



CogVLA: Cognition-Aligned Vision-Language-Action Models via Instruction-Driven Routing & Sparsification

Appendix

This appendix provides comprehensive supplementary material to support the methodology, analysis, and findings presented in the main paper.

- Section A describes implementation details, including model and training details.
- Section B outlines experimental details for both simulation and real-world settings.
- Section C presents extended quantitative analyses, including multi-seed evaluations, additional ablation studies, and expanded real-world results.
- Section D provides supplementary qualitative analyses, such as diverse task executions and instruction-to-observation attention visualizations.
- Section E discusses additional insights into the motivation behind CogVLA, highlights its current limitations, and reflects on the broader societal implications and potential risks.
- We provide third-person view videos in the supplementary materials showing CogVLA performing manipulation tasks in a fully autonomous mode, played at 1× speed. Due to the need for remote communication during each action chunk prediction, slight delays are introduced by network latency. In future deployments, we plan to run CogVLA locally on hardware with more than 20GB of GPU memory (e.g., RTX 4090 with 24GB) to eliminate this latency.

A Implementation Details

A.1 Model Details

EFA-Routing. In Step 1, each of the two vision encoders uses 64 aggregation tokens, thereby reducing the number of visual tokens to 25% of the original. In addition, the scale and shift vectors for FiLM, γ_i and β_i , are derived from a linear transformation of the text embedding. In Step 2, a two-layer MLP is applied to the text embedding to produce routing weights for the two vision encoders.

LFP-Routing. In this module, we employ a shifted cosine schedule [19] to control the proportion of visual tokens retained at each layer. The formulation is as follows:

$$\beta_l = \frac{1}{2} \cos \frac{\pi l}{L} + \eta, \quad l = 1, 2, \dots, L \quad (1)$$

where L denotes the total number of layers in the LLM, which is $L = 32$ for CogVLA. The constant η is a shift factor that vertically adjusts the cosine decay curve, providing a flexible mechanism to control the overall computational cost of the model. In our implementation, η is set to 0.5. Specifically, we apply a clamp operation to constrain β_l within the range $[0.05, 0.85]$. As a result, LFP-Routing achieves approximately a 50% token pruning rate.

In addition, the instruction-conditioned scaling and shifting functions $\gamma_{\text{LLM}}(\cdot)$ and $\beta_{\text{LLM}}(\cdot)$ in LFP-Routing are both implemented using two-layer MLPs, with a hidden layer dimension of 2048, resulting in a parameter count almost identical to that of a direct linear layer.

A.2 Training Details

LIBERO Training Setup. We adopt OpenVLA [9] as the backbone model and set the action chunk size to $K = 8$. Fine-tuning is performed using Low-Rank Adaptation (LoRA) with a rank of 32 and an α value of 64. The model is trained for 60K steps with a batch size of 64 and an initial learning rate of $5e-4$. Checkpoints are evaluated every 10K steps, and the best-performing checkpoint is selected for reporting.

Real-World Training Setup. For the real-world experiments, we set the chunk size to $K = 25$ and fine-tune OpenVLA using LoRA with a rank of 32 and an alpha value of 64. The model is trained

41 with a batch size of 32 for a total of 80K steps. The initial learning rate was set to 5e-4, which is
42 reduced to 5e-5 after 50K steps. Starting from step 60K, we evaluate checkpoints every 10K steps
43 and report the best-performing checkpoint.

44 B Experimental Details

45 B.1 Simulation Benchmark

46 We evaluate CogVLA on the LIBERO simulation benchmark [11], a standardized suite of language-
47 conditioned robotic manipulation tasks. Unlike earlier benchmarks such as RL Bench [6], LIBERO
48 features more complex and diverse instructions, averaging 10.48 words per command compared to
49 only 3.34 in RL Bench. This makes it a more suitable testbed for assessing the model’s capacity
50 in language grounding and multimodal reasoning. LIBERO comprises four task suites—**Spatial**,
51 **Object**, **Goal**, and **Long**—each containing 10 tasks with 50 human-teleoperated demonstrations.
52 These suites are designed to probe distinct reasoning capabilities:

