BeliefVLA: Implicit Belief-State Driven Vision-Language-Action Models for Robust Manipulation

Yu Liu[1,†], Tianlv Huang[1,†], Wei Han[1,2,†], Hetian Guo[1], Chengyu Miao[1], Fei Yan[1], Zipei Fan[1,*], and Xuan Song[1]

***Abstract*— While Vision-Language-Action (VLA) models exhibit remarkable open-vocabulary generalization in robotic manipulation, they remain fundamentally constrained by a structurally reactive paradigm. Inferring actions from static visual representations or simple frame stacking forces policies to learn temporal dynamics from scratch, lacking the physical grounding necessary for robust real-world control. We introduce BeliefVLA, a framework that shifts robotic control from purely reactive mapping to reasoning within a continuous implicit belief state. By leveraging a Video Joint-Embedding Predictive Architecture (V-JEPA2) to infer these latent states from historical context, BeliefVLA injects dynamics-aware priors directly into the policy. This approach seamlessly integrates physical reasoning and spatiotemporal dynamics without sacrificing the high-level semantic capabilities of foundational VLMs. Extensive experiments establish BeliefVLA's state-of-the-art performance across simulated and real-world benchmarks. Furthermore, five observation-perturbation tasks explicitly validate its capacity for temporal memory, action-centric features, and physical priors.**

I. Introduction

Vision-Language-Action (VLA) models inherit the aligned visual perception and natural language understanding capabilities of large-scale pretrained Vision-Language Models (VLMs), significantly enhancing open-vocabulary semantic generalization in robotics [1], [2], [3], [4], [5], [6]. Despite these advantages, current VLAs still struggle to generalize effectively to real-world environments.

A major bottleneck in current VLAs originates from their underlying VLMs [7], [8], [9], particularly the prevailing paradigm of their upstream visual representations. Most rely on encoders pretrained via image-text contrastive learning (e.g., CLIP [10], SigLIP [11]) on massive datasets of static web images. While excelling at cross-modal semantic alignment and object recognition, these representations exhibit two fundamental limitations for physical embodiment: (i) *Deficiency in physical priors:* Standard contrastive objectives heavily prioritize high-level semantic abstraction, thereby failing to encode the underlying physical laws governing object interactions and environmental dynamics. (ii) *Absence of temporal context:* The inherent frame-independent design discards temporal dependencies, causing the policy to implicitly treat a POMDP as fully observable. Consequently, this static paradigm forces the system into a memoryless, single-step reactive control regime, rendering it fundamentally incapable of capturing the dynamic evolution of the physical world. This motivates a central question: *How can we inject physical priors and temporal context into visual representations without sacrificing their open-vocabulary generalization to advance downstream VLA policies?*

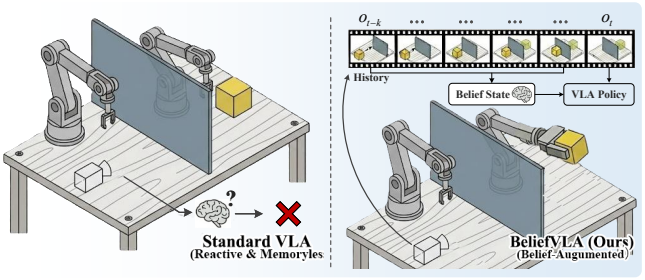

Fig. 1. **Paradigm shift via implicit belief states.** (Left) Standard VLA: Relies solely on current observations. As the object falls out of view, the model fails to execute the correct action due to the absence of memory context. (Right) BeliefVLA (Ours): Leverages historical frame sequences. By inferring a persistent physical belief of the object's location behind the barrier, the model successfully completes the task despite severe occlusion.

To bridge this gap, we argue that rather than relying solely on immediate observations, embodied agents should maintain a belief state inferred from historical context. While the belief state is a fundamental concept for solving Partially Observable Markov Decision Processes (POMDPs) in classical control theory [12], we find that Joint-Embedding Predictive Architectures are inherently suited to fulfill this role in high-dimensional, continuous visual spaces. Unlike methods reliant on pixel-level reconstruction, JEPAs (e.g., V-JEPA2 [13]) optimize the encoder to learn physical dynamics and model temporal evolution by predicting missing patches within an abstract feature space. This latent predictive representation not only filters out task-irrelevant visual noise but, crucially, equips the policy network with an implicit belief state prior that encapsulates temporal consistency.

We introduce BeliefVLA, a Vision-Language-Action (VLA) framework driven by an implicit belief state. While conventional methods rely predominantly on static, frame-independent visual features that fail to capture complex real-world dynamics, BeliefVLA overcomes this by integrating a pretrained V-JEPA2 encoder to infer a continuous implicit belief state from historical observations. Functioning as a physical and temporal prior, this belief state shifts the control policy away from reactive, toward a grounded understanding of environmental dynamics. By injecting this learned prior

†These authors contributed equally to this work.
*Corresponding author (fanzipei@jlu.edu.cn)
[1]School of Artificial Intelligence, Jilin University. [2]IO-AI TECH.
This research was funded by the Science and Technology Development Program of Jilin Province (No. 20250203122SF).

as a dynamic belief-augmentation module, the framework achieves complementary dual-perception capabilities: standard vision-language encoders extract high-dimensional semantic and spatial features, while the V-JEPA2 belief state captures the physical evolution essential for dynamic interaction. The synergy between static semantic comprehension and dynamic physical reasoning ultimately yields highly robust closed-loop robotic control.

