\mathbf{K}_{\text{reg}}$) and value ($\mathbf{V}_{\text{reg}}$) pairs. Concatenating these with the standard VLM pairs ($\mathbf{K}_{\text{vlm}}, \mathbf{V}_{\text{vlm}}$) lets the Action Expert access both local details and global context:

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

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

c) Gating Mechanism: Because global context can distract the policy during precision tasks, we introduce a learnable gate parameter g, passed through a sigmoid σ, to control the influence of the Register Tokens. This allows the model to adaptively balance local precision and global context.

D. Training Objective: Conditional Flow Matching

We train RetoVLA using conditional flow matching [22] to map pure noise to robot actions, conditioned on image and text inputs. Let $\mathbf{a}_0$ be the real robot action and $\mathbf{a}_1 \sim \mathcal{N}(0, \mathbf{I})$ be random noise. We define $\mathbf{a}_t = (1-t)\mathbf{a}_0 + t\mathbf{a}_1$ for $t \in [0, 1]$. The target vector is $\mathbf{u}_t = \mathbf{a}_1 - \mathbf{a}_0$.

The model θ predicts this vector as $\mathbf{v}_\theta(\mathbf{a}_t, t, c)$, optimized via MSE:

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

This objective trains the Action Expert to recover the correct action using spatial cues from the Register Tokens [4].

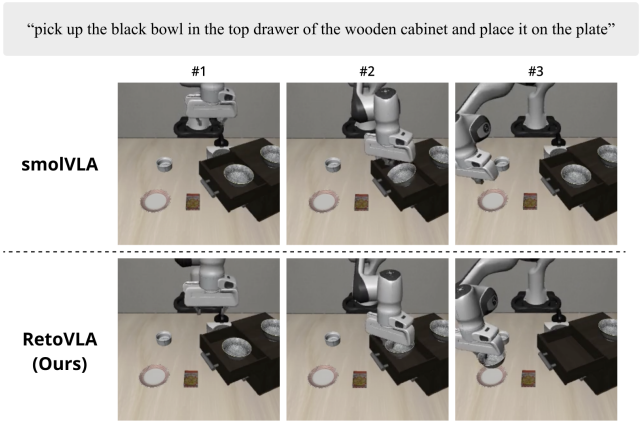

Fig. 5: Reusing Register Tokens [4] enables complex 3D spatial reasoning. The baseline SmolVLA [3] grasps a visually similar but incorrect object, whereas RetoVLA correctly interprets the instruction "in the top drawer" by using the injected spatial context. This example highlights the model's ability to handle complex multi-step manipulation commands.

IV. EXPERIMENTS

We evaluated RetoVLA on the LIBERO benchmark, a real robot arm, and a custom simulation environment (Fig. 2).

TABLE II: Overall success rates on the four main categories of the LIBERO benchmark

Category	Task Characteristics	SmolVLA [3] (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%

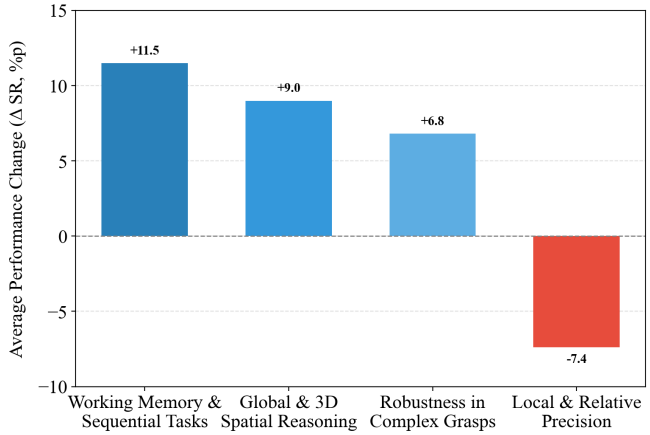

Fig. 6: Breakdown of RetoVLA's performance by skill type. We plot the change in success rate relative to the baseline across all 40 LIBERO tasks. RetoVLA performs better on tasks that require memory and 3D understanding, but it loses some accuracy on tasks that demand extreme local precision.