- 53 • **LIBERO-Spatial** evaluates spatial reasoning capabilities by presenting identical objects
54 arranged in different spatial configurations. The agent must interpret spatial relations (e.g.,
55 left/right, front/behind) described in the instruction to complete the task correctly.
- 56 • **LIBERO-Object** measures the model’s ability to generalize across object categories. While
57 spatial layouts remain fixed, the manipulated objects vary in type, shape, or color, requiring
58 the agent to ground object-referential language and adapt its actions accordingly.
- 59 • **LIBERO-Goal** tests task-oriented comprehension by altering the goal specification while
60 keeping object types and spatial layouts constant. The agent must disambiguate subtle
61 differences in instruction semantics to execute distinct manipulation outcomes.
- 62 • **LIBERO-Long** challenges the agent with multi-step, long-horizon tasks involving diverse
63 objects and environments. Success requires not only grounded perception and instruction
64 following, but also sequential planning.

65 CogVLA is trained and evaluated under the same setting as OpenVLA [8] to ensure comparability.
66 We report results on all four suites to validate the model’s generalization, efficiency, and semantic
67 grounding capabilities.

68 B.2 Real-World Setup

69 We deploy CogVLA on Cobot Agilex ALOHA [3] manipulation platform, to validate its real-world
70 applicability. The real-world evaluation consists of five diverse tasks involving both single-arm and
71 coordinated dual-arm manipulation. To assess robustness and generalization, we introduce moderate
72 data augmentation by varying object attributes (e.g., size, color) and rearranging spatial layouts.

73 B.2.1 Task Descriptions

74 We report the results of **Tasks 1–3** in the main paper, and provide additional results for **Tasks 4–5** in
75 this appendix. The instructions and descriptions for **Tasks 1–5** are provided below:

- 76 • **Task 1:** “*Put the cube into the plate, and then put the toy into the bowl.*”
77 A two-step pick-and-place task involving object category understanding and temporal
78 sequencing. This is a dual-arm task consisting of two sequential subtasks: 1) “*Put the cube*
79 *into the plate*” with the left arm, and 1) “*Put the toy into the bowl*” with the right arm. Task
80 success is achieved only when both subtasks are completed successfully. We report success
81 rates for each subtask and the overall task.
- 82 • **Task 2:** “*Open the drawer, place the toy into the drawer, and then close it.*”
83 A composite task requiring interaction with articulated objects and multi-stage execution.
84 This is a dual-arm task consisting of three sequential subtasks: 1) “*Open the drawer*” with
85 the left arm, 2) “*Place the toy into the drawer*” with the right arm, and 3) “*Close the drawer*”
86 with the left arm. Task success requires all three subtasks to be completed. We report success
87 rates for each subtask and the overall task.
- 88 • **Task 3:** “*Fold the T-shirt.*”
89 A soft-body manipulation task that evaluates the system’s ability to handle deformable