- We formulate JEPA-style predictive video representations as an implicit belief state for embodied control, providing a principled way to inject temporal context and physical priors into VLA policies.
- We propose BeliefVLA, a Vision-Language-Action (VLA) framework driven by an implicit belief state. By grounding control policies in history-derived predictive world states, it significantly enhances physical reasoning and dynamics modeling while fully preserving open-vocabulary semantic capabilities.
- We achieve state-of-the-art performance across standard simulated and real-world benchmarks. Additionally, we introduce five observation-perturbation scenarios to explicitly validate BeliefVLA's capabilities in temporal memory, extracting action-centric visual representations, and leveraging physical priors.

II. Related Work

A. Vision-Language-Action Models

Recent Vision-Language-Action (VLA) models transfer the aligned perceptual and linguistic capabilities of large-scale pretrained Vision-Language Models (VLMs) [7], [8], [9] to robotic decision-making, enabling open-vocabulary task understanding and semantic generalization across diverse scenes and instructions [1], [2], [4], [5]. By grounding action prediction in VLM-derived representations, these approaches substantially reduce the need for task-specific perception engineering and improve robustness to novel objects, synonyms, and compositional language [3], [6]. Despite this progress, real-world deployment still exposes a persistent gap: many VLAs remain brittle under distribution shifts involving contact-rich interactions, dynamics, occlusions, and long-horizon closed-loop execution [14], [15], [16]. This brittleness is partly driven by the fact that large-scale robotic datasets, while rapidly growing, still struggle to fully cover the diversity of in-the-wild environments and multi-step task dependencies [17], [18], [19].

Unlike existing VLAs that incorporate historical context via simple frame stacking [3], [20], which forces the policy to infer temporal dynamics from scratch, we explicitly construct a dynamics-aware latent belief state from predictive representations, thereby enabling a highly data-efficient VLA paradigm.

B. Visual Representations for VLA

The dominant paradigm in modern VLMs pretrains visual encoders (e.g., CLIP [10], SigLIP [11]) using image-text contrastive objectives on internet-scale data, producing cross-modal alignment and high-level semantic abstraction [10], [11], [7], [8], [9]. However, due to a lack of physical grounding, these paradigms remain suboptimal for embodied agents [21]. Contrastive objectives tend to extract coarse semantic invariances, often discarding the action-relevant physical priors and dynamic structures required for real-world manipulation [13], [22], [23]. Furthermore, their reliance on single-frame observations neglects spatiotemporal kinematics, reducing continuous robotic control to a memoryless Markov Decision Process [24].

To bridge this critical gap, we propose treating JEPA-based predictive visual representations as implicit belief states. This formulation structurally shifts VLA models from a strictly memoryless, reactive paradigm into a belief-augmented regime that inherently encodes physical laws and temporal dynamics.

C. World State Representations

Classical control theory formalizes partial observability through Partially Observable Markov Decision Processes (POMDPs). Within this framework, optimal decision-making requires a belief state, which aggregates past observations into a sufficient statistic to model the uncertainty of latent world states [12]. Extending these belief states to high-dimensional visual control typically relies on learned world models [25], [21], [26], [27]. However, pixel-level reconstruction in such models is computationally expensive and prone to overfitting task-irrelevant details [28], [29].

Conversely, Joint-Embedding Predictive Architectures (JEPAs) [13] learn temporally coherent representations by predicting abstract latent features. This mechanism avoids explicit reconstruction while effectively capturing core environmental dynamics. From a POMDP perspective, we formulate these predictive latents as implicit belief states that efficiently integrate historical context and encode dynamics-aware priors. Unlike standard VLA pipelines relying on static VLM features, our approach explicitly injects pretrained JEPA-derived belief states into the policy. This integration improves real-world closed-loop robustness without sacrificing open-vocabulary semantic generalization.

III. Method

A. Problem Formulation & Architecture Overview

1) Problem Formulation: Real-world robotic manipulation fundamentally constitutes a Partially Observable Markov Decision Process (POMDP), defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{R}, \gamma)$. Here, $s_t \in \mathcal{S}$ denotes the unobservable true physical state, $o_t \in \mathcal{O}$ is the current visual observation, and $a_t \in \mathcal{A}$ is the action at time step t. Existing VLA policies simplify the physical environment into a fully observable MDP. They rely on a static visual encoder Φ_{static} to extract features from the immediate observation (e.g., a single frame or a simple stack of recent frames) and optimize the policy π_θ given a natural language instruction l as follows:

$$a_t \sim \pi_\theta(a_t \mid \Phi_{\text{static}}(o_t), l) \tag{1}$$

While this paradigm achieves reasonable reactive control, it inherently fails to fully capture the underlying temporal

