BRAINVLA: TOWARDS VLA-DRIVEN HIGH-LEVEL EMBODIED PLANNING FOR MOTOR BRAIN-COMPUTER INTERFACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Intracortical Brain-Computer Interfaces (iBCIs) restore motor function for paralyzed patients by decoding neural signals into control commands. However, current decoding models lack environmental awareness and require users to continuously control spatial trajectories, causing cognitive fatigue and compounding control errors. To address this limitation, we introduce BrainVLA, a hybrid-intelligence framework that integrates iBCIs with pre-trained Vision-Language-Action (VLA) models by leveraging neural signals as high-level planning priors. BrainVLA extracts intent tokens from neural activity during motor preparation and maps them into the VLA's semantic space. This integration establishes a control hierarchy in which neural signals dictate high-level goals, guiding the VLA to leverage visual context for autonomous fine-grained action generation. Experiments on the maze reach task demonstrate that BrainVLA achieves a 31.6% higher success rate than classical neural decoders on average. It outperforms the VLA model π_0 by 17.7% in unseen scenarios, highlighting its impressive generalizability. Our approach establishes a new embodied iBCI paradigm where robotic agents harness cognitive abilities from biological systems.

1 INTRODUCTION

Intracortical brain-computer interfaces (iBCIs) have demonstrated remarkable success in restoring motor function for paralyzed patients by reconstructing a direct communication pathway between the brain and external devices (Lebedev & Nicolelis, 2017; Oby et al., 2020). By using implanted electrodes to record neural population activity from the motor cortex, iBCIs translate neural signals into control commands for external devices through decoding algorithms, including Kalman filters (Gilja et al., 2012), recurrent neural networks (Sussillo et al., 2012), and transformer-based architectures (Ye & Pandarinath, 2021). Recently, iBCIs have achieved impressive performance across a wide range of applications, enabling paralyzed patients to control robotic arms (Wessberg et al., 2000; Velliste et al., 2008; Hochberg et al., 2012; Flesher et al., 2021), navigate computer cursors (Kennedy et al., 2000; Hochberg et al., 2006; Pandarinath et al., 2017), and perform handwriting (Willett et al., 2021; Xu et al., 2025).

Despite these advances, traditional iBCIs face two fundamental challenges (Fig. 1, a). First, they impose a high cognitive load on users, who must continuously manage and correct spatial trajectories (Lee et al., 2025). Studies have shown that neural cursor trajectories exhibit frequent deviations from optimal paths (Simeral et al., 2011; Gilja et al., 2012). Second, the iBCI decoder operates solely on neural signals, detached from any awareness of the physical environment, preventing iBCIs from leveraging contextual information that could otherwise assist control; incorporating sensory (Flesher et al., 2021) and visual (Muelling et al., 2017) information from the environment has been shown to significantly improve robotic arm control. Recent advances have begun to address these challenges by introducing AI copilots that identify target objects in the scene and autonomously guide the robotic arm toward the target (Lee et al., 2025). However, this approach relies on a limited, predefined set of targets and cannot generalize to real-world environments. To overcome these limitations, iBCIs must move beyond traditional decoding paradigms toward interactive systems that leverage embodied AI to perceive, reason, and interact with the physical world.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

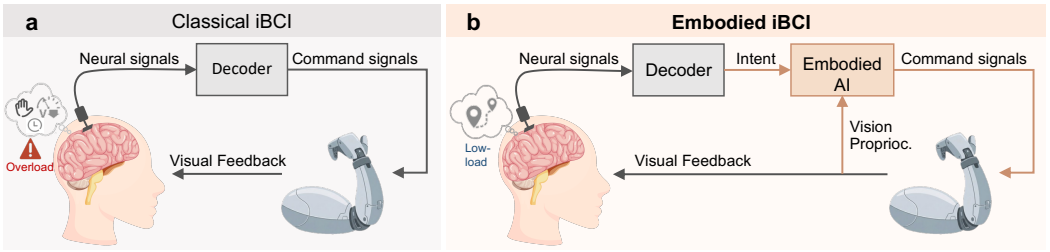

Figure 1: Comparison between classical and embodied iBCIs. Classical iBCIs decode neural signals into control commands for robotic arm, causing high cognitive burden. Differently, our proposed embodied iBCIs utilize neural signals to extract high-level intent priors, which are then fused with visual information by an embodied AI model to generate robotic movements.

Concurrently, the field of embodied AI has advanced rapidly with the development of vision–language–action (VLA) models (Kim et al., 2024; Zitkovich et al., 2023; Black et al., 2024; Bjorck et al., 2025; Li et al., 2026), which unify visual perception, language understanding, and robotic manipulation within a unified framework. Pre-trained on large-scale robotic demonstration datasets, VLAs have demonstrated remarkable capabilities in complex real-world manipulation tasks. Their robust physical execution makes them ideally suited to serve as intelligent copilots for iBCIs. However, existing VLAs are primarily driven by textual and visual prompts, lacking the mechanism to incorporate human intent for brain decoding.

To bridge this gap, we propose BrainVLA, a hybrid iBCI decoding framework that integrates pre-trained VLA models with neural decoding for embodied iBCIs (Fig. 1, b). Unlike classical iBCI paradigms that decode detailed kinematics, BrainVLA reframes the neural decoder as an intent planner. In this architecture, neural signals provide high-level plan with coarse spatial paths, while the VLA conducts low-level motor execution and environmental interaction. This design is grounded on our finding that preparation-phase neural activity in the motor cortex encodes rich path-planning information beyond the target position alone (Fig. 4). Building on this insight, we design a neural intent encoder to extract compact intent embeddings that capture the spatial planning of the upcoming movement. A cross-modal projection module maps neural embeddings into the VLM token space, enabling the VLA model to jointly attend to neural intent, visual observations, and task prompts for motor command generation. To ensure that the VLA correctly interprets neural tokens into spatial semantics, we design a spatial grounding loss that supervises fused neural tokens to predict relative direction vectors toward key waypoints, anchoring neural priors as spatial maps. To evaluate BrainVLA, we adapt the maze task (Pei et al., 2021) into a neural-planning and robotic-execution paradigm and construct a physical workspace where a robotic arm captures visual observations and executes reaching. The main contributions of this work are:

- We propose BrainVLA, the first embodied iBCI framework that introduces neural signals as high-level intent priors into a pre-trained VLA model, shifting the iBCI paradigm from kinematic decoding to neural intent planning.
- We design a neural intent encoder that maps preparatory neural activity into compact intent tokens aligned with the VLA model's semantic space, enabling action generation jointly conditioned on neural intent and visual observations.
- We demonstrate that BrainVLA achieves bi-directional empowerment through a hybrid-intelligence paradigm: maze reaching experiments demonstrate that neural intent priors improve π_0 performance by 17.7% in unseen scenarios, while the VLA model compensates for neural decoding errors, achieving a 31.6% higher than conventional decoders on average.