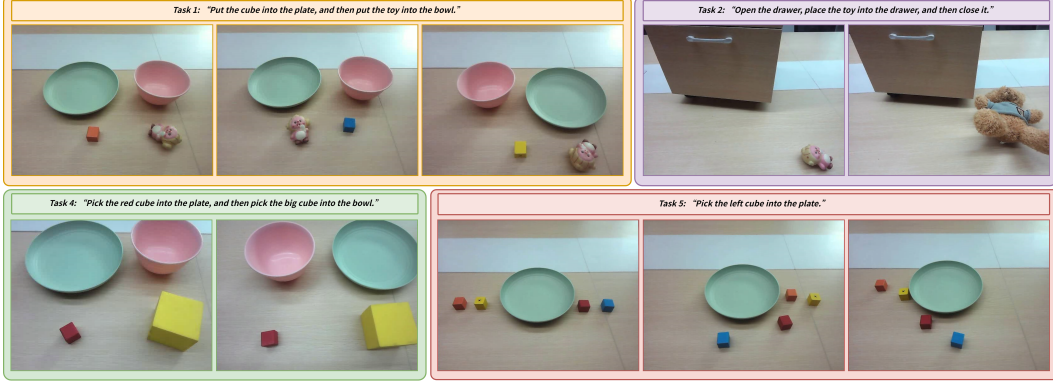


Figure 1: **Initial Observations for Structured Demonstration Groups.** Representative initial states for each demonstration group in Tasks 1–2 and Tasks 4–5. Across groups, object attributes (e.g., color, size), spatial configurations, and object categories are varied to introduce controlled diversity. This design enhances cross-instance generalization while maintaining intra-task consistency.

90 objects. This is a dual-arm task consisting of three sequential folding steps. Task success is
 91 determined by the successful execution of all three steps. We report intermediate success
 92 rates for each step and the overall task performance.

- 93 • **Task 4:** “Pick the red cube into the plate, and then pick the big cube into the bowl.”
 94 A multi-attribute grounding task requiring comprehension of both color and size references.
 95 This is a dual-arm task consisting of two sequential subtasks: 1) “Pick the red cube into the
 96 plate” with the left arm, and 2) “Pick the big cube into the bowl” with the right arm. Task
 97 success is achieved only when both subtasks are completed. We report success rates for
 98 each subtask and the overall task.
- 99 • **Task 5:** “Pick the left cube into the plate.”
 100 A spatial reasoning task focusing on relative positioning and egocentric understanding. This
 101 is a single-arm task consisting of one pick-and-place action. We report the final task success
 102 rate.

103 B.2.2 Data Collection and Augmentation

104 We collect real-world training data for the Cobot Agilex ALOHA robot via human teleoperation.
 105 For Tasks 1–5, we gather 45, 45, 30, 30, and 45 expert demonstrations, respectively. To enhance
 106 generalization, we apply data augmentation by varying initial object poses, object attributes, and
 107 tabletop layouts. Specifically:

- 108 • **Task 1:** Demonstrations are grouped into three sets of 15. Across groups, the cube properties,
 109 the relative positions of the plate and bowl, and the spatial layout of four objects vary.
- 110 • **Task 2:** The dataset is divided into two groups with 25 and 20 demonstrations, respectively.
 111 The toy differs between groups in category, size, and color.
- 112 • **Task 4:** Two groups of 15 demonstrations are collected, with differences in cube properties
 113 across groups.
- 114 • **Task 5:** Demonstrations are grouped into three sets of 15. The leftmost cube, the relative
 115 positions among cubes, and the overall object layout vary across groups.

116 This structured grouping strategy introduces controlled variability across tasks, enabling the model to
 117 learn more robust cross-instance representations. It also facilitates stable convergence during training
 118 by balancing intra-task consistency with inter-group diversity. The initial observations for each group
 119 in Tasks 1–2 and Tasks 4–5 are illustrated in **Fig. 1**.

Table 1: **Multi-seed evaluation results in simulation.** Task success rates (SR) are compared across four task categories on the LIBERO benchmark. “†” denotes our reproduced results. CogVLA demonstrates strong and consistent performance.

Method	Spatial SR ↑	Object SR ↑	Goal SR ↑	Long SR ↑	Average SR ↑	RK ↓
OpenVLA [CoRL’24] [8]	84.7 ± 0.9	88.4 ± 0.8	79.2 ± 1.0	53.7 ± 1.3	76.5 ± 0.6	5
SpatialVLA [RSS’25] [13]	88.2 ± 0.5	89.9 ± 0.7	78.6 ± 0.6	55.5 ± 1.0	78.1 ± 0.7	4
STAR [ICML’25] [5]	95.5 ± 0.6	98.3 ± 0.2	95.0 ± 0.7	88.5 ± 0.3	94.3 ± 0.1	2
CoT-VLA [CVPR’25] [20]	87.5 ± 1.4	91.6 ± 0.5	87.6 ± 0.6	69.0 ± 0.8	83.9 ± 0.6	3
CogVLA	98.5 ± 0.5	98.8 ± 0.4	96.5 ± 0.6	95.2 ± 1.1	97.4 ± 0.4	1

Table 2: **Extended real-world results on Tasks 4–5.** Performance comparison on the Cobot Agilex ALOHA tasks. “†” indicates our reproduced results.