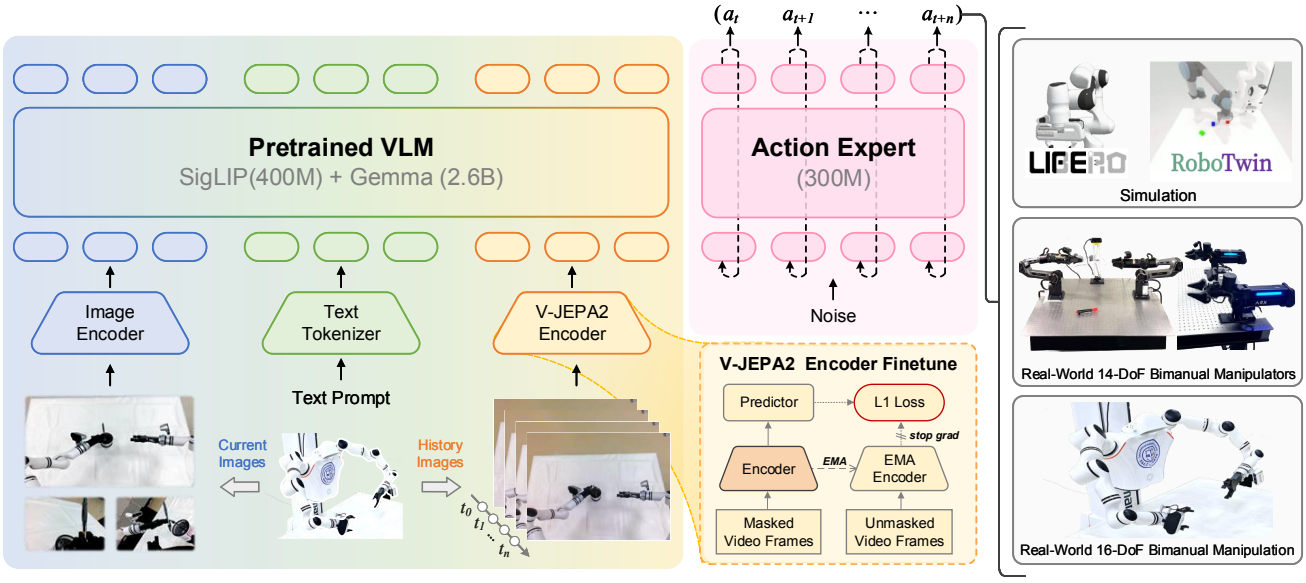

Fig. 2. **Framework of BeliefVLA.** The model integrates a pretrained vision-language front-end to extract instantaneous semantic context and a domain-adapted V-JEPA2 latent world encoder to infer physical dynamics from historical observations. These multimodal representations are fused to condition a continuous action expert, which employs conditional flow matching to iteratively refine random noise into precise multi-step action sequences.

environmental dynamics $\mathcal{T}(s_{t+1}|s_t, a_t)$. Even when extending the input o_t to a stack of historical frames, such naive temporal aggregation merely provides superficial motion cues. It fundamentally struggles to construct a robust understanding of complex physical interactions and long-horizon spatiotemporal constraints, leaving the policy in a myopic regime.

Under the POMDP framework, optimal decision-making requires inferring a belief state b_t that reflects the true physical environment [12]. To achieve this, we employ a Joint-Embedding Predictive Architecture (V-JEPA2 [13]) as a spatiotemporal encoder Ψ_{JEPA} to extract an implicit belief prior $b_t = \Psi_{\text{JEPA}}(o_{t-K:t})$. While initially derived from visual observations, this prior interacts directly with the VLM's latent action space via deep self-attention layers. Through this process, it is progressively refined into an implicit belief state representation, capturing the essential interplay between historical observations and executed actions. Consequently, our closed-loop policy is reformulated to condition on both static semantic $\Phi_{\text{static}}(o_t)$ features and this physics-driven belief state b_t:

$$a_t \sim \pi_\theta(a_t \mid \Phi_{\text{static}}(o_t), b_t, l) \tag{2}$$

2) Architecture Overview: The core objective of BeliefVLA is to map multimodal inputs, comprising both images and language, into continuous robotic actions. As illustrated in Fig. 2, the proposed framework consists of two primary components: multimodal perception and action generation. For multimodal perception, the architecture is built upon a pretrained Vision-Language Model (VLM) backbone. It processes a natural language instruction and the current visual frame to construct a rich semantic understanding of the scene, while concurrently processing a historical sequence of visual observations to formulate a dynamic physical belief state. Subsequently, in the action generation phase, the unified representations from the VLM—fusing the semantic understanding with physical priors and temporal dynamics—condition a Flow Matching-based action expert to synthesize continuous motor commands.

B. Multimodal Perception

1) Dynamic Belief State Embedding: To overcome the memoryless limitation of standard VLMs, we maintain a continuous belief state that encapsulates the physical dynamics and temporal evolution of the environment. We leverage the Joint-Embedding Predictive Architecture, specifically V-JEPA2 [13], which is naturally suited for modeling state transitions in latent space.

Instead of training from scratch, we initialize the V-JEPA2 encoder with its pretrained weights, which already encode strong general physical priors from diverse video datasets. To bridge the domain gap between general videos and specific robotic manipulation tasks, we further fine-tune the V-JEPA2 encoder using domain-specific robotic trajectory data. Training details are provided in Sec. III-D.1.

In our forward pass, given a history of observations, the fine-tuned V-JEPA2 encoder extracts a sequence of abstract feature embeddings. We treat these temporal embeddings as dynamic visual tokens, denoted as $\mathcal{Z}_{\text{belief}}$. Crucially, these tokens function as an implicit physical belief state, capturing the trajectory's momentum, object interactions, and temporal consistency, ready to be ingested by the downstream VLM.

2) Unified Multimodal Condition Encoding: Parallel to the belief state construction, we extract embeddings from the current sensory inputs. The natural language instruction is processed by the text embedding layer of PaliGemma [7] to produce semantic tokens $\mathcal{Z}_{\text{text}}$. Simultaneously, the current

static image observation is encoded by the internal SigLIP visual encoder to generate static visual tokens $\mathcal{Z}_{\text{static}}$.

To form a unified representation, we project $\mathcal{Z}_{\text{belief}}$ into the same latent dimension as the PaliGemma embeddings via a linear projection. We then concatenate these three distinct token sequences along the sequence dimension:

$$\mathcal{Z}_{\text{input}} = [\mathcal{Z}_{\text{text}}; \mathcal{Z}_{\text{static}}; \mathcal{Z}_{\text{belief}}] \tag{3}$$

This concatenated sequence $\mathcal{Z}_{\text{input}}$ is then fed into the deep transformer layers of the PaliGemma-3B backbone. Through the self-attention mechanism, the model learns to entangle the open-vocabulary semantic concepts with the temporal physical priors. The final hidden state of the transformer serves as the comprehensive multimodal condition $\mathcal{Z}_{\text{cond}}$ for downstream action generation.