2 Related Work

2.1 Neural decoding for motor control

Traditional iBCIs have primarily treated neural decoding as a regression problem, translating spiking activity or local field potentials into low-level kinematics such as direction (Taylor et al., 2002), velocity (Hochberg et al., 2012; Flesher et al., 2021) and position (Hochberg et al., 2006). Early studies relied on classical machine learning algorithms like Kalman filters (Hochberg et al., 2012; Gilja et al., 2012) and support vector machine (Wessberg et al., 2000). Recently, deep learning approaches have significantly enhanced decoding performance. Multilayer perceptrons (Wessberg et al., 2000; Willsey et al., 2022), recurrent neural network (Sussillo et al., 2012), gated recurrent unit (GRU) (Wilson et al., 2020), Long Short-Term Memory (Willett et al., 2021), and Transformers (Ye & Pandarinath, 2021; Ye et al., 2023) have been widely adopted. However, current methods work under a paradigm that demands the brain to manage precisely low-level kinematics (e.g., spatial coordinates) during motor imagery, imposing a high cognitive burden. Incorporating AI assistance is therefore essential to alleviate this burden without sacrificing performance.

2.2 Vision-language-action models

Embodied AI has advanced rapidly with VLA models, which leverage multimodal pretraining to translate visual perception and language instructions into continuous motor actions, establishing a cornerstone for generalizable robotic learning. Early VLA models (Zitkovich et al., 2023; Kim et al., 2024) adapted VLMs to robotic control by discretizing continuous actions into tokens for autoregressive prediction. To realize smooth control, diffusion-based approaches (Chi et al., 2025; Zhu et al., 2025) model action generation as an iterative denoising process from Gaussian noise, but induce high inference latency. Recently, flow-matching architectures, such as π serials (Black et al., 2024; Intelligence et al., 2025) enabled real-time robotic execution with low inference latency. This capability offers a complement to iBCIs: VLA models lack access to the user's neural activity, whereas iBCIs access this signal but struggle to decode it robustly. Therefore, we propose a bio-hybrid intelligence paradigm: artificial systems inherit high-level planning and intention from the brain, while the brain utilizes the movement execution capability of Embodied AI systems.

3 BrainVLA Framework

BrainVLA treats preparatory motor cortical activity as a high-level spatial plan and leaves low-level execution to a pre-trained VLA. This design aligns with the dynamical systems perspective of cortical control, where neural activity does not directly encode movement parameters but generates a basis set of patterns for movements (Churchland et al., 2012; Shenoy et al., 2013; Vyas et al., 2020; Jiang et al., 2025). As illustrated in Fig. 2a, BrainVLA establishes this neural-to-execution mapping through three components: (1) a neural encoder that extracts compact intent tokens from preparation-phase neural activity (Sec. 3.1); (2) a cross-modal projection module that maps these tokens into the VLA token space for joint attention with vision and prompt text(Sec. 3.2); (3) a downstream action expert that interprets the fused tokens and produces smooth robot actions.

3.1 Neural Encoder

The neural encoder (Fig. 2b) maps neural spiking activity during the preparation phase into a compact set of intent tokens. These tokens encode a coarse spatial plan of the upcoming reach: two intermediate *keypoints* plus the target (Fig. 4a). A keypoint is a turning waypoint on the behavioral trajectory where the hand detours around a barrier, as opposed to the final goal. Together they form a piecewise path start $\rightarrow \mathbf{k}_1 \rightarrow \mathbf{k}_2 \rightarrow$ target. We extract the two keypoints from each behavioral trajectory and represent each as a $[\sin, \cos]$ direction (Appendix A). The encoder predicts these directions rather than a full kinematic trajectory.

Causal temporal encoding. Let $\mathbf{S} \in \mathbb{R}^{T \times N_{\text{neuron}}}$ denote spike counts in 20 ms bins over the T-bin preparatory period from N_{neuron} neurons, with padding masks for variable-length trials. Each bin is embedded into $\mathbb{R}^{d_{\text{neural}}}$ by a learned projection with sinusoidal positional encodings. A causal

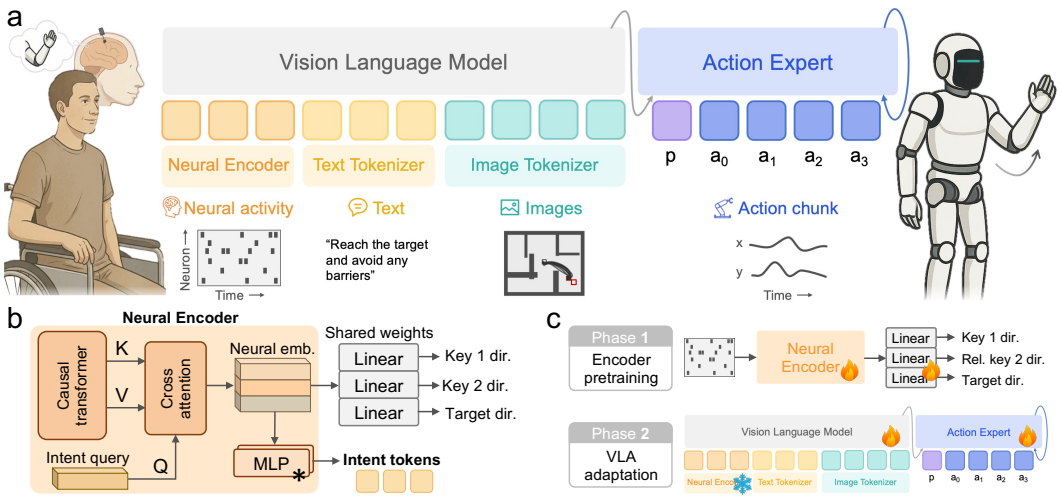

Figure 2: Overview of the BrainVLA. (a) BrainVLA architecture. Motor cortical activity, wrist-view images, and text are jointly processed through corresponding tokenizer and VLM, and then concatenated with state tokens into the action expert model to generate robot actions. (b) Network structure of the neural encoder module. (c) Two-stage training paradigm of BrainVLA.

Transformer then produces $\mathbf{M} \in \mathbb{R}^{T \times d_{\text{neural}}}$, capturing the temporal dynamics of cortical preparation without access to future activity.