Method	Task 4		Task 5	Average
	Red Cube→Plate	+Big Cube→Bowl	Left Cube→Plate	SR
PD-VLA† [16]	7/10	5/10	6/10	60.0%
OpenVLA-OFT† [7]	7/10	6/10	6/10	63.3%
CogVLA	8/10	7/10	8/10	76.7%

C Supplementary Quantitative Analysis

C.1 Multi-Seed Evaluation

To evaluate the statistical robustness and consistency of CogVLA’s performance, we conduct multi-seed evaluations on the LIBERO benchmark. For each of the four task suites (Spatial, Object, Goal, and Long), we run experiments using three independent random seeds and report the mean success rate along with the standard deviation.

As shown in **Tab. 1**, CogVLA exhibits consistently high performance across different seeds, with standard deviations ranging from 0.2% to 0.6%. This indicates stable learning behavior and further validates the strong generalization capability of CogVLA’s three-stage instruction-driven architecture across diverse task types.

To assess statistical significance, we conduct two-tailed paired t -tests comparing CogVLA against OpenVLA, under identical random seed settings. For each task within each suite, 50 evaluation trials are performed. OpenVLA exhibits higher variability, with standard deviations ranging from 0.6% to 1.3%. Across all task suites, CogVLA’s improvements are statistically significant ($p < 0.05$), confirming that the observed gains are not due to random fluctuation. These findings are consistent with our observations in the main paper and highlight CogVLA’s ability to balance efficiency and performance through high-ratio visual sparsification.

C.2 Extended Real-World Task Results

In addition to the results reported in the main paper, we present the performance of CogVLA on Tasks 4 and 5, as shown in **Tab. 2**.

- **Task 4** (“Pick the red cube into the plate, and then pick the big cube into the bowl”) evaluates the model’s ability to ground multi-attribute language and execute sequential actions. CogVLA achieves the highest success rates across both subtasks and the overall task, demonstrating strong compositional understanding of attribute references such as color and size.
- **Task 5** (“Pick the left cube into the plate”) focuses on egocentric spatial reasoning, requiring precise interpretation of relative spatial references from the agent’s visual perspective. CogVLA maintains a high success rate in this setting, indicating robust grounding of spatial concepts.

These results further validate CogVLA’s ability to generalize to real-world tasks that demand fine-grained language grounding and spatial understanding.

C.3 Extended Ablation Studies

We extend the sparsification analysis by evaluating additional Stage 1/Stage 2 configurations: 2×2 and 4×4 , while keeping the total sparsification ratio fixed at $4 \times$ and $16 \times$, respectively. These configurations are compared alongside the baseline $8 \times$ setting with different asymmetric allocations (e.g., $2 \times - 4 \times$ and $4 \times - 2 \times$), allowing us to systematically assess how the distribution of sparsity across stages impacts downstream performance.

As shown in **Tab.3**, the 2×2 setting provides a favorable trade-off between performance and computational efficiency. In contrast, the 4×4 setting leads to a slight degradation in performance, suggesting that excessive sparsification across both stages may hinder the preservation of task-relevant information.

Interestingly, the asymmetric configurations, particularly the $4 \times - 2 \times$ setup, outperform their symmetric counterparts, achieving the highest spatial

success rate of 98.6. This highlights the advantage of applying a more aggressive token reduction in Stage 1 (EFA-Routing), where redundant visual tokens can be effectively compressed via instruction-guided aggregation. Subsequently, Stage 2 (LFP-Routing) performs finer-grained token pruning in a context-aware manner within the language model, allowing for better preservation of task-relevant information.

These findings support the core design principle of CogVLA: progressive sparsification with an asymmetric allocation tailored to the representational characteristics of each stage. By balancing early-stage compression and late-stage selectivity, the model achieves both computational efficiency and high task accuracy, reinforcing the importance of stage-aware sparsity scheduling in multimodal architectures.