C. Action Generation

Conditioned on the fused multimodal embedding $\mathcal{Z}_{\text{cond}}$, the action generation module decodes a sequence of motor commands (action chunks) using a conditional flow matching [30] formulation. Let $a \in \mathbb{R}^{D_a}$ denote the continuous action vector. The action expert is parameterized to predict a time-dependent vector field $v_\theta(a_\tau, \tau, \mathcal{Z}_{\text{cond}})$, which defines a probability path between a tractable Gaussian noise distribution $p_0(a_0)$ and the expert data distribution $q(a_1)$, where $\tau \in [0, 1]$ represents the flow matching timestep.

At inference time, we generate actions by integrating the learned vector field starting from random noise $a_0 \sim \mathcal{N}(0, I)$. We employ the forward Euler method to solve the ordinary differential equation (ODE):

$$a_{\tau+\delta} = a_\tau + \delta \cdot v_\theta(a_\tau, \tau, \mathcal{Z}_{\text{cond}}) \tag{4}$$

where δ denotes the integration step size. By grounding this ODE-based refinement in the unified semantic and physical context, BeliefVLA achieves precise and multimodal robotic control.

D. Training Recipe and Implementation Details

1) Physical Representation Alignment: We fine-tune the V-JEPA2 [13] visual encoder on a mixture of robotic datasets, including LIBERO [15], RoboTwin [31], and RealSource World [32]. This adaptation is performed with a masked latent prediction objective: conditioned on the historical visual context, the model predicts the latent representations of masked spatiotemporal regions. This self-supervised phase encourages the encoder to capture domain-specific spatial structure and physical relevant dynamics.

2) End-to-End Policy Fine-Tuning: Following the physical representation alignment phase, the domain-adapted V-JEPA2 encoder is strictly frozen. It serves to extract physically grounded visual tokens, which are seamlessly fused into the VLM prefix. Conditioned on this enriched context, we jointly fine-tune the VLM fusion layers and the continuous action expert end-to-end on task-specific demonstrations. To achieve this, we model continuous action generation as a conditional vector field regression problem. Following the Conditional Flow Matching (CFM) paradigm [30], we define a probability density path between a tractable noise prior $\epsilon \sim \mathcal{N}(0, I)$ and the ground-truth continuous action chunk $a \in \mathbb{R}^{T_a \times D}$, where T_a represents the temporal horizon and D denotes the control dimension. The time-dependent interpolated action a_τ for a continuous time variable $\tau \sim \mathcal{U}(0, 1)$ is constructed as:

$$a_\tau = (1 - \tau)\epsilon + \tau a \tag{5}$$

The objective is to train the action expert v_θ to approximate the constant velocity vector field, defined by the time derivative $\frac{d}{d\tau} a_\tau = a - \epsilon$. Crucially, this vector field regression is strictly conditioned on our physically-enriched multimodal prefix $\mathcal{Z}_{\text{cond}}$, forcing the action generation to proactively align with both instantaneous semantics and underlying physical laws. The overall optimization minimizes the following expected Euclidean distance:

$$\mathcal{L}_{FM} = \mathbb{E}_{\tau, a, \epsilon, \mathcal{Z}_{\text{cond}}} \left[\| v_\theta(a_\tau, \tau, \mathcal{Z}_{\text{cond}}) - (a - \epsilon) \|_2^2 \right] \tag{6}$$

3) Implementation Details: To leverage robust foundation priors, both the vision-language front-end and the continuous action expert inherit their initial weights from the pretrained $\pi_{0.5}$ model [2]. Specifically, the vision-language backbone is initialized with PaliGemma [7], employing a 400M-parameter SigLIP encoder [11], while the action expert is parameterized by a 300M-parameter Gemma-based decoder-only transformer [7]. Furthermore, the latent world encoder is built upon a pretrained V-JEPA2 (ViT-Large) encoder [13] to capture underlying physical dynamics.

IV. Experiments

We evaluate BeliefVLA in both simulated and real-world environments to address three core research questions: **Q1:** Does BeliefVLA achieve state-of-the-art (SOTA) performance in robotic manipulation tasks? **Q2:** Can the system maintain robust performance when subjected to corrupted observations and distribution shifts? **Q3:** Are the performance gains fundamentally driven by the implicit belief state, and does a domain-adapted belief encoder combined with an effective multimodal fusion strategy yield further improvements?

To answer **Q1**, we benchmark BeliefVLA across standardized simulation and real-world tasks (Sec. IV-A). To address **Q2**, we designed real-world experiments featuring diverse observation perturbations across three heterogeneous robot configurations (Sec. IV-B). Finally, we answer **Q3** through comprehensive ablation studies that isolate the individual contributions of the implicit belief state and the fusion mechanism (Sec. IV-C).

A. Evaluation on Simulation Benchmarks

We evaluate BeliefVLA against state-of-the-art Vision-Language-Action (VLA) models across two comprehensive simulation benchmarks: LIBERO [15] for single-arm manipulation and RoboTwin [16] for bimanual coordination.