A. Experimental Setup

a) Standardized Benchmark: We used the 'Spatial', 'Object', 'Goal', and '10 (Long)' task groups from the LIBERO benchmark [23], [24] to assess diverse skills.

b) Real-World Environment and Task Design: We evaluated seven tasks on a custom 7-DOF robot arm (Table I), ranging from basic pick-and-place to long-horizon planning ('Build Domino Line') and 3D understanding ('Close Drawer'). We collected 1,804 human demonstrations to train both models on our custom hardware.

c) Custom Simulation Environment: To evaluate the policy without physical noise or lighting shifts, we digitally replicated our real-world setup using Unity's MuJoCo engine.

d) Model and Training Setup: For efficiency, we retained only the first 16 layers of SmolVLM2-500M [25]. Training lasted 100k steps with a batch size of 64. RetoVLA used two Register Tokens [4] in its Action Expert; the baseline used none.

B. LIBERO Benchmark Results

While the overall scores in Table II improve only slightly, Figure 6 reveals clear gains in Working Memory (+11.5%p) and Global & 3D Spatial Reasoning (+9.0%p). These results suggest that token reuse improves 3D understanding. Figure 5 shows that RetoVLA opens the correct drawer by capturing room layout, unlike the baseline. A minor drop on tasks that require extreme local precision suggests that broad scene context can occasionally interfere with fine control (see Appendix and Fig. 7).

C. Real-World Experiments

Real-world experiments (Table IV) show clear improvements. The mean success rate increases by 17.14%p, from 50.28% to 67.42%. RetoVLA performs especially well on tasks that require deeper spatial understanding, such as 'Build Domino Line' (+28%p) and 'Close Drawer' (+36%p). The improvement on 'Jenga' (+18%p) further indicates that spatial context helps careful object interaction.

TABLE III: Success rates on Simulation Environment manipulation tasks

Task Name	**SmolVLA [3] (SR)**	**RetoVLA (SR)**	**Performance Change (Δ)**
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 [3] (SR)**	**RetoVLA (SR)**	**Performance Change (Δ)**
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**

D. Custom Simulation Experiments

To isolate the impact of token reuse from physical noise, we conducted custom simulation experiments. Table III shows an MSR gain of 12.0%p, from 62.8% to 74.8%, with the largest improvements on 'Build Domino Line' (+24.0%p) and 'Jenga' (+16.0%p). The consistency across simulation, LIBERO, and real-world evaluation suggests that injecting Register Tokens [4] improves spatial task performance.

E. Analytical Studies: Understanding Attention and Causality

Attention maps (Fig. 4) show that the model actively uses Register Tokens. They receive high attention weights across tasks (Fig. 4a, b). Sweeping the gate value (g) changes the predicted action (Fig. 4c), and replacing the tokens with noise reduces success rates (Fig. 4d). These results indicate that the tokens contain meaningful spatial information.