Table 3: **Supplementary ablation on sparsification ratio allocation.** Spf.Ratio denotes the sparsification ratio, which can be adjusted based on the performance–efficiency trade-off. CogVLA achieves better performance when a relatively higher sparsification ratio is allocated to Stage 1 compared to Stage 2.

Stage 1	Stage 2	Spf.Ratio	Spatial SR	FLOPs
$2 \times$	$2 \times$	$4 \times$	96.4 (-2.2)	3.87 T
$4 \times$	$4 \times$	$16 \times$	93.2 (-5.4)	2.30 T
$2 \times$	$4 \times$	$8 \times$	94.6 (-4.0)	2.72 T
$4 \times$	$2 \times$	$8 \times$	98.6	2.72 T

D Supplementary Qualitative Analysis

D.1 Additional Visualizations of Simulation and Real-World Results

We present additional qualitative results from both simulation and real-world experiments to illustrate CogVLA’s generalization and execution capabilities. As shown in **Fig. 5**, the model consistently completes multi-step tasks across diverse environments, object configurations, and instruction variants.

In real-world tasks with varying instructions, CogVLA accurately interprets long-horizon commands and produces coherent action sequences. These examples further highlight the model’s ability to maintain cross-modal consistency and temporal reasoning, as well as its robustness in simulation-to-reality transfer. **Fig. 2** illustrates the real-world manipulation workflows for Tasks 1-5. For Task 1, we provide multi-view observations from the *Front Camera*, *Left Wrist Camera*, and *Right Wrist Camera*. For Tasks 2-5, only *Front Camera* observations are shown for clarity. In **Fig. 3**, we present a third-person view demonstration of CogVLA performing a manipulation task in the lab. The corresponding MP4 video file is provided in the supplementary materials.

D.2 Instruction-to-Observation Attention Maps

To gain deeper insights into how CogVLA aligns language instructions with visual observations, we visualize the attention maps generated by the cross-modal attention modules. As shown in **Fig. 4**, the attention weights highlight task-relevant regions in the input image.

These visualizations demonstrate that CogVLA’s instruction-aware routing mechanisms effectively guide the perception module to attend to semantically meaningful areas, enabling robust visual grounding even in cluttered or ambiguous scenes.



Figure 2: **Real-world Manipulation Workflows and Visualizations for Tasks 1–5.** Each task panel illustrates the initial setup and CogVLA’s execution process based on the natural language instruction. For Task 1, multi-view observations from the *Front Camera*, *Left Wrist Camera*, and *Right Wrist Camera* are provided to capture dual-arm coordination. For Tasks 2–5, representative frames from the *Front Camera* highlight key manipulation stages. These visualizations support interpretation of task complexity and grounding behavior.

E Discussion

E.1 Supplementary Details on the Motivation

CogVLA is motivated by the need to improve both computational efficiency and cross-modal semantic alignment in instruction-conditioned robotic systems. Its architectural design is informed by cognitive science research on how humans process language, perceive their environment, and execute actions in a coherent and goal-directed manner.

Cognitive studies suggest that humans rely on structured inductive biases—often termed “intuitive theories”—to interpret the world, including intuitive physics, causality, and theory of mind [10, 18]. While recent multimodal large language models exhibit partial competence in these areas, they often lack robustness in compositional reasoning and causally grounded behavior [14].

To address these limitations, CogVLA adopts a biologically inspired architecture that reflects the division of functional roles observed in the human brain. Specifically, we draw connections between the model’s three routing modules and key components in human multimodal cognition: the **Visual Attention System (VAS)**, the **Supplementary Motor Area (SMA)**, and the **Premotor Cortex (PMC)**.