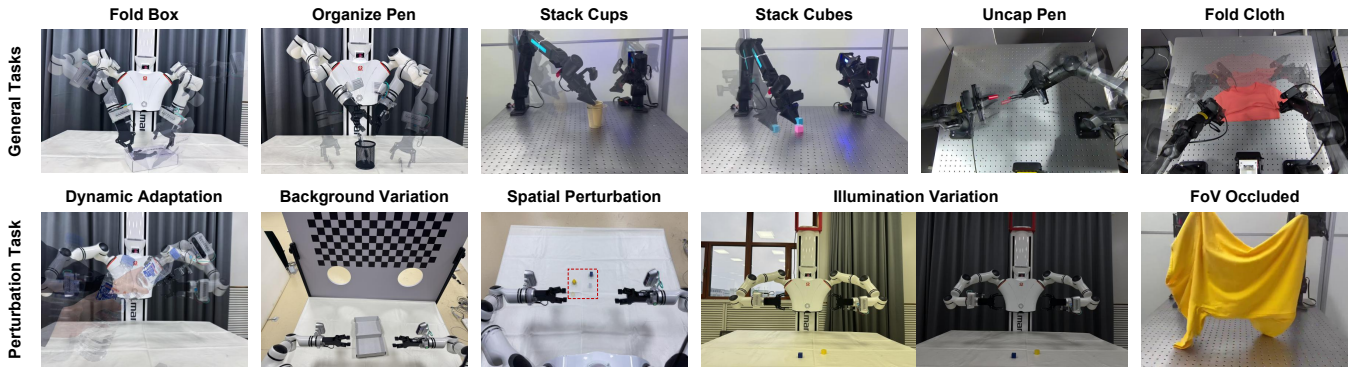

Fig. 3. **Overview of real-world experimental tasks.** Top: General Tasks. We evaluate six general-purpose bimanual manipulation tasks under standard laboratory conditions on three heterogeneous physical platforms (Realman, ARX R5, and AgileX PIPER). Bottom: Robustness. We design five challenging scenarios to assess policy robustness to environmental perturbations, including dynamic adaptation (bottle handover), background variation (switching from the training curtain to a calibration-board background), spatial perturbation (mobile base/platform motion), illumination changes, and field-of-view occlusion (occluding the base camera).

TABLE I
SUCCESS RATES ON THE LIBERO SIMULATION BENCHMARK.

Model	Spatial	Object	Goal	Long	Average
DP [33]	77.8%	91.2%	67.5%	49.8%	71.6%
Octo [34]	78.9%	85.7%	84.6%	51.1%	75.1%
OpenVLA [35]	84.7%	88.4%	79.2%	53.7%	76.5%
SpatialVLA [36]	88.2%	89.9%	78.6%	55.5%	78.1%
CoT-VLA [20]	87.5%	91.6%	87.6%	69.0%	83.9%
π_0 [1]	96.8%	**98.8%**	95.8%	85.2%	94.2%
$\pi_{0.5}$ [2]	**98.8%**	98.2%	**98.0%**	92.4%	96.9%
OpenVLA-OFT [4]	97.6%	98.4%	97.9%	**94.5%**	97.1%
BeliefVLA (Ours)	**98.8%**	98.6%	97.3%	94.2%	**97.2%**

Notes: We report the performance across four distinct suites and their overall average. Best results are **bolded**.

1) LIBERO Benchmark: The LIBERO benchmark primarily evaluates models across four distinct capabilities: spatial reasoning, object-centric understanding, instruction following, and long-horizon planning. We benchmark BeliefVLA against a comprehensive set of baselines, including Diffusion Policy (DP) [33], Octo [34], OpenVLA [35], SpatialVLA [36], OpenVLA-OFT [4], CoT-VLA [20], π_0 [1], and $\pi_{0.5}$ [2]. To ensure statistical reliability, all reported results are averaged across three random seeds. As shown in Table I, BeliefVLA achieves the best average performance, establishing a new state-of-the-art (SOTA) performance. It outperforms the strongest baseline with an average success rate of 97.2%.

2) RoboTwin Benchmark: To evaluate bimanual coordination, we adopt the Aloha-AgileX dual-arm setup [37] from the RoboTwin benchmark [16] across eight representative tasks. Following the official configuration, we fine-tune all policies using 50 demonstration trajectories per task. Performance is evaluated under two distinct protocols: the *Easy* setting, which matches the training environment, and the *Hard* setting, which introduces domain randomization. To ensure statistical significance, we execute 100 evaluation rollouts per task for both settings. We compare BeliefVLA against established baselines, including Diffusion Policy (DP) [33], ACT [37], RDT [5], π_0 [1], and $\pi_{0.5}$ [2]. As summarized in Table II, BeliefVLA consistently outperforms all prior methods. Specifically, compared to the most competitive baseline, $\pi_{0.5}$, our model achieves absolute success rate improvements of 7.8% in the *Easy* setting and 6.9% in the *Hard* setting.

3) Analysis: The state-of-the-art performance established by BeliefVLA across both benchmarks can be primarily attributed to the fact that implicitly modeled belief states, when effectively grounded in physical and spatiotemporal memory constraints, significantly enhance downstream action execution. Specifically, the spatiotemporal memory mitigates compounding errors in long-horizon tasks, while physical grounding ensures robust bimanual coordination under domain randomization (reflected in the 6.9% improvement in RoboTwin *Hard*).

B. Real-World Experiments

1) Experiment Setup: As shown in Fig. 3, we benchmark BeliefVLA in the real world across three heterogeneous robotic platforms on eleven manipulation tasks, covering both general manipulation and vision-perturbation settings:
Bimanual ARX. Comprising two 6-DoF arms and a tri-camera system (two wrist cameras and one third-person base camera), this setup defines a 14-dimensional continuous state and action space. We evaluate general manipulation capabilities via *Stack Cups* and *Stack Cubes*. Furthermore, we evaluate the FoV Occluded perturbation setting by introducing a *Complex Fold Cloth* task, enabling rigorous testing of the policy's robustness when a large garment intermittently obstructs the cameras' view of the manipulators.
Bimanual Realman. This setup comprises two 7-DoF arms and a tri-camera vision system (two wrist-mounted and one egocentric base camera), parameterizing a 16-dimensional continuous state and action space. General manipulation capabilities are first assessed using the *Fold Box* and *Organize Pen* tasks. We subsequently evaluate the policy's robustness against visual perturbations across four conditions: (1) Dynamic Adaptation: The system must adapt to continuous, dynamic shifts in the target position as a human operator

TABLE II

QUANTITATIVE EVALUATION ON ROBOTWIN BENCHMARKS.