Intent token extraction. The preparatory period varies in time length across trials. Attending over the entire period would spread attention onto bins unrelated to the planned reach. Studies have shown that preparatory activity converge to a movement-specific state and become stable before the go cue (Churchland et al., 2010; Shenoy et al., 2013; Even-Chen et al., 2019). Thus, we read out from the last $N = 32$ timesteps of $\mathbf{M}$, denoted $\mathbf{T}_{\text{gather}} \in \mathbb{R}^{N \times d_{\text{neural}}}$. This aligns trials of different delay lengths to the go cue while preserving information from earlier neural activity. To capture a compact sequential trajectory plan, we introduce $K = 6$ learnable intent queries, $\mathbf{Q}_{\text{intent}} \in \mathbb{R}^{K \times d_{\text{neural}}}$, organized into three consecutive pairs: $\mathbf{Q}_{1:2}$ for the 1st keypoint, $\mathbf{Q}_{3:4}$ for the 2nd keypoint, and $\mathbf{Q}_{5:6}$ for the target and barrier context. The intent tokens $\mathbf{X} \in \mathbb{R}^{B \times K \times d_{\text{neural}}}$ are obtained through cross-attention over the gathered neural representations, $\mathbf{X} = \text{CrossAttn}(\mathbf{Q} = \mathbf{Q}_{\text{intent}}, \mathbf{K} = \mathbf{T}_{\text{gather}}, \mathbf{V} = \mathbf{T}_{\text{gather}})$. To associate each query pair with its corresponding spatial goal, we supervise it using the direction of its corresponding keypoint or target. A shared linear head ϕ decodes each direction from the corresponding query pair, $\hat{\mathbf{k}}_1 = \phi(\bar{\mathbf{X}}_{1:2}), \hat{\mathbf{k}}_2 = \phi(\bar{\mathbf{X}}_{3:4}), \hat{\mathbf{g}} = \phi(\bar{\mathbf{X}}_{5:6})$, where $\bar{\mathbf{X}}_{i:j}$ denotes the mean of queries i through j. A separate head predicts barrier presence $\hat{b}$ from $\mathbf{X}_{1:4}$. These heads are used only during encoder training (Sec. 3.3). The intent tokens $\mathbf{X}$ are then passed directly to the VLA.

3.2 BrainVLA Integration Mechanism

Neural signals convey the user's movement intention, while vision provides the scene context needed to interpret it. We therefore inject the intent tokens into the VLM, building on its pretrained capacity for vision–language alignment to relate the neural plan to the observed scene. This mechanism integrates neural signals, vision and task prompts, grounding the user's intention into the physical state of the environment.

Visual observations are first encoded by a pre-trained vision encoder to yield visual tokens $\mathbf{Z}_v \in \mathbb{R}^{N_v \times D_{model}}$. Concurrently, the global text prompt describing the main task is tokenized into text tokens $\mathbf{Z}_t \in \mathbb{R}^{N_t \times D_{model}}$. A multi-layer perceptron (MLP) projects the intent tokens to neural tokens that match the token dimension expected by the VLM (PaliGemma (Beyer et al., 2024)). The neural tokens $\mathbf{Z}_n \in \mathbb{R}^{N_n \times D_{model}}$ are concatenated with the visual and text tokens as $\mathbf{E}_{in} = [\mathbf{Z}_v; \mathbf{Z}_t; \mathbf{Z}_n] \in \mathbb{R}^{(N_v+N_t+N_n) \times D_{model}}$. This joint sequence undergoes token-to-token attention in the VLM, yielding fused multimodal representations $\mathbf{E}_{fused} = \text{VLM}(\mathbf{E}_{in})$.

Following π_0 paradigm (Black et al., 2024), we construct a suffix sequence $\mathbf{E}_{\text{suffix}} \in \mathbb{R}^{(1+H)\times D}$, including state tokens $s_t \in \mathbb{R}^D$ and H action-time tokens derived from noisy action chunk $a_t^\tau \in \mathbb{R}^{H\times D}$ and timestep τ. The concatenated sequence $[\mathbf{E}_{\text{fused}}; \mathbf{E}_{\text{suffix}}]$ is forwarded through action expert and predict velocity field $\mathbf{v}_\theta(a_t^\tau, \mathbf{o}_t) \in \mathbb{R}^{H\times D_a}$.

3.3 Training Strategy

Jointly training a randomly initialized neural encoder and the pre-trained π_0 can lead to unstable gradients and degrade the pre-learned representations. Therefore, we employ a two-stage training paradigm (Fig. 2c).

Stage 1: Neural encoder pre-training. In the first stage, the neural encoder is trained independently. Let $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$ denote the ground-truth and predicted directions, respectively, for $i \in \{k_1, k_2, g\}$, and let $b \in \{0, 1\}$ denote the barrier-presence label. The training objective is defined as

$$\mathcal{L}_{\text{S1}} = \sum_i \left(1 - \frac{\hat{\mathbf{y}}_i^\top \mathbf{y}_i}{\|\hat{\mathbf{y}}_i\|\|\mathbf{y}_i\|}\right) + \mathcal{L}_{\text{BCE}}(\hat{b}, b), \tag{1}$$

where the first term penalizes angular prediction errors and the second computes the binary cross-entropy loss for barrier prediction.

Stage 2: Neural-driven VLA adaptation. Stage 2 aligns the neural representations with the VLA while retaining its pretrained manipulation capabilities. We freeze the neural encoder, fine-tune the projection MLP and action expert, and adapt the VLM using LoRA (Hu et al., 2022). The training objective combines a flow-matching loss for action generation with a spatial grounding loss that encourages the VLA to use the neural intent tokens alongside visual observations:

$$\mathcal{L}_{\text{S2}} = \mathcal{L}_{\text{flow}} + \lambda\, \mathcal{L}_{\text{spatial}}, \qquad \lambda = 0.2. \tag{2}$$

Flow-matching loss. Given a ground-truth action chunk $\mathbf{A}_t$, noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\tau \in [0, 1]$, the noisy chunk is constructed as $\mathbf{A}_t^\tau = \tau \mathbf{A}_t + (1-\tau)\boldsymbol{\epsilon}$. The action expert is optimized to predict the corresponding target velocity $\mathbf{A}_t - \boldsymbol{\epsilon}$ through the conditional flow-matching objective (Black et al., 2024):

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{t,\tau,\boldsymbol{\epsilon}} \left\| \mathbf{v}_\theta(\mathbf{A}_t^\tau, \mathbf{o}_t) - (\mathbf{A}_t - \boldsymbol{\epsilon}) \right\|_2^2. \tag{3}$$

Spatial grounding loss. To ground neural intent in the current robot state, we extract the fused neural tokens $\mathbf{E}_{\text{neu}}$ from the VLM outputs and apply a linear prediction head Φ to predict the directions from the current end-effector (EEF) position $\mathbf{x}_t$ to each waypoint $\mathbf{p}_i$, where $i \in \{k_1, k_2, g\}$:

$$[\hat{\mathbf{p}}_{k_1}, \hat{\mathbf{p}}_{k_2}, \hat{\mathbf{p}}_g] = \Phi(\text{Pool}(\mathbf{E}_{\text{neu}})). \tag{4}$$

Let $\hat{\mathbf{d}}_i$ and $\mathbf{d}_i^*$ denote the unit vectors of $\hat{\mathbf{p}}_i - \mathbf{x}_t$ and $\mathbf{p}_i - \mathbf{x}_t$, respectively. We penalize their L_2 discrepancy, weighted by the true distance to reduce the influence of unstable direction estimates near the waypoint:

$$\mathcal{L}_{\text{spatial}} = \frac{1}{3} \sum_{i \in \{k_1, k_2, g\}} \|\mathbf{p}_i - \mathbf{x}_t\|_2 \, \|\hat{\mathbf{d}}_i - \mathbf{d}_i^*\|_2^2. \tag{5}$$

4 Experiments and Results

4.1 Experimental Setting

We evaluate BrainVLA on MC_Maze from the NLB (Pei et al., 2021), which contains spiking activity recorded from the motor cortex of a rhesus monkey performing delayed reaches. Crucially, BrainVLA receives neural activity only from the preparatory period to infer the intended path. Implementation details are provided in Appendix B.

