GazeProjectionVLA: Multi-View Gaze-Guided Attention for Humanoid Robot Manipulation

The 6th Military Robotics Society Annual Conference

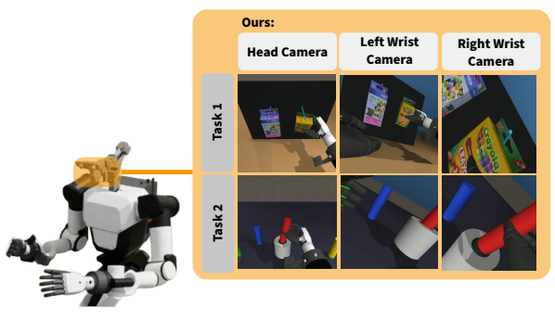
YoungChae Son, DaeKwan Ko, YoonJi Choi, JungWoo Lee, Soo-Chul Lim*

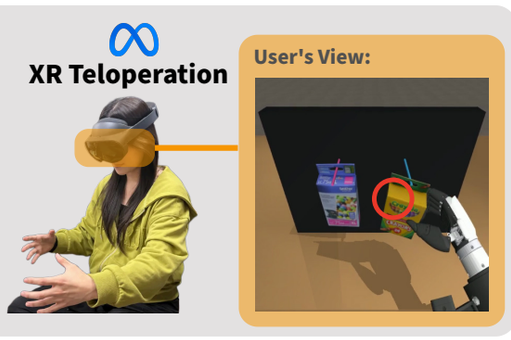
Dongguk University

Background

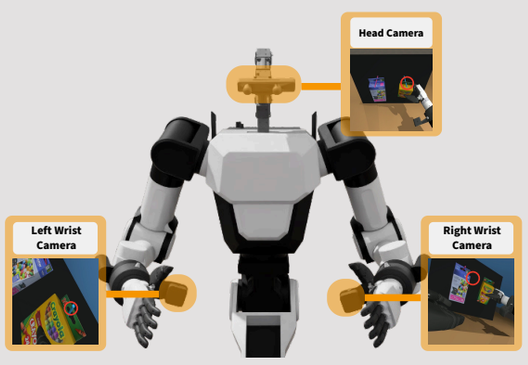
- Recent Vision-Language-Action (VLA) models use **multi-view inputs from head and wrist cameras** for manipulation. However, since each camera views the same target from a different viewpoint and scale, existing models must implicitly learn cross-view spatial correspondence and task-relevant visual regions.
- In contrast, humans **selectively focus their gaze on task-critical regions** during object manipulation, such as the target object, grasp point, and upcoming contact area, which lets them efficiently gather the visual information needed for the next action.
- Therefore, multi-view VLA policies for humanoid robots can benefit from human gaze supervision that provides geometrically consistent attention targets across camera views.

Introduction





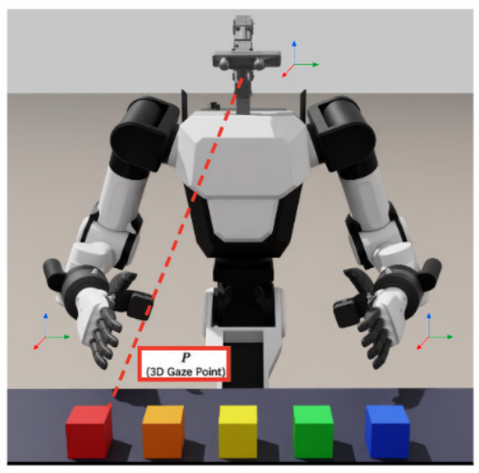
Not only the head view but also the wrist cameras contain important information.