Method	Hanging Mug		Move Pillbottle Pad		Pick Dual Bottles		Place Object Stand	
	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard
DP [33]	8.0%	0.0%	1.0%	0.0%	24.0%	0.0%	22.0%	0.0%
ACT [37]	7.0%	0.0%	0.0%	0.0%	31.0%	0.0%	1.0%	0.0%
RDT [5]	11.0%	5.0%	3.0%	0.0%	19.0%	2.0%	6.0%	0.0%
π_0 [1]	5.0%	1.0%	10.0%	0.0%	26.0%	1.0%	17.0%	3.0%
$\pi_{0.5}$ [2]	8.0%	3.0%	26.0%	1.0%	28.0%	8.0%	27.0%	7.0%
BeliefVLA (Ours)	**12.0%**	**6.0%**	**39.0%**	**5.0%**	**38.0%**	**12.0%**	**35.0%**	**20.0%**

Blocks Ranking Size		Shake Bottle		Stamp Seal		Place Burger Fries		Average	
Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard
1.0%	0.0%	65.0%	8.0%	2.0%	0.0%	72.0%	0.0%	24.4%	1.0%
0.0%	0.0%	74.0%	10.0%	2.0%	0.0%	49.0%	0.0%	20.5%	1.3%
0.0%	0.0%	53.0%	21.0%	1.0%	0.0%	21.0%	7.0%	14.3%	4.4%
0.0%	0.0%	71.0%	60.0%	3.0%	0.0%	30.0%	0.0%	20.3%	8.1%
9.0%	1.0%	96.0%	73.0%	8.0%	4.0%	63.0%	16.0%	33.1%	14.1%
13.0%	**4.0%**	**99.0%**	**81.0%**	**16.0%**	**7.0%**	**75.0%**	**33.0%**	**40.9%**	**21.0%**

Notes: We report the performance across eight tasks and their overall average. Models are all directly fine-tuned on 50 clean trajectories without in-domain large-scale pre-training. Best results are **bolded**.

hands a water bottle to the robot. (2) Background Variation: The operational background transitions from a clean surface to highly textured environments (e.g., a calibration board). (3) Spatial Perturbation: The initial relative position between the robot base and the target object is shifted. (4) Illumination Variation: Experiments are conducted across three distinct time periods (morning, noon, and evening) to evaluate resilience to natural lighting fluctuations.

Bimanual AgileX. This platform consists of two 6-DoF arms and a quad-camera system (comprising two wrist-mounted cameras and two egocentric cameras positioned at different elevations), establishing a 14-dimensional continuous state and action space. General manipulation capabilities are assessed via the *Fold Cloth* and *Uncap Pen* tasks.

To ensure reproducibility and statistical robustness, we collect expert demonstrations for each task under the standard observation setting via human teleoperation. All models are fine-tuned on the same amount of data using an identical fine-tuning protocol (i.e., matched training procedures and hyperparameters) and are evaluated with 30 independent rollouts per setting. For comparison, we adopt the representative state-of-the-art VLA model $\pi_{0.5}$ [2], which reports strong performance in simulation, as our primary baseline.

2) General-Purpose Manipulation: As illustrated in Fig. 4(a), BeliefVLA achieves the best performance across all six real-world tasks, attaining an average success rate of 78.3% and outperforming the strongest baseline, $\pi_{0.5}$, by an average of 10.0%. This improvement stems primarily from the fundamental differences in state representation mechanisms. Relying solely on single-frame visual observations, $\pi_{0.5}$ [2] lacks temporal memory, making it highly susceptible to state aliasing in tasks involving reciprocating motions (e.g., uncapping a pen) and long-horizon tasks such as folding clothes or boxes. In contrast, BeliefVLA takes a sequence of historical frames as input and constructs an implicit belief state via the V-JEPA2. The empirical results demonstrate that this implicit state effectively integrates physical priors with temporal memory, thereby substantially enhancing the model's generalization capabilities in real-world environments.

3) Perturbed-Observation Manipulation: As illustrated in Fig. 4(b), BeliefVLA achieves the best performance across all five real-world observation perturbation tasks, attaining an average success rate of 67.3% and outperforming the strongest baseline, $\pi_{0.5}$, by an average margin of 15.4%. These five tasks are designed to systematically evaluate the model's robustness across three distinct perturbation categories, each probing a core capability. First, dynamic and field-of-view (FoV) occlusions are used to validate temporal memory. For instance, in human-robot handovers, integrating historical frames allows the model to accurately infer future state evolutions rather than relying on a single, incomplete observation; similarly, during the complex cloth-folding task, the model effectively leverages historical context to recover the true state when the robotic arm is temporarily occluded by garments. Second, lighting and background variations assess the model's action-relevant representations. Because V-JEPA2 prioritizes latent-space feature encoding over pixel-level reconstruction, it successfully disentangles action-relevant features from environmental noise, exhibiting superior adaptability to out-of-distribution (OOD) visual perturbations. Finally, spatial perturbations—introduced by altering the initial relative positions between the robot and target objects—verify the physical priors embedded within the model. This confirms BeliefVLA's robust spatial reasoning and generalization capabilities in navigating complex 3D relationships.