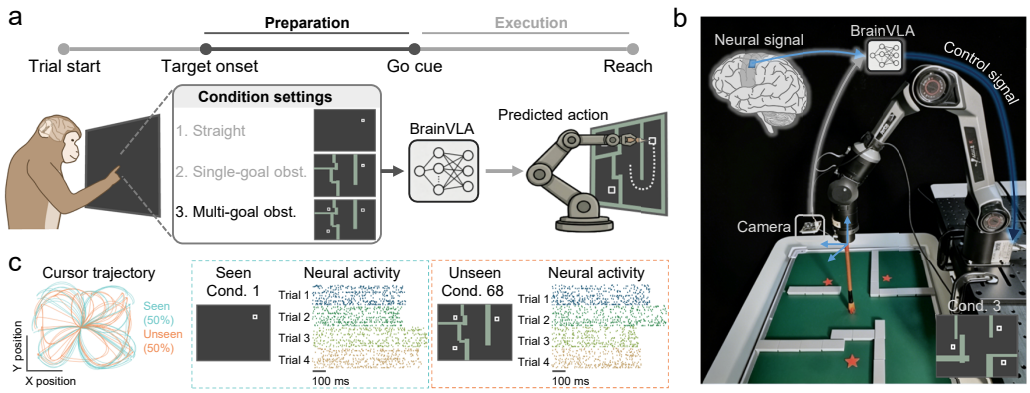

Figure 3: Experiment setting. (a) Neural-planning and robotic-execution workflow. The monkey plans a reaching path during the preparatory phase, and a go cue triggers robotic arm execution. (b) Experiment platform. The 2D maze coordinates are physically mapped onto a horizontal blackboard workspace. The robotic arm performs the task guided by neural activity recorded from monkey brain and visual observations from a camera. (c) Neural dataset. Left: dataset splitting to ensure unbiased directional representation. Right: spiking activity during preparation for two example conditions.

Experimental paradigm. The experimental workflow operates on a synchronized neural-planning and robotic-execution paradigm (Fig. 3a). After fixation, a particular configuration is presented. Experimental conditions vary by the inclusion of distractor targets and barriers. During the variable-duration preparatory period, the monkey has to determine the reachable target and plan a reach to bypass the barriers. Meanwhile, the monkey remains stationary and we utilize the neural activity during this preparatory phase. Upon the go cue, the control loop shifts to the embodiment phase, prompting a robotic arm to execute the planned reach.

Experimental setup. To bridge the gap between monkey's and robotic observations, we mapped the 2D maze coordinates onto a physical workspace, consisting of a blackboard resting on a table. Red and grey magnetic blocks are placed within this workspace to serve as targets and barriers, mirroring the spatial layout and coordinates of the original maze configuration (Fig. 3b).

Task execution is performed by a 6-DOF robotic arm (Agile, Piper), which was programmed to track the cursor trajectories from the maze dataset within the mapped physical workspace during data collection. Visual observations of the workspace are captured at 30 fps by a wrist-mounted RGB camera (640×480 resolution).

Dataset construction. To evaluate generalization, we select 54 of the 108 conditions (seen) for training by sampling half of target directions that broadly cover the directional space (Fig. 3c). The remaining 54 conditions are treated as unseen. Models are trained exclusively on the seen conditions, while robotic evaluation focuses on the unseen conditions (Appendix C). The dataset comprises $\sim$1000 trajectories. Each trajectory contains synchronized streams of: neural signals during the motor preparation phase, RGB images, and end-effector 6D poses.

Baselines. We compare BrainVLA with two categories of baselines. First, we fine-tune π_0 with the same hyper-parameters(Black et al., 2024), as an embodied AI baseline to evaluate whether neural intent improves visually guided action generation. Second, we compare against POYO (Azabou et al., 2023), NDT (Ye & Pandarinath, 2021), GRU to assess whether BrainVLA's hierarchical planning-and-execution paradigm is more effective than conventional neural decoding models. Since continuous movement trajectories cannot be decoded from preparation activity, neural decoding models use neural activity during movement execution. This serves as a system-level comparison between two BCI paradigms: embodied framework versus traditional decoding methods.

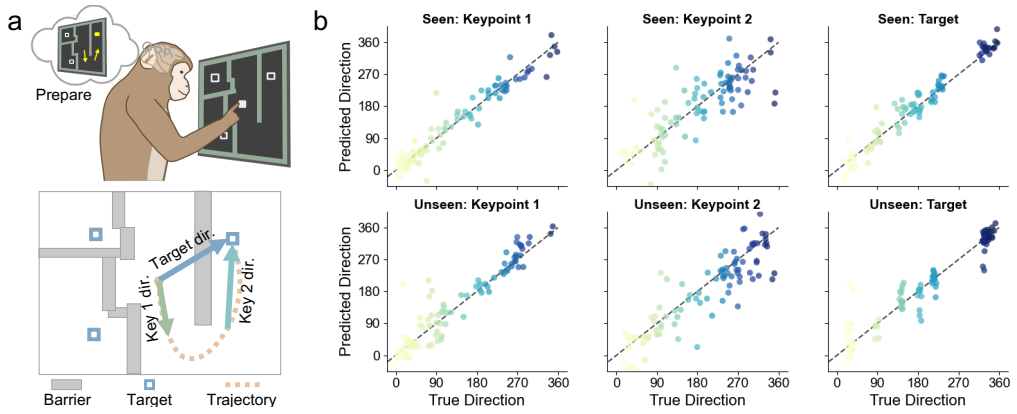

Figure 4: Neural-to-intent decoding. (a) Conceptual illustration of path-planning representation in motor cortex during preparation (top). The lower schematic defines the decoded neural intent: kinematic keypoint vectors (key 1 dir., key 2 dir.) and the target vector (Target dir.), outlining a structural path around intervening barriers. (b) Scatter plots of predicted and true angular directions (0° to 360°) for keypoint 1, keypoint 2, and the target across both seen (top) and unseen (bottom).

4.2 Main Experiment Results

Neural-to-intent decoding. BrainVLA is based on the assumption that preparatory motor cortical activity encodes the planned route rather than merely the reach target (Fig. 4a). We first decode the directions of the target and the two keypoints from spiking activity (Fig. 4b). Table 1 reports the Mean Absolute Error (MAE) and the coefficient of determination (R^2) across 191 total test trials, split into seen and unseen conditions. The target and the first keypoint are decoded with comparable accuracy (23.6° and 25.2° on unseen conditions), whereas the second keypoint is decoded less precisely (43.3°). Accuracy degrades mildly from seen to unseen conditions, indicating that the encoder generalizes to novel conditions.

Table 1: Neural decoding accuracy (MAE in degrees and R^2) for the target direction and the two path keypoints on seen and unseen conditions.