Visual Attention System (VAS) → Encoder-FiLM. The human visual attention system selectively enhances perception of task-relevant features while suppressing distractors [1]. Top-down signals from frontal and parietal cortices bias visual processing toward objects or regions mentioned in language or necessary for action. This selective modulation improves efficiency and semantic grounding in complex scenes. In CogVLA, the **Encoder-FiLM** module mimics VAS by dynamically modulating visual encoder features conditioned on instructions, focusing perception on semantically

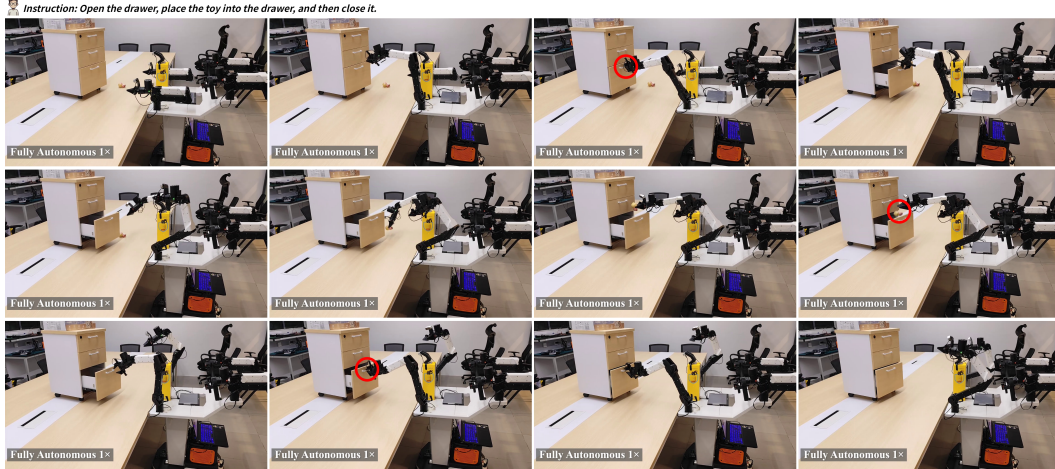


Figure 3: **Third-person visualization of CogVLA performing a manipulation task.** The corresponding video is provided in the supplementary materials. Gripper details are highlighted with red circles.

relevant regions and reducing redundancy [12]. This allows the model’s perception to be grounded in context, much as the brain’s attention system tunes visual processing to relevant aspects of a scene during coordinated vision-language tasks.

Supplementary Motor Area (SMA) → LLM-FiLM. The SMA plays a key role in planning and sequencing actions, even in the absence of physical movement [15, 17]. It integrates multimodal information and high-level goals to shape future motor behavior, before engaging primary motor circuits. In CogVLA, the **LLM-FiLM** module can be seen as the “intention planner” of the model and serves a similar function: it injects task-specific intent into the language model, pruning irrelevant visual-linguistic tokens and steering the model toward generating appropriate action plans. This enables more efficient and intention-aligned reasoning, analogous to how the SMA organizes abstract motor programs before execution.

Premotor Cortex (PMC) → V-L-A Coupled Attention. The premotor cortex is involved in translating perceptual cues into executable motor plans [4, 2]. It contains visuomotor neurons that represent both the perception of object affordances and the intended grasping actions, enabling visuomotor grounding. CogVLA’s **V-L-A Coupled Attention** module reflects this mechanism by integrating visual, linguistic, and action representations through a unified attention mechanism. This ensures that generated actions are causally and temporally coherent with respect to both the observed scene and the given instruction.

By aligning its modular design with biologically plausible cognitive functions, CogVLA offers not only performance and efficiency gains, but also a cognitively grounded pathway for improving generalization and interpretability in embodied multimodal agents.

E.2 Limitation and Future Work

While CogVLA demonstrates strong performance across simulation and real-world tasks, several limitations remain. First, the current instruction-to-vision routing relies on predefined sparsity ratios and fixed token pruning schedules, which may not adapt optimally to varying instruction complexity or scene difficulty. Second, although the model generalizes well within the LIBERO and ALOHA settings, its performance under out-of-distribution instructions or unseen manipulation categories is yet to be thoroughly evaluated.