4) Inference Efficiency: To assess the feasibility of real-world deployment, we evaluate the pure inference speeds of both models. As shown in Fig. 4(c), the baseline $\pi_{0.5}$ operates at approximately 25 Hz (40–45 ms per forward pass), whereas BeliefVLA maintains around 20 Hz (about

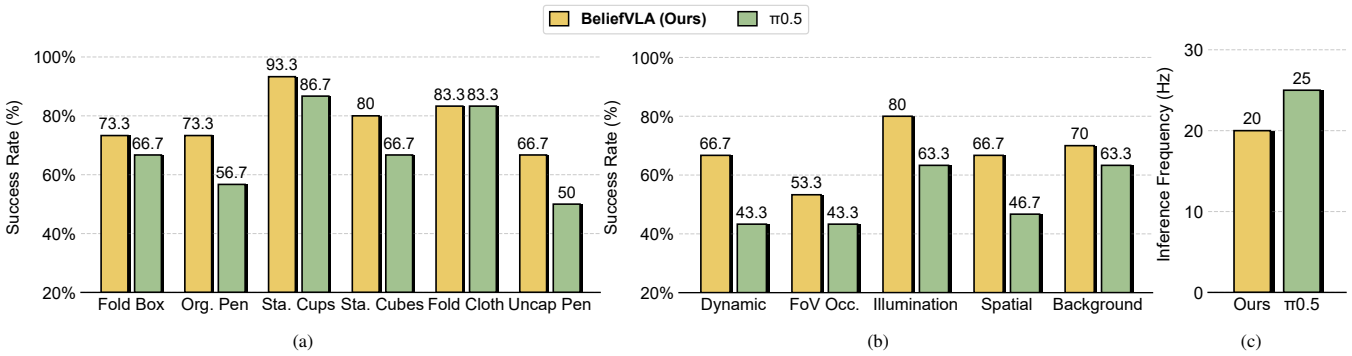

Fig. 4. **Real-world experimental results.** Success rates for (a) general-purpose and (b) perturbed-observation manipulation tasks. (c) Inference frequency (Hz) comparison between BeliefVLA (ours) and $\pi_{0.5}$.

50 ms). This demonstrates a highly practical trade-off: a marginal and acceptable increase in inference latency yields substantial improvements in structural robustness and task success rates under highly challenging scenarios.

C. Ablation Studies

To isolate the impact of our key design choices, we ablate (i) the belief-state representation and (ii) the multimodal fusion mechanism. Simulation results are reported as average success rates over eight representative bimanual tasks in RoboTwin and the suite-level average in LIBERO; real-world ablations are conducted on *Cube Stacking*.

1) Ablation for Belief State Representation: We investigate the efficacy of the V-JEPA2 encoder [13]. Our full model employs a V-JEPA2 encoder explicitly pretrained on robotics data to extract domain-specific physical priors. We compare this against two variants: a **w/o domain adaptation** variant (utilizing off-the-shelf V-JEPA2 weights without robotics fine-tuning) and a **w/o V-JEPA2** variant (which completely removes the belief representation).

As detailed in Table III, BeliefVLA yields an average improvement of 3.9% over the w/o domain adaptation variant across both real and simulated environments. This performance gap underscores the critical role of domain adaptation; while the original V-JEPA2 encoder [13] leverages internet-scale data for broad generalization, its generic feature representations are insufficient for the physical dynamics required in robotics. Furthermore, compared to the w/o V-JEPA2 variant, our full model achieves a 7.1% improvement across all evaluated settings. This unequivocally validates the effectiveness of incorporating the proposed belief state representation, which is highly consistent with the findings in our main comparative evaluations.

2) Ablation for Fusion Mechanism: We analyze the impact of our fusion strategy by evaluating how the Vision-Language Model integrates multimodal inputs. Our default BeliefVLA employs bidirectional visibility within the PaliGemma self-attention over the concatenated sequence $[\mathcal{Z}_{\text{text}}; \mathcal{Z}_{\text{static}}; \mathcal{Z}_{\text{belief}}]$, representing the language instructions, current-frame SigLIP [11] embeddings, and V-JEPA2 features, respectively. This allows for full, unrestricted token-to-token interactions. To isolate this effect, we compare it against a **causal (belief-isolated) masking strategy**. By applying a block attention mask in the PaliGemma self-attention, queries from the V-JEPA2 tokens ($\mathcal{Z}_{\text{belief}}$) are prevented from attending to the instantaneous semantic observations ($\mathcal{Z}_{\text{text}}$ and $\mathcal{Z}_{\text{static}}$).

TABLE III
ABLATION STUDY RESULTS.

Model Variant	Simulation		Real-World
	RoboTwin	LIBERO	Cube Stack
Belief State Representation			
(a) w/o V-JEPA2 ($\pi_{0.5}$)	33.1%	96.9%	66.7%
(b) w/o domain adaptation	36.0%	97.1%	73.3%
Fusion Mechanism			
(c) Causal	17.0%	94.5%	46.7%
BeliefVLA (Ours)	**40.9%**	**97.2%**	**80.0%**

As detailed in Table III, the default BeliefVLA outperforms this causal baseline by significant margins of 23.9% in RoboTwin, 2.7% in LIBERO, and 33.3% in real-world evaluations. This substantial improvement is primarily attributed to the rich contextual interactions enabled by bidirectional visibility. Although allowing the V-JEPA2 tokens to attend to the current-step observations introduces larger optimization variance and requires an extended training schedule to reach convergence, it ultimately yields a far superior attention distribution.

V. CONCLUSION

We present BeliefVLA, a novel framework that addresses the lack of physical priors and temporal context in standard Vision-Language-Action (VLA) models. By integrating a pretrained V-JEPA2 encoder to extract spatiotemporal belief priors, which are then fused with language and visual representations within the VLM, our approach effectively bridges open-vocabulary semantic comprehension with an implicit understanding of physical dynamics. Extensive evaluations demonstrate state-of-the-art performance across simulation and real-world benchmarks, highlighting the framework's enhanced robustness and reliable closed-loop control against dynamic perturbations.