	KP1 Direction		KP2 Direction		Target Direction	
Condition	MAE ↓	R^2 ↑	MAE ↓	R^2 ↑	MAE ↓	R^2 ↑
Seen	18.3°	0.91	39.3°	0.65	19.4°	0.92
Unseen	25.2°	0.81	43.3°	0.61	23.6°	0.86
Overall	21.7°	0.86	41.3°	0.68	21.5°	0.89

To verify that these keypoints cannot be inferred from the target alone, we permute neural signals among test trials that share a target but differ in maze layout (Appendix D). Keypoint error roughly doubles (KP1: $21.7^\circ \rightarrow 49.2^\circ$; KP2: $41.3^\circ \rightarrow 67.4^\circ$; $p < 0.001$), confirming that motor cortex encodes the planned route. This rich, multi-keypoint neural intention provides instruction for integrating VLA model, replacing step-by-step coordinate commands with a robust intent prompt.

Robotic task performance. We evaluate all models on the 96 unseen test trials, spanning three contexts of increasing complexity: (1) **Straight** (N=33), which requires a direct reach toward the target; (2) **Single-goal obstacle** (N=31), which requires the robot arm to navigate around obstacles to reach a single target; (3) **Multi-goal obstacle** (N=32), which adds distractor targets to the single-goal obstacle setting.

We first compare BrainVLA with traditional neural decoders to assess the advantage of embodied iBCI over traditional kinematic decoding. Qualitatively, BrainVLA consistently produces goal-directed trajectories that navigate around barriers, while traditional neural decoders exhibit unstable, deviating paths and π_0 occasionally select incorrect routes (Fig. 5). This advantage is further demon-

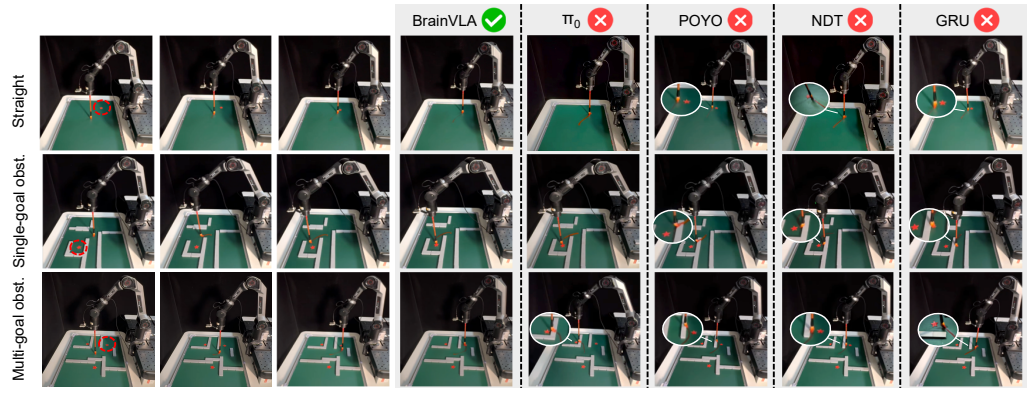

Figure 5: Qualitative results on maze reach task, compared with embodied AI (π_0) and neural decoding models (POYO, NDT, GRU). BrainVLA completes this task with higher success rate.

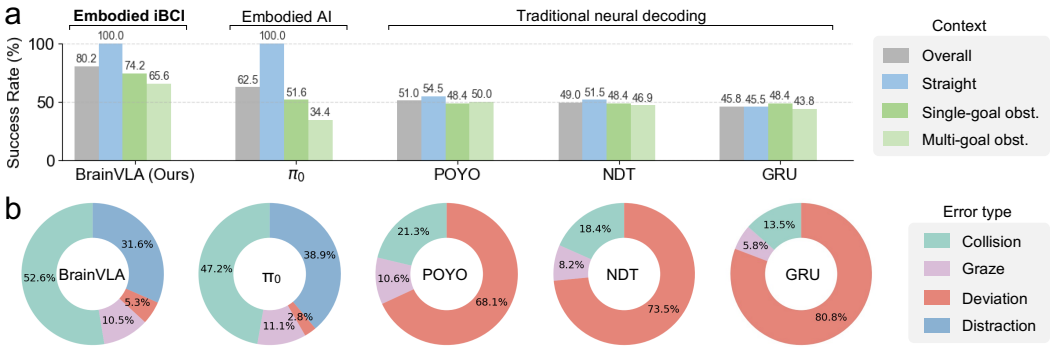

Figure 6: Quantitative evaluation and error analysis on the maze reaching task. (a) Success rate comparison across four evaluation conditions: overall (pooled across trials), straight, single-goal obstacle, and multi-goal obstacle. BrainVLA outperforms both embodied AI (π_0) and traditional continuous neural decoders (POYO, NDT, GRU) in unseen conditions, demonstrating superior generalization. (b) Distribution of error types across models.

strated by quantitative result. BrainVLA achieves an overall success rate of 80.2%, outperforming the strongest conventional neural decoder, POYO (51.0%), by 29.2% (Fig. 6a). Complete success counts and 95% Wilson confidence intervals are reported in Appendix Table 5.

Furthermore, we compare BrainVLA with π_0 to assess the contribution of neural intent to VLA performance. Both models achieve 100% success in straight trials, indicating that the VLA backbone reliably identifies the intended target and executes direct reaches when the route is unambiguous. The advantage of neural intent emerges in obstacle conditions: BrainVLA achieves 74.2% success versus 51.6% for π_0 in single-goal trials, and 65.6% versus 34.4% in multi-goal trials. Overall, BrainVLA improves the success rate by 17.7%. The increasing with scene complexity suggests that preparatory neural activity provides path-planning information that cannot be reliably inferred from vision alone, guiding BrainVLA toward the intended route in held-out configurations.

We categorize failed trials as *distraction* (moving toward a blocked distractor), *collision* (directly hitting an obstacle), *graze* (contacting an obstacle edge), or *deviation* (failing to stop around the target). Detailed criteria are provided in Appendix F. Trajectory deviation is the predominant failure mode of traditional decoders, accounting for 68.1%–88.8% of their errors (Fig. 6b). This pattern is consistent with the accumulation of low-level decoding errors over time. In contrast, deviation constitutes only 5.3% of BrainVLA's errors, which are instead dominated by collision (52.6%) and distraction (31.6%). This shift suggests that embodied iBCI largely resolves unstable low-level control. π_0 shows the same error profile but fails nearly twice as often (36 vs. 19 trials), indicating that neural intent reduces high-level route errors rather than execution errors.

4.3 Ablation Studies

Impact of textual tokenization of neural activity. To assess the benefit of projecting neural activity directly into the VLA latent space, we compare BrainVLA with a neural-to-text translation approach. We replace the intent tokens with text: the Stage-1 head decodes the three directions $(\hat{\theta}_{k_1}, \hat{\theta}_{k_2}, \hat{\theta}_g)$, which are written in radians inside the prompt (e.g., "1.2, 0.5, 3.14 rad"), and tokenized by PaliGemma's text tokenizer. Stage 2 is retrained on this prompt with the same data and hyper-parameters. On unseen obstacle trials, the success rate drops from 69.8% to 50.8% (Table 6): from 74.2% to 58.1% in single-goal trials and from 65.6% to 43.8% in multi-goal trials. Error analysis shows that text-based variant is more easily distracted by other goals and selecting the wrong target. These results suggest that discrete text bottleneck may discard rich, continuous probabilistic information in the neural preparation phase.