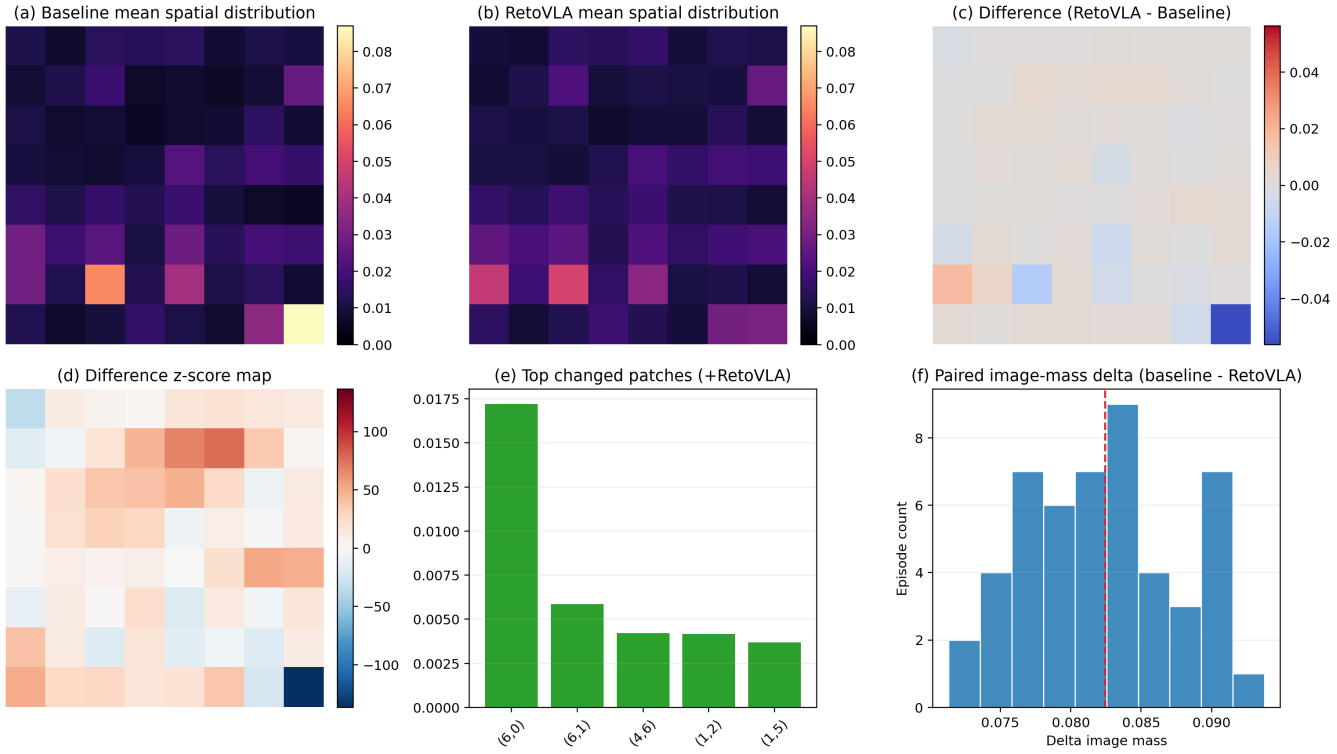

Fig. 7: Visualization of efficient attention redistribution. (a, b) The mean spatial attention distributions show where the model is looking. (c, d) The difference maps (RetoVLA minus Baseline) reveal a fascinating pattern: RetoVLA actively ignores the broad, featureless background regions (shown in blue). It offloads this global context processing to the Register Tokens. (e) Instead, it sharpens its focus. The top increased patches (green bars) correspond precisely to the grippers and the target objects. (f) The consistent positive shift in image-mass delta confirms that RetoVLA relies less on raw image tokens overall, trusting the injected tokens to handle the "big picture."

By relying on Register Tokens for global context, the model frees visual attention for task-relevant regions. Figure 7 shows a clear reduction in attention on flat background regions (blue areas in c, d). The saved attention shifts toward grippers and target objects (green bars in e), which helps explain the observed performance gains.

V. CONCLUSIONS

RetoVLA reuses Register Tokens [4] to provide spatial context for robotic action generation, improving the performance of lightweight models on complex tasks.

During evaluation, RetoVLA remains relatively robust to moving shadows, likely because Register Tokens capture broad layout information and reduce sensitivity to lighting changes. However, highly reflective objects remain challenging, which indicates that complex texture perception is still difficult for small models.

Our method also has limitations. Performance drops slightly on tasks that require extreme local precision, which suggests the need for a more selective gating mechanism. In addition, we evaluate the approach only on a small model. Future work should test the same idea on larger backbones such as OpenVLA [2] and on other robotic platforms, including mobile robots.

We will share our code, model weights, data, and hardware designs to support further research.

ACKNOWLEDGMENT

This work was supported by the Future Challenge Defense Technology Research and Development Program through the Agency For Defense Development(ADD) grant funded by the Defense Acquisition Program Administration(DAPA) in 2025 (No.915134201) and by Gachon University research fund (GCU-202500670001).

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] 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.

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

[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] T. Cho, R. Kim, and A. J. Choi, "Research on enhancing model performance by merging with korean language models," *Engineering Applications of Artificial Intelligence*, vol. 159, p. 111686, 2025.

[15] 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.

[16] 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.

[17] T. Cho, J. Koo, and A. J. Choi, "Enhancing imitation learning for space robotic arm via monocular vision-based spatial embedding," in *16th Asia-Pacific International Symposium on Aerospace Technology (APISAT 2025)*, 2025.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

APPENDIX

A. Ablation Study

TABLE V: Ablation study on the number of Register Tokens [4].