In future work, we aim to explore adaptive sparsification mechanisms conditioned on task semantics and environmental uncertainty. Moreover, integrating lifelong learning or online adaptation strategies may further enhance CogVLA’s robustness in open-world deployment scenarios. Lastly, extending the framework to support multimodal feedback (e.g., haptic or force sensing) could improve its applicability to fine-grained manipulation tasks.

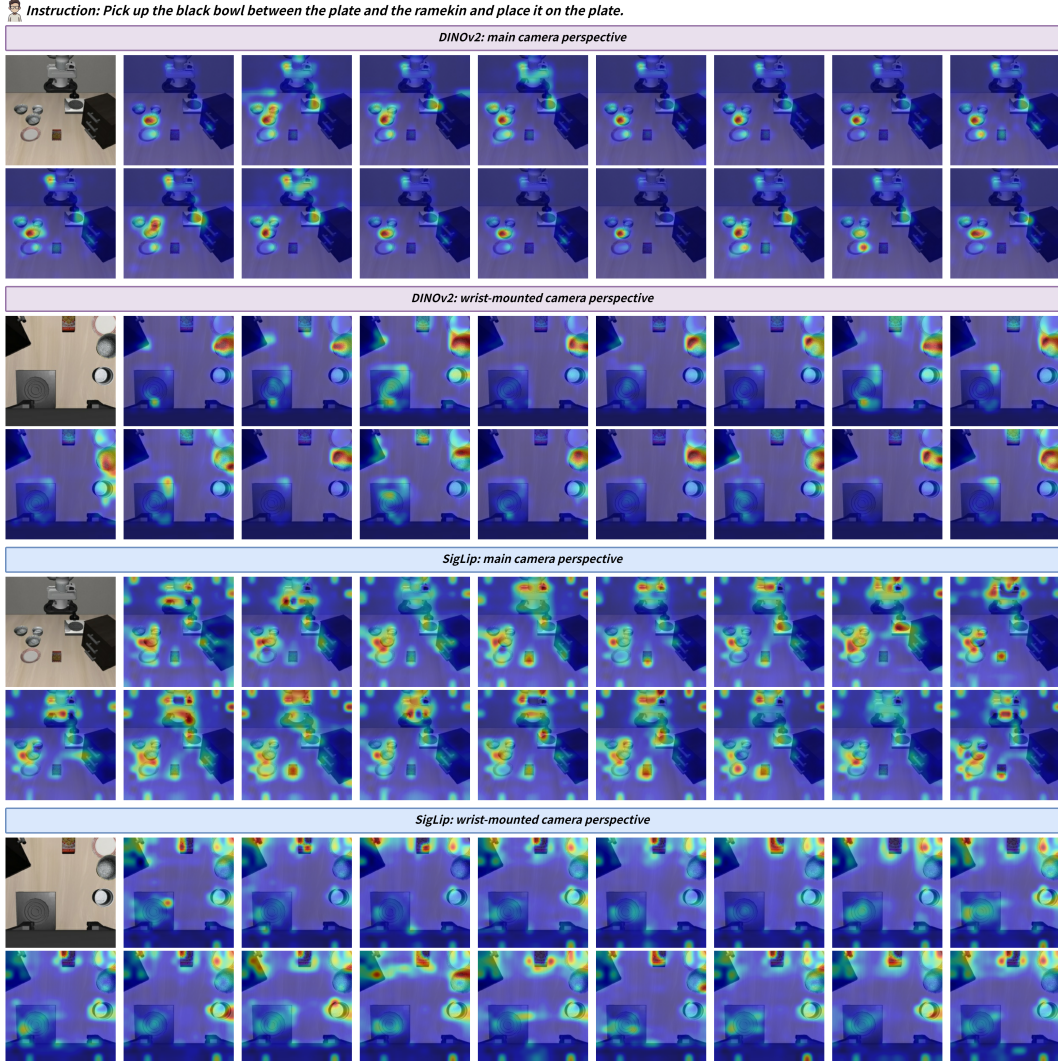


Figure 4: **Attention maps between aggregation tokens and patch tokens in DINOv2 and SigLIP.** We visualize the attention maps from 17 out of 64 aggregation tokens to the patch tokens of the observation, covering four sets of visualizations across two visual encoders and two camera views. The input language instruction is: “Pick up the black bowl between the plate and the ramekin and place it on the plate.” Both DINOv2 and SigLIP exhibit varying degrees of focused attention on patch tokens relevant to the instruction.