Impact of the auxiliary spatial grounding loss. To test whether the VLA can effectively use intent tokens without auxiliary spatial supervision, we remove the $\mathcal{L}_{spatial}$ during encoder training. Success rate in obstacle scene reduces the overall success rate from 69.8% to 50.8%, with decreases from 74.2% to 58.1% in single-goal trials and from 65.6% to 43.8% in multi-goal trials (Appendix Table 6). Error analysis reveals that omitting this loss increases *distraction* and *collision* errors. Without the spatial grounding loss, the VLA model tends to over-rely on visual features and occasionally ignores the user's high-level intent when encountering unfamiliar spatial obstacles. These results highlight that $\mathcal{L}_{spatial}$ is critical for maintaining robust grounding between the biological controller and the embodied agent.

4.4 Information controls and oracle upper bound

Shuffled intent tokens. To verify that BrainVLA benefits from neural intent tokens rather than the auxiliary supervision $\mathcal{L}_{spatial}$, we randomly permute neural activity across trials during training and evaluation, destroying their correspondence with the actual behavior. Success rate on unseen obstacle trials falls to 41.3% (Appendix Table 6), comparable to the performance of π_0 (42.9%) without neural input. This confirms that the advantage of BrainVLA is driven by valid neural guidance.

Oracle upper bound. To quantify how strongly the decoding error limits performance, we train an oracle variant in which the six neural tokens are replaced by the six ground-truth $[\sin, \cos]$ values of $\mathbf{k}_1$, $\mathbf{k}_2$, and $\mathbf{g}$, passed through the same projection MLP and Stage 2 training. Trained and evaluated with ground-truth directions, this variant reaches 85.7% on unseen obstacle trials (Table 6). Compared with BrainVLA. The 15.9 point gap is attributable to the $\sim 25^\circ$ decoding error (Sec. 4.2), which accounts for most of the remaining collision and distraction failures, whereas the residual gap to 100% reflects the VLA's own execution problems (e.g., limited scene understanding), which even a perfect intent signal cannot resolve.

5 Discussion

We introduce BrainVLA, a novel paradigm for iBCIs that shifts the neural decoding objective from low-level kinematics to high-level intention planning. By extracting the intended path from neural activity during preparation, BrainVLA fuses this neural guidance with visual observations to generate robot actions. Evaluations across complex reaching task demonstrate that BrainVLA achieves an overall success rate of 80.2%, outperforming state-of-the-art embodied AI and traditional neural decoders, particularly in unseen scenes containing obstacles and distractor targets. Furthermore, error and ablation analyses confirm that our neural tokens preserve graded information about planned paths. Thus, BrainVLA learns neural representations that connect motor intention with perception and action rather than simply adding neural signals as another input modality. Our work bridges the gap between biological intent and embodied robotics, establishing a foundation for more resilient, context-aware, and natural neural-driven control systems. However, evaluation was conducted offline using prerecorded neural activity. Future work will evaluate BrainVLA in an online setting and extend the framework to more complex manipulation tasks.

AI USE STATEMENT

In this work, AI tools were used only to edit the manuscript for clarity, readability, and grammatical correctness. All scientifc content (research design, experiments, and analysis) was produced entirely by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses the publicly available MC_Maze dataset from the Neural Latents Benchmark (Pei et al., 2021) collects robot execution data on our physical platform. No animal or human experiments were conducted.

REPRODUCIBILITY STATEMENT

The neural data are public (Pei et al., 2021), and the VLA backbone is the open-source π_0 (Black et al., 2024). The detailed architecture, training hyperparameters, and the prompt template in Appendix B, dataset splits and the physical workspace mapping in Appendix C. All datasets used in this work are publiclyavailable,and we will release the source code upon acceptance.

REFERENCES

Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh Nachimuthu, Michael Mendelson, Blake Richards, Matthew Perich, Guillaume Lajoie, and Eva Dyer. A unified, scalable framework for neural population decoding. *Advances in Neural Information Processing Systems*, 36:44937–44956, 2023.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. Paligemma: A versatile 3b vlm for transfer, 2024. URL `https://arxiv.org/abs/2407.07726`.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π0: A vision-language-action flow model for general robot control. *arXiv preprint ARXIV.2410.24164*, 2024.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

Mark M Churchland, Byron M Yu, John P Cunningham, Leo P Sugrue, Marlene R Cohen, Greg S Corrado, William T Newsome, Andrew M Clark, Paymon Hosseini, Benjamin B Scott, et al. Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nature neuroscience*, 13(3):369–378, 2010.

Mark M Churchland, John P Cunningham, Matthew T Kaufman, Justin D Foster, Paul Nuyujukian, Stephen I Ryu, and Krishna V Shenoy. Neural population dynamics during reaching. *Nature*, 487 (7405):51–56, 2012.

Nir Even-Chen, Blue Sheffer, Saurabh Vyas, Stephen I Ryu, and Krishna V Shenoy. Structure and variability of delay activity in premotor cortex. *PLOS Computational Biology*, 15(2):e1006808, 2019.

Sharlene N Flesher, John E Downey, Jeffrey M Weiss, Christopher L Hughes, Angelica J Herrera, Elizabeth C Tyler-Kabara, Michael L Boninger, Jennifer L Collinger, and Robert A Gaunt. A brain-computer interface that evokes tactile sensations improves robotic arm control. *Science*, 372(6544):831–836, 2021.

Vikash Gilja, Paul Nuyujukian, Cindy A Chestek, John P Cunningham, Byron M Yu, Joline M Fan, Mark M Churchland, Matthew T Kaufman, Jonathan C Kao, Stephen I Ryu, et al. A high-performance neural prosthesis enabled by control algorithm design. *Nature neuroscience*, 15(12): 1752–1757, 2012.

Leigh R Hochberg, Mijail D Serruya, Gerhard M Friehs, Jon A Mukand, Maryam Saleh, Abraham H Caplan, Almut Branner, David Chen, Richard D Penn, and John P Donoghue. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature*, 442(7099):164–171, 2006.

Leigh R Hochberg, Daniel Bacher, Beata Jarosiewicz, Nicolas Y Masse, John D Simeral, Joern Vogel, Sami Haddadin, Jie Liu, Sydney S Cash, Patrick Van Der Smagt, et al. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature*, 485(7398):372–375, 2012.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Hongru Jiang, Xiangdong Bu, Qianqian Shi, Zhiyan Zheng, Di Zhu, Huajin Tang, and Yao Chen. Rethink the motor cortical control via the experiment-analysis-model flywheel: an overview. *Med-X*, 3(1):15, 2025.

Philip R Kennedy, Roy AE Bakay, Melody M Moore, Kim Adams, and John Goldwaithe. Direct control of a computer from the human central nervous system. *Ieee Transactions on Rehabilitation Engineering*, 8(2):198–202, 2000.

