NOSTRA: Enabling Robust Robot Imitation via Multimodal Latent Imagination

Vaibhav Saxena[1], Yunhao Luo[1], Yotto Koga[2], and Danfei Xu[1]

***Abstract*—Similar to humans, robots benefit from multiple sensing modalities when performing complex manipulation tasks. Current behavior cloning (BC) policies typically fuse learned observation embeddings from multimodal inputs before decoding them into actions. This approach suffers from two key limitations: 1) it requires all modalities to be present and in-distribution at test time, otherwise corrupting the latent state and leading to fragile execution; and 2) naive fusion across all inputs hinders learning from large-scale heterogeneous datasets, where only a subset of modalities may be informative at different phases of a task. We introduce NOSTRA, a multimodal state-space model that learns a modular per-modality latent representation, enabling flexible action prediction with or without specific inputs. BC-NOSTRA improves robustness to unseen noise by using KL divergence between inferred and imagined multimodal latents as a noise measure, and by employing latent imagination to predict action trajectories over arbitrary horizons. On a suite of MuJoCo-based tasks, BC-NOSTRA fits expert demonstrations up to six input modalities (multi-view RGB, depth, and proprioception), achieving over 20% higher performance under noisy evaluation. Furthermore, NOSTRA adaptively down-weights non-informative inputs, facilitating effective co-training on large heterogeneous robotics datasets with $\mathcal{O}(10k)$ demonstrations spanning diverse tasks and visual conditions. Finally, we demonstrate real-world deployment, where BC-NOSTRA achieves up to a 40% performance gain under camera occlusions on multiple manipulation tasks.**

***Index Terms*—Imitation Learning, Probabilistic Inference**

I. INTRODUCTION

HUMANS rely on a rich variety of sensing modalities – vision, hearing, touch, pressure, temperature – to perform everyday tasks. Crucially, not all modalities are necessary at once: when one becomes unavailable (e.g., vision under occlusion), we draw on experience and dead-reckoning to bridge the gap, and seamlessly return to the more informative signal once it becomes available again. Achieving human-level robustness in robotic object manipulation requires similar capabilities. Robots must be able to process diverse inputs such as multiple camera views, depth, and proprioception, each of which occupies a distinct subspace and often warrants its own encoder within a visuomotor policy. Equally important, they must learn to identify which inputs are useful at a given time and remain robust when some become noisy, unreliable, or entirely missing due to sensor failures. While state-of-the-art behavior cloning (BC) policies do incorporate multiple input modalities, they typically fuse their learned representations naively to predict actions [1, 2], a strategy that falls short of these requirements.

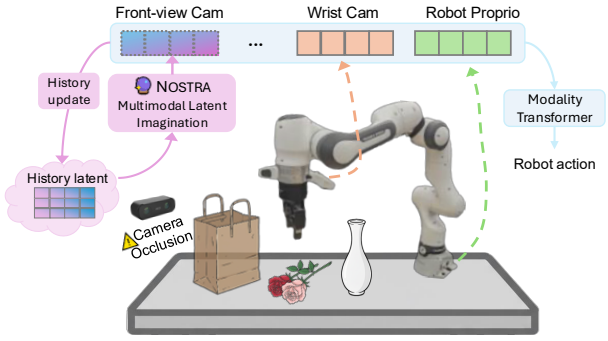


Figure 1: NOSTRA uses latent imagination to handle unforeseen noise in a multimodal robotic system. This allows it to maintain performance and robustness even with noisy or corrupt inputs.

Incorporating multimodal observations into visuomotor policies presents significant challenges, particularly in deciding when and how each modality should contribute to decision-making. Not all task phases require every modality, and incorporating irrelevant or unreliable signals can inject noise, corrupt shared representations, and degrade policy performance. Prior work has attempted to address this through task-specific heuristics or modality-specific gating – [3] proposed a learned contact predictor to selectively incorporate force/torque feedback only during contact-rich phases, while [4] use information-theoretic criteria to identify useful modalities – the approach being limited to just two modalities. Despite these advances, existing strategies remain limited in generality and robustness, particularly in real-world settings where multiple modalities may intermittently fail or degrade. This highlights the need for a more scalable framework that can flexibly manage modality reliability while preserving strong visuomotor control.

To address these challenges, we propose a modular latent representation strategy that explicitly partitions the latent space into subspaces dedicated to individual input modalities. This design disentangles modality-specific information, ensuring that the degradation of one modality does not contaminate the entire latent state, while enabling the policy to adaptively weight each input based on its relevance to the task. To remain robust under missing or noisy inputs, we introduce a marginal latent state conditioned on a memory module, which leverages historical context to infer plausible substitutes. Latent variable models have demonstrated strong capabilities in capturing long-term dependencies for video prediction [5], 3D navigation [6], and complex Atari gameplay [7], making them a natural foundation

Manuscript received: March, 13, 2026; Revised July, 4, 2026; Accepted July, 22, 2026.

This paper was recommended for publication by Wei Pan upon evaluation of the Associate Editor and Reviewers' comments.

[1]School of Interactive Computing, Georgia Institute of Technology, Atlanta, USA; correspondence to vsaxena33@gatech.edu

[2]Autodesk Research

Digital Object Identifier (DOI): see top of this page.