E.3 Broader Impact and Potential Risk

CogVLA advances the efficiency and interpretability of instruction-driven robotic manipulation, offering potential benefits in applications such as assistive robotics, household automation, and industrial assembly. Its biologically inspired sparsification and routing mechanisms reduce computation cost, making it more accessible for resource-constrained platforms. However, as with any vision-language-action system, risks include misinterpretation of ambiguous instructions, failure in unpredictable environments, and bias amplification from training data. If deployed in safety-critical settings without appropriate safeguards, such failures could lead to unintended behaviors or physical harm. We encourage the community to adopt robust evaluation protocols, prioritize transparency in model behavior, and consider human-in-the-loop designs to mitigate such risks. Broader societal considerations—including data diversity, accessibility, and responsible deployment—should guide future development of systems built upon CogVLA.



Figure 5: Manipulation Workflows and Visualizations in the LIBERO Simulation Environment. We present the execution processes of CogVLA across LIBERO-Spatial, LIBERO-Object, LIBERO-Goal, and LIBERO-Long, demonstrating its strong performance under diverse instructions and a wide range of tasks.

References

- [1] Maurizio Corbetta and Gordon L Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, 3(3):201–215, 2002.
- [2] Leonardo Fogassi, Pier Francesco Ferrari, Benno Gesierich, Stefano Rozzi, Fabian Chersi, and Giacomo Rizzolatti. Parietal lobe: from action organization to intention understanding. *Science*, 308(5722):662–667, 2005.
- [3] Zipeng Fu, Tony Z. Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. In *Conference on Robot Learning (CoRL)*, 2024.
- [4] Vittorio Gallese, Luciano Fadiga, Leonardo Fogassi, and Giacomo Rizzolatti. Action recognition in the premotor cortex. *Brain*, 119(2):593–609, 1996.
- [5] Li Hao, Lv Qi, Shao Rui, Deng Xiang, Li Yinchuan, HAO Jianye, and Nie Liqiang. Star: Learning diverse robot skill abstractions through rotation-augmented vector quantization. *International Conference on Machine Learning (ICML)*, 2025.
- [6] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J. Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 2020.
- [7] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025.
- [8] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.
- [9] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. In *8th Annual Conference on Robot Learning*, 2024.
- [10] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.
- [11] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *arXiv preprint arXiv:2306.03310*, 2023.
- [12] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
- [13] Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025.
- [14] Luca M Schulze Buschoff, Elif Akata, Matthias Bethge, and Eric Schulz. Visual cognition in multimodal large language models. *Nature Machine Intelligence*, pages 1–11, 2025.
- [15] Michael Schwartze, Kathrin Rothermich, and Sonja A Kotz. Functional dissociation of pre-sma and sma-proper in temporal processing. *Neuroimage*, 60(1):290–298, 2012.
- [16] Wenxuan Song, Jiayi Chen, Pengxiang Ding, Han Zhao, Wei Zhao, Zhide Zhong, Zongyuan Ge, Jun Ma, and Haoang Li. Accelerating vision-language-action model integrated with action chunking via parallel decoding. *arXiv preprint arXiv:2503.02310*, 2025.
- [17] Shoji Tanaka and Eiji Kirino. Dynamic reconfiguration of the supplementary motor area network during imagined music performance. *Frontiers in human neuroscience*, 11:606, 2017.
- [18] Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011.
- [19] Jun Zhang, Desen Meng, Ji Qi, Zhenpeng Huang, Tao Wu, and Limin Wang. p-mod: Building mixture-of-depths mllms via progressive ratio decay. *arXiv preprint arXiv:2412.04449*, 2024.
- [20] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. *arXiv preprint arXiv:2503.22020*, 2025.