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.

Mikhail A Lebedev and Miguel AL Nicolelis. Brain-machine interfaces: from basic science to neuroprostheses and neurorehabilitation. *Physiological reviews*, 97(2):767–837, 2017.

Johannes Y Lee, Sangjoon Lee, Abhishek Mishra, Xu Yan, Brandon McMahan, Brent Gaisford, Charles Kobashigawa, Mike Qu, Chang Xie, and Jonathan C Kao. Brain–computer interface control with artificial intelligence copilots. *Nature machine intelligence*, 7(9):1510–1523, 2025.

Yang Li, Zhaxizhuoma, Hongru Jiang, Junjie Xia, Hongquan Zhang, Jinda Du, Yunsong Zhou, Jia Zeng, Ce Hao, Jieji Ren, Qiaojun Yu, Cewu Lu, Yu Qiao, and Jiangmiao Pang. Forcevla2: Unleashing hybrid force-position control with force awareness for contact-rich manipulation, 2026. URL https://arxiv.org/abs/2603.15169.

Katharina Muelling, Arun Venkatraman, Jean-Sebastien Valois, John E Downey, Jeffrey Weiss, Shervin Javdani, Martial Hebert, Andrew B Schwartz, Jennifer L Collinger, and J Andrew Bagnell. Autonomy infused teleoperation with application to brain computer interface controlled manipulation. *Autonomous Robots*, 41(6):1401–1422, 2017.

Emily R. Oby, Jay A. Hennig, Aaron P. Batista, Byron M. Yu, and Steven M. Chase. *Intracortical Brain–Machine Interfaces*, pp. 185–221. Springer International Publishing, Cham, 2020.

Chethan Pandarinath, Paul Nuyujukian, Christine H Blabe, Brittany L Sorice, Jad Saab, Francis R Willett, Leigh R Hochberg, Krishna V Shenoy, and Jaimie M Henderson. High performance communication by people with paralysis using an intracortical brain-computer interface. *eLife*, 6: e18554, 2017.

Felix Pei, Joel Ye, David Zoltowski, Anqi Wu, Raeed H Chowdhury, Hansem Sohn, Joseph E O'Doherty, Krishna V Shenoy, Matthew T Kaufman, Mark Churchland, et al. Neural latents benchmark'21: evaluating latent variable models of neural population activity. *arXiv preprint arXiv:2109.04463*, 2021.

Krishna V Shenoy, Maneesh Sahani, and Mark M Churchland. Cortical control of arm movements: a dynamical systems perspective. *Annual review of neuroscience*, 36(1):337–359, 2013.

JD Simeral, Sung-Phil Kim, MJ Black, JP Donoghue, and LR Hochberg. Neural control of cursor trajectory and click by a human with tetraplegia 1000 days after implant of an intracortical microelectrode array. *Journal of neural engineering*, 8(2):025027, 2011.

David Sussillo, Paul Nuyujukian, Joline M Fan, Jonathan C Kao, Sergey D Stavisky, Stephen Ryu, and Krishna Shenoy. A recurrent neural network for closed-loop intracortical brain–machine interface decoders. *Journal of neural engineering*, 9(2):026027, 2012.

Dawn M Taylor, Stephen I Helms Tillery, and Andrew B Schwartz. Direct cortical control of 3d neuroprosthetic devices. *science*, 296(5574):1829–1832, 2002.

Meel Velliste, Sagi Perel, M Chance Spalding, Andrew S Whitford, and Andrew B Schwartz. Cortical control of a prosthetic arm for self-feeding. *Nature*, 453(7198):1098–1101, 2008.

Saurabh Vyas, Matthew D Golub, David Sussillo, and Krishna V Shenoy. Computation through neural population dynamics. *Annual review of neuroscience*, 43(1):249–275, 2020.

Johan Wessberg, Christopher R Stambaugh, Jerald D Kralik, Pamela D Beck, Mark Laubach, John K Chapin, Jung Kim, S James Biggs, Mandayam A Srinivasan, and Miguel AL Nicolelis. Real-time prediction of hand trajectory by ensembles of cortical neurons in primates. *Nature*, 408(6810): 361–365, 2000.

Francis R Willett, Donald T Avansino, Leigh R Hochberg, Jaimie M Henderson, and Krishna V Shenoy. High-performance brain-to-text communication via handwriting. *Nature*, 593(7858): 249–254, 2021.

Matthew S Willsey, Samuel R Nason-Tomaszewski, Scott R Ensel, Hisham Temmar, Matthew J Mender, Joseph T Costello, Parag G Patil, and Cynthia A Chestek. Real-time brain-machine interface in non-human primates achieves high-velocity prosthetic finger movements using a shallow feedforward neural network decoder. *Nature Communications*, 13(1):6899, 2022.

Guy H Wilson, Sergey D Stavisky, Francis R Willett, Donald T Avansino, Jessica N Kelemen, Leigh R Hochberg, Jaimie M Henderson, Shaul Druckmann, and Krishna V Shenoy. Decoding spoken english from intracortical electrode arrays in dorsal precentral gyrus. *Journal of neural engineering*, 17(6):066007, 2020.

Guangxiang Xu, Zebin Wang, Kedi Xu, Junming Zhu, Jianmin Zhang, Yueming Wang, and Yaoyao Hao. Decoding handwriting trajectories from intracortical brain signals for brain-to-text communication. *Advanced Science*, 12(40):e05492, 2025.

Joel Ye and Chethan Pandarinath. Representation learning for neural population activity with neural data transformers. *Neurons, Behavior, Data Analysis, and Theory*, 5(3):1–18, 2021.

Joel Ye, Jennifer Collinger, Leila Wehbe, and Robert Gaunt. Neural data transformer 2: multi-context pretraining for neural spiking activity. *Advances in Neural Information Processing Systems*, 36:80352–80374, 2023.

Minjie Zhu, Yichen Zhu, Jinming Li, Junjie Wen, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, Yaxin Peng, Feifei Feng, et al. Scaling diffusion policy in transformer to 1 billion parameters for robotic manipulation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10838–10845. IEEE, 2025.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.

A KEYPOINT EXTRACTION

Given a behavioral trajectory $\mathbf{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_T\}$ with velocities $\{\mathbf{v}_1, \ldots, \mathbf{v}_T\}$, where $\mathbf{p}_t \in \mathbb{R}^2$ is the hand position at time t, let $\mathbf{u} = (\mathbf{p}_T - \mathbf{p}_1)/\|\mathbf{p}_T - \mathbf{p}_1\|_2$ denote the unit vector from start to target and $\mathbf{u}_\perp$ its normal. We project the velocity onto the normal, $v_t^\perp = \mathbf{v}_t^\top \mathbf{u}_\perp$, and take its two extrema,

$$t_a = \arg\min_t v_t^\perp, \qquad t_b = \arg\max_t v_t^\perp, \tag{6}$$

which mark the moments at which the hand deviates from and returns to the direct path with the largest lateral velocity. The two keypoints are $\mathbf{p}_{k_1} = \mathbf{P}[\min(t_a, t_b)]$ and $\mathbf{p}_{k_2} = \mathbf{P}[\max(t_a, t_b)]$, ordered in time.