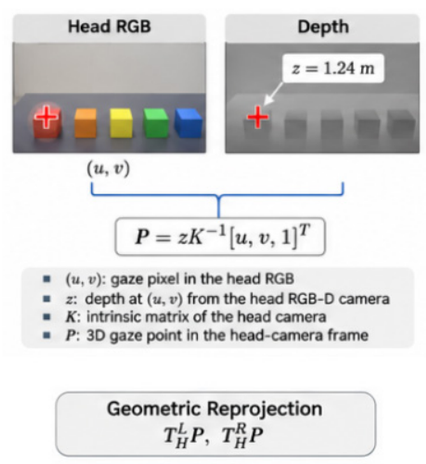


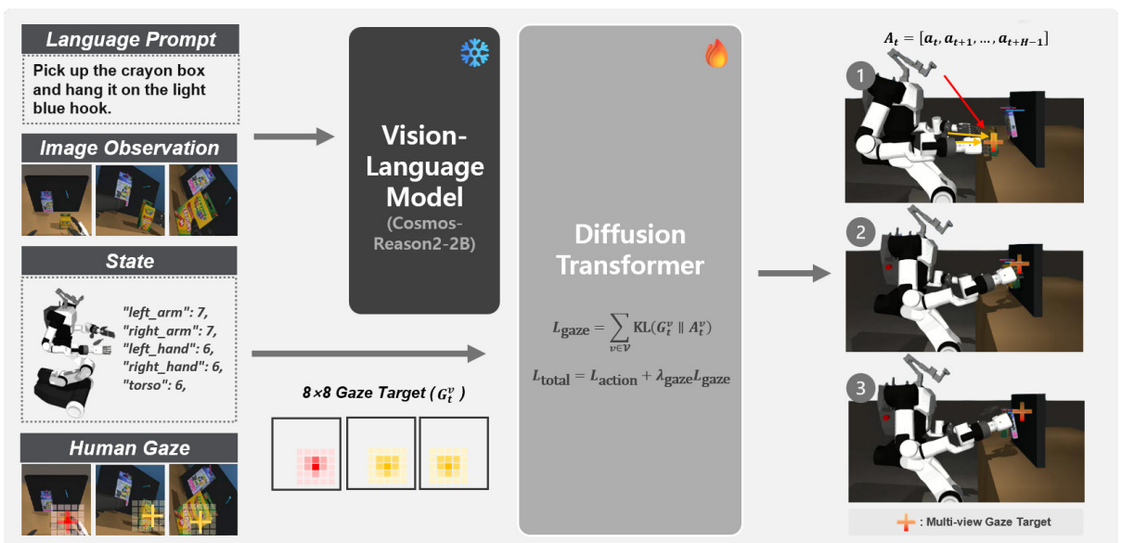
- **Robot:** Rainbow Robotics RB-Y1 , 6-DoF Inspire RH56F1 dexterous hand
- **Dataset size:** 25 episodes per task
- **State:** 32-dim joint-state vector (14 arm + 6 body + 12 hand DoF)
- **Visual observation:** head-mounted ZED Mini stereo camera + left/right wrist Intel RealSense D405 cameras (RGB/depth, 640×480, 30 Hz)
- **Gaze:** eye tracking via Meta Quest Pro

- **GazeProjectionVLA** leverages human gaze as training-time supervision by geometrically projecting gaze targets from the head view onto both wrist views.
- A gaze-guided attention alignment loss encourages the VLA to attend to task-relevant regions across all three camera views, **without requiring gaze or depth inputs at inference time.**

Method







Cross-Attention of the Diffusion Action Head Jointly Learning Gaze Information

Multi-View Gaze Projection

- Using a Meta Quest Pro, we collect the demonstrator's gaze during VR teleoperation.
- The gaze location is lifted into 3D space using the RGB-D data from the robot's head camera.
- Using the robot's forward kinematics and camera calibration, the 3D gaze point is projected onto the left and right wrist views.

Token-Level Gaze Supervision

- The gaze point in each view is converted into a Gaussian distribution over the visual token grid. This yields a separate gaze target for each of the head, left-wrist, and right-wrist views.
- Cross-attention is extracted from the first action query of the GR00T N1.7 action head.
- The visual attention distribution is aligned with the projected human gaze distribution using a KL divergence loss.
- The gaze alignment loss is applied only to valid camera views.

Training and Inference

The gaze alignment loss is added to the original GR00T N1.7 flow-matching loss for fine-tuning:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{flow}} + \lambda_{\text{gaze}}\, \mathcal{L}_{\text{gaze}}$$

Training: uses multi-view RGB, language instructions, robot state, and gaze-based auxiliary supervision.

Inference: generates action chunks from the same inputs as the original VLA, with no gaze or depth input.

Results

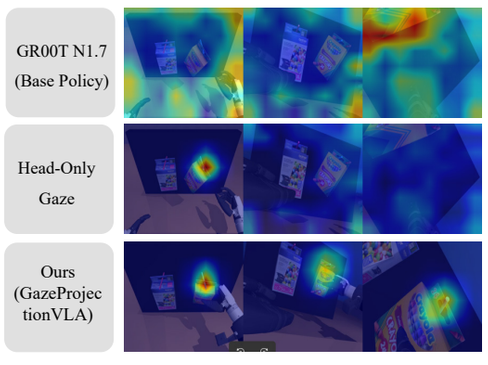


Table. 1 Task 1 success rates under different peg-radius conditions.

Model	*R*=3 (mm)	*R*=3.5 (mm)	*R*=3.8 (mm)
No Gaze	60% (12/20)	50% (5/10)	0% (0/10)
Head-Only Gaze	75% (15/20)	70% (7/10)	**70% (7/10)**
Ours (GazeProjectionVLA)	**95% (19/20)**	**80% (8/10)**	**70% (7/10)**

Table.2 Task success rates on the main manipulation tasks.

Task	No Gaze	Head-Only Gaze	Ours
Task 1 (Crayon box)	60% (12/20)	75% (15/20)	**95% (19/20)**
Task 2 (Red cylinder)	75% (15/20)	70% (14/20)	**90% (18/20)**

Table.3 Success rates under out-of-distribution conditions.

Task	No Gaze	Head-Only Gaze	Ours
Task 1 (Crayon box)	50% (5/10)	50% (5/10)	**70% (7/10)**
Task 2 (Red cylinder)	60% (6/10)	60% (6/10)	**70% (7/10)**

*OOD: Cluttered environment with five or more distractor objects.