Despite these advancements, two primary limitations remain. Currently, belief state inference relies on a fixed-length observation window, limiting its effectiveness in long-horizon tasks where temporal memory decay occurs. Furthermore, the reliance on V-JEPA2 representations introduces computational overhead that increases both training and deployment costs. Future research will investigate the integration of dynamic memory modules or State Space Models to sustain and update physical belief states over extended temporal horizons, scaling the framework for complex, long-term planning.

REFERENCES

[1] 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.

[2] Physical Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai *et al.*, "$\pi_{0.5}$: a vision-language-action model with open-world generalization," 2025. [Online]. Available: https://arxiv.org/abs/2504.16054

[3] 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.

[4] M. J. Kim, C. Finn, and P. Liang, "Fine-tuning vision-language-action models: Optimizing speed and success," *arXiv preprint arXiv:2502.19645*, 2025.

[5] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu, "Rdt-1b: a diffusion foundation model for bimanual manipulation," *arXiv preprint arXiv:2410.07864*, 2024.

[6] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu *et al.*, "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[7] A. Steiner, A. S. Pinto, M. Tschannen, D. Keysers, X. Wang, Y. Bitton, A. Gritsenko, M. Minderer, A. Sherbondy, S. Long *et al.*, "Paligemma 2: A family of versatile vlms for transfer," *arXiv preprint arXiv:2412.03555*, 2024.

[8] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[9] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[11] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 11 975–11 986.

[12] M. T. Spaan, "Partially observable markov decision processes," in *Reinforcement learning: State-of-the-art*. Springer, 2012, pp. 387–414.

[13] M. Assran, A. Bardes, D. Fan, Q. Garrido, R. Howes, M. Muckley, A. Rizvi, C. Roberts, K. Sinha, A. Zholus *et al.*, "V-jepa 2: Self-supervised video models enable understanding, prediction and planning," *arXiv preprint arXiv:2506.09985*, 2025.

[14] S. Fei, S. Wang, J. Shi, Z. Dai, J. Cai, P. Qian, L. Ji, X. He, S. Zhang, Z. Fei *et al.*, "Libero-plus: In-depth robustness analysis of vision-language-action models," *arXiv preprint arXiv:2510.13626*, 2025.

[15] 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.

[16] T. Chen, Z. Chen, B. Chen, Z. Cai, Y. Liu, Z. Li, Q. Liang, X. Lin, Y. Ge, Z. Gu *et al.*, "Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation," *arXiv preprint arXiv:2506.18088*, 2025.

[17] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," *arXiv preprint arXiv:2403.12945*, 2024.

[18] F. Ebert, Y. Yang, K. Schmeckpeper, B. Bucher, G. Georgakis, K. Daniilidis, C. Finn, and S. Levine, "Bridge data: Boosting generalization of robotic skills with cross-domain datasets," *arXiv preprint arXiv:2109.13396*, 2021.

[19] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 6892–6903.

[20] Q. Zhao, Y. Lu, M. J. Kim, Z. Fu, Z. Zhang, Y. Wu, Z. Li, Q. Ma, S. Han, C. Finn *et al.*, "Cot-vla: Visual chain-of-thought reasoning for vision-language-action models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1702–1713.

[21] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, "Mastering diverse domains through world models," *arXiv preprint arXiv:2301.04104*, 2023.

[22] J. Zhang and K. Ma, "Rethinking the augmentation module in contrastive learning: Learning hierarchical augmentation invariance with expanded views," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 650–16 659.

[23] J. Zeng, Q. Bu, B. Wang, W. Xia, L. Chen, H. Dong, H. Song, D. Wang, D. Hu, P. Luo *et al.*, "Learning manipulation by predicting interaction," *arXiv preprint arXiv:2406.00439*, 2024.

[24] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[25] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson, "Learning latent dynamics for planning from pixels," in *International conference on machine learning*. PMLR, 2019, pp. 2555–2565.

[26] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," *arXiv preprint arXiv:1912.01603*, 2019.

[27] D. Hafner, T. Lillicrap, M. Norouzi, and J. Ba, "Mastering atari with discrete world models," *arXiv preprint arXiv:2010.02193*, 2020.

[28] F. Deng, I. Jang, and S. Ahn, "Dreamerpro: Reconstruction-free model-based reinforcement learning with prototypical representations," in *International conference on machine learning*. PMLR, 2022, pp. 4956–4975.

[29] N. Hansen, H. Su, and X. Wang, "Td-mpc2: Scalable, robust world models for continuous control," *arXiv preprint arXiv:2310.16828*, 2023.

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

[31] Y. Mu, T. Chen, Z. Chen, S. Peng, Z. Lan, Z. Gao, Z. Liang, Q. Yu, Y. Zou, M. Xu *et al.*, "Robotwin: Dual-arm robot benchmark with generative digital twins," in *Proceedings of the computer vision and pattern recognition conference*, 2025, pp. 27 649–27 660.

[32] RealSource, "Realsource world: A large-scale real-world dual-arm manipulation dataset," https://huggingface.co/datasets/RealSourceData/RealSource-World, 2025.

[33] 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*, vol. 44, no. 10-11, pp. 1684–1704, 2025.

[34] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu *et al.*, "Octo: An open-source generalist robot policy," *arXiv preprint arXiv:2405.12213*, 2024.

[35] 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.

[36] D. Qu, H. Song, Q. Chen, Y. Yao, X. Ye, Y. Ding, Z. Wang, J. Gu, B. Zhao, D. Wang *et al.*, "Spatialvla: Exploring spatial representations for visual-language-action model," *arXiv preprint arXiv:2501.15830*, 2025.

[37] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.