The three supervision directions are $\mathbf{k}_1 = \mathbf{p}_{k_1} - \mathbf{p}_1$ (departure), $\mathbf{k}_2 = \mathbf{p}_T - \mathbf{p}_{k_2}$ (approach), and $\mathbf{g} = \mathbf{p}_T - \mathbf{p}_1$ (target), each normalized as $[\sin\theta, \cos\theta]$ to preserve angular continuity.

B IMPLEMENTATION DETAILS

Neural encoder. Table 2 lists the encoder configuration. Trials shorter than the maximum length of preparation ($T = 77$) bins are left-zero-padded and masked during attention.

Table 2: **Neural encoder architecture.**

Parameter	Value
Number of Neurons	182
Hidden Dimension	64
Number of Layers	2
Number of Heads	2
Dropout	0.25

Training setup. BrainVLA is trained on 2 A100 GPUs with a batch size of 32 for 18,000 steps, which takes about 3.5 hours. Table 3 lists the training settings. At test time, BrainVLA runs at 30 Hz on a single RTX 4090 GPU with an action chunk size of 12, which is fast enough for real-time control.

Table 3: **Training details of BrainVLA.**

Parameter	Value
Learning Rate Schedule	Cosine Decay
Optimizer	AdamW
EMA Decay	0.99
Random Seed	42
Batch Size	32
Training Steps	18,000
Vision token dropout	20%
Projection MLP	$64 \to 1024 \to 2048$, GELU

Prompt template. The neural intent tokens are placed inside a fixed prompt template, between a system instruction and a task instruction. The template starts with a system instruction, followed by a task instruction that describes the goal. The six neural intent tokens, two tokens for each of the two keypoints and two for the target, are then inserted at a tag marking the start of the neural path, followed by a matching tag:

```
prompt_prefix = (
"<System> You are a brain-controlled agent. "
"Execute the trajectory following the neural path. </System>\n"
"<Task> Reach red target and avoid any grey objects </Task>\n"
"<Neural_Path>"
```

```
)
neural_tokens = [key11, key12, key21, key22, targ1, targ2]
prompt_suffix = "</Neural_Path>"
```

C Dataset and Platform Details

Dataset split. The training set contains 985 trials from 54 selected conditions. The validation and test sets each contain 191 trials sampled from all 108 conditions, enabling separate evaluation on seen and unseen conditions. Robot arm data, consisting of camera images paired with actions, are collected only for the training split to fine-tune the VLA model. No images or actions from unseen conditions are used in training. The validation split contains spiking data and direction labels only and is used to select the neural encoder and the neural decoding baselines. The test split is evaluated directly on the physical robot.

Coordinate mapping. The maze dataset gives the target position and the obstacle center, width, and height in millimeters. We place the physical target and obstacle blocks at these same millimeter coordinates on the workspace, with no extra scale factor between the dataset coordinates and the physical workspace.

D Target-Matched Shuffle Control

To verify that the motor cortex encodes keypoints, we group test trials by the target and permute the neural signals among trials. Each trial is paired with neural data recorded for the same target but a different maze layout. We repeat the permutation $1{,}000$ times and report the mean and 95% interval of the resulting keypoint MAE (Table 4).

Table 4: **Target-matched shuffle control.** Neural inputs are permuted among trials with the same target. Shuffled values are the mean and 95% interval over $1{,}000$ permutations; all $p < 0.001$.

Split	Keypoint	Aligned MAE	Shuffled MAE [95% CI]	Δ
Overall (N=191)	KP1	21.7°	49.2° [47.2, 50.9]	$+27.4^\circ$
	KP2	41.3°	67.4° [64.8, 70.0]	$+26.2^\circ$
Seen (N=95)	KP1	18.3°	47.1° [44.7, 49.9]	$+28.9^\circ$
	KP2	39.3°	67.5° [64.4, 70.2]	$+28.2^\circ$
Unseen (N=96)	KP1	25.2°	51.2° [48.4, 53.9]	$+26.0^\circ$
	KP2	43.3°	67.2° [62.6, 72.0]	$+24.0^\circ$

E Robotic Evaluation

Table 5 reports the success counts and binomial uncertainty for the robotic evaluation. Overall rates are pooled across the 33 straight, 31 single-goal obstacle, and 32 multi-goal obstacle trials. Confidence intervals are computed using the Wilson score method.

F Error Categorization Criteria

To evaluate model failure modes, we categorize failed trials in the robotic evaluation into four types based on the following criteria:

Distraction: The robot moves toward a distractor rather than the intended target, regardless of whether it contacts the obstacle blocking that distractor.

Collision: In the absence of distraction, the robot directly hits an obstacle in the scene.

Graze: In the absence of distraction or collision, the robot touches or grazes the edge of an obstacle without a direct impact.

Table 5: **Robotic task success rates with 95% Wilson confidence intervals.** Overall results are pooled across all three contexts.

Model	Context	Successes/Trials	Success Rate (%)	95% CI
BrainVLA	Overall	77/96	80.2	[71.1, 86.9]
	Straight	33/33	100.0	[89.6, 100.0]
	Single-goal obstacle	23/31	74.2	[56.8, 86.3]
	Multi-goal obstacle	21/32	65.6	[48.3, 79.6]
π_0	Overall	60/96	62.5	[52.5, 71.5]
	Straight	33/33	100.0	[89.6, 100.0]
	Single-goal obstacle	16/31	51.6	[34.8, 68.0]
	Multi-goal obstacle	11/32	34.4	[20.4, 51.7]
POYO	Overall	49/96	51.0	[41.2, 60.8]
	Straight	18/33	54.5	[38.0, 70.2]
	Single-goal obstacle	15/31	48.4	[32.0, 65.2]
	Multi-goal obstacle	16/32	50.0	[33.6, 66.4]
NDT	Overall	47/96	49.0	[39.2, 58.8]
	Straight	17/33	51.5	[35.2, 67.5]
	Single-goal obstacle	15/31	48.4	[32.0, 65.2]
	Multi-goal obstacle	15/32	46.9	[30.9, 63.6]
GRU	Overall	44/96	45.8	[36.2, 55.8]
	Straight	15/33	45.5	[29.8, 62.0]
	Single-goal obstacle	15/31	48.4	[32.0, 65.2]
	Multi-goal obstacle	14/32	43.8	[28.2, 60.7]

Deviation: If none of the above events occurs, the trial is categorized as deviation when the robot fails to stop within a 2.5 cm radius of the intended target center.

G Additional Ablation Results

Table 6 restates the ablation study results from the main text in table form for a clearer comparison.

Table 6: **Success rate (%) on unseen obstacle trials for information controls and oracle.**

Model	Single-goal	Multi-goal	Obstacle avg.
Text-based neural tokens	58.1	43.8	50.8
w/o spatial grounding loss	54.8	34.4	44.4
Shuffled tokens	51.6	31.3	41.3
Oracle	90.3	81.3	85.7
BrainVLA	**74.2**	**65.6**	**69.8**