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

Table V shows that two Register Tokens are optimal on the LIBERO Spatial benchmark (76.2% vs. 75.8% for the baseline). Adding more tokens (≥ 4) degrades performance, likely because excessive global tokens act as noise and overwrite important local details. A small number of tokens provides the best balance between global context and local precision.

B. Detailed LIBERO Benchmark Results

Tables VI-IX compare RetoVLA and the baseline at their best checkpoints (40k–50k for SmolVLA and 60k–70k for RetoVLA). SmolVLA often overfits after 50k steps, whereas RetoVLA remains stable and improves, indicating Register Tokens aid training stability. These results confirm RetoVLA excels in spatial memory tasks with minimal local precision trade-offs.

C. Formal Definition of Analytical Metrics

We define *Attention Mass* (M) as the sum of cross-attention weights for a specific token group. Let $\alpha_{i,j}^{(l,h)}$ be the weight from Action Expert query i to Vision Encoder key j at layer l and head h. Due to softmax, $\sum_j \alpha_{i,j}^{(l,h)} = 1$.

Register Mass (M_{reg}) and Image Mass (M_{img}) are the average attention across all queries (N_q), heads (H), and layers (L):

$$M_{\text{reg}} = \frac{1}{N_q H L} \sum_{l=1}^{L} \sum_{h=1}^{H} \sum_{i=1}^{N_q} \sum_{j \in \mathcal{K}_{\text{reg}}} \alpha_{i,j}^{(l,h)} \tag{5}$$

$$M_{\text{img}} = \frac{1}{N_q H L} \sum_{l=1}^{L} \sum_{h=1}^{H} \sum_{i=1}^{N_q} \sum_{j \in \mathcal{K}_{\text{img}}} \alpha_{i,j}^{(l,h)} \tag{6}$$

where $\mathcal{K}_{\text{reg}}$ and $\mathcal{K}_{\text{img}}$ are the index sets for Register and Image Tokens.

To analyze visual focus shifts (Fig. 7f), we calculate *Delta Image Mass*:

$$\Delta M_{\text{img}} = M_{\text{img}}^{(\text{Baseline})} - M_{\text{img}}^{(\text{RetoVLA})} \tag{7}$$

A positive ΔM_{img} means RetoVLA pays less attention to normal patches, successfully offloading broad background understanding to Register Tokens.

Action Delta (Fig. 4c, d) is the L_2 distance between the normal prediction $\mathbf{v}_t$ and the changed prediction $\hat{\mathbf{v}}_t$ (e.g., when adding noise):

$$\Delta A = \|\mathbf{v}_t - \hat{\mathbf{v}}_t\|_2 \tag{8}$$

TABLE VI: Detailed comparison on the LIBERO Spatial benchmark.

Index	SmolVLA (SR)	RetoVLA (SR)	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 (SR)	RetoVLA (SR)	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 (SR)	RetoVLA (SR)	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 (SR)	RetoVLA (SR)	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

D. Detailed Causal Analysis of Register Tokens

As outlined in the main text, we conducted a causal analysis to determine whether the tokens directly influence the robot's decisions (Fig. 4). Empirical results confirm the active utilization of these tokens, which consistently exhibit high attention weights across both successful and failed executions (a). By relying on these tokens for global context, the model correspondingly allocates less attention mass to standard image patches (b).

Furthermore, we investigated whether perturbing the tokens alters the generated actions. Sweeping the gate parameter (g) yields a direct shift in the predicted action (c), and substituting the tokens with random noise significantly degrades the success rate (d). These findings demonstrate that Register Tokens encode meaningful spatial representations rather than acting merely as empty noise absorbers.

E. Detailed Visualization of Spatial Attention Shift

Figure 7 illustrates the redistribution of visual attention within RetoVLA. The difference maps (c, d) reveal a distinct reduction in attention mass over featureless background regions, such as tables or walls (indicated in blue). This shift suggests that the model effectively offloads global context processing to the Register Tokens.

By conserving attention capacity in these background regions, the model can focus more precisely on task-relevant features. Panel (e) demonstrates that this reclaimed attention is redirected specifically toward the robot grippers and target objects. Finally, panel (f) confirms this trend across numerous episodes: RetoVLA exhibits a consistent decrease in its reliance on standard image patches for global scene comprehension, thereby enabling the Action Expert to concentrate on fine-grained manipulation.