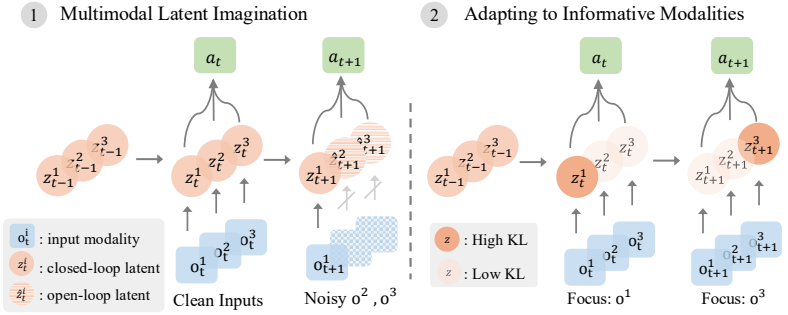


Figure 2: **Model Capabilities.** NOSTRA enables: (1) per-modality robustness to noise via Multimodal Latent Imagination – a novel capability due to our modular stochastic latent space that learns open- and closed-loop latents for each modality, and (2) an emergent adaption to informative modalities in its latent space that adapt to the heterogeneity in multimodal inputs, allowing for robust pre- or co-training on diverse robotics datasets.

for modeling history in visuomotor control.

We present NOSTRA, a multimodal state-space model that enables a visuomotor policy to (1) ignore individual input modalities when they become noisy, (2) use per-modality latent imagination for open-loop execution in extended periods of noisy inputs, and (3) selectively attend to informative modalities when learning from datasets with non-informative inputs (i.e. heterogeneous). Our visuomotor policy, BC-NOSTRA, trains to fit sequences of expert demonstrations for robotic manipulation tasks. NOSTRA learns modular stochastic representations regularized by an information bottleneck. To maintain task-specific behavior both with and without inputs, we leverage a learned marginal (prior) latent distribution, or *open-loop* latent, for open-loop rollouts, and the variational approximation to the input-conditioned posterior (per modality) for *closed-loop* rollouts. In our experiments, we show *multimodal latent imagination* (MLI) allows BC-NOSTRA to be robust against out-of-distribution noises in RGB images (such as occlusion, changing textures), depth, and robot proprioception. NOSTRA also enables efficient pre- and co-training with large-scale heterogeneous datasets that may contain non-informative inputs. We deploy BC-NOSTRA on a real robot showing sample-efficient policy learning from 30 human-teleoperated demos, as well as robustness to camera occlusion. To summarize:

1) We present NOSTRA, a novel observation trunk for BC that learns *per-modality* stochastic latents for both *closed-loop* (input-conditioned) and *open-loop* (marginal) states, allowing for action generation with missing sensor inputs using *multimodal latent imagination* (MLI), and efficient learning from heterogeneous datasets that contain non-informative inputs.
2) We create an adaptive mechanism to tackle noisy input modalities, called AdaMLI, that uses the per-modality KL divergence between inferred and imagined latents as a metric to identify noise, and adaptively switches between latent imagination and closed-loop prediction.
3) We train BC-NOSTRA on 12 MuJoCo-based tasks (up to 3 noisy evaluation variants, and 6 modalities including RGB, depth, and robot proprio), showing our method identifies and ignores noisy RGB, depth, and low-dim inputs, achieving higher success by up to 60%, >20% on avg. (Table I & II), in a suite of high-precision, prehensile, and long-horizon tasks.
4) We analyze how BC-NOSTRA trains on tasks where not all inputs are informative (heterogeneous). When pre-training on a dataset with non-informative RGB inputs, BC-NOSTRA is able to better utilize a finite network capacity by re-attributing extracted nats to informative modalities, resulting in better fine-tuning with higher success by up to 18% (Table VI), as well as better co-training with diverse datasets ($\mathcal{O}(10k)$ demos), by up to 20% (Table VII).

II. RELATED WORKS

Behavior cloning from diverse datasets. Learning from offline expert demonstrations has become an increasingly popular strategy for training robots for manipulation [2, 8–11] – end-to-end policies take multimodal observations as input and output robotic actions [12]. However, lack of standardized embodiments and datasets poses significant challenges. Several works have leveraged a variety of visual inputs to aid learning, such as video [13–15], voxel-grids [16, 17], and point-clouds [18, 19]. However, there is a missing unified approach to learning from heterogeneous datasets, or even an information-theoretic understanding of the challenge itself. As such, current methods constrain themselves to specific input modalities [20] and simply fuse different modalities together (feature-level concatenation), which limit their performance to certain tasks or make them sensitive to perception noise. In this work, we propose a general multi-modality visuomotor policy framework that learns a modular per-modality stochastic latent space, explicitly regularized by history-conditioned marginal latents, paving the way for robust multimodal policy learning.

Latent-variable models in policy learning. Recent advances in latent representation learning have enabled compact modeling of sequential decision-making through latent state-space models, facilitating long-horizon prediction [5], planning [21], and policy learning from high-dimensional inputs [7]. These ideas have recently been adopted in robot imitation learning and world models to improve visuomotor prediction and temporal reasoning [1, 8, 13]. In parallel, multimodal latent variable models have demonstrated that structured latent representations can disentangle shared and modality-specific information across heterogeneous observations, improving representation learning and cross-modal reasoning [22–24]. Unlike these works, our objective is not to improve latent prediction or planning performance, but to enable robust policy deployment under sensor corruption. We leverage modular latent dynamics to selectively imagine missing observations at inference time and adaptively replace unreliable sensory inputs, allowing policies to remain robust without corruption-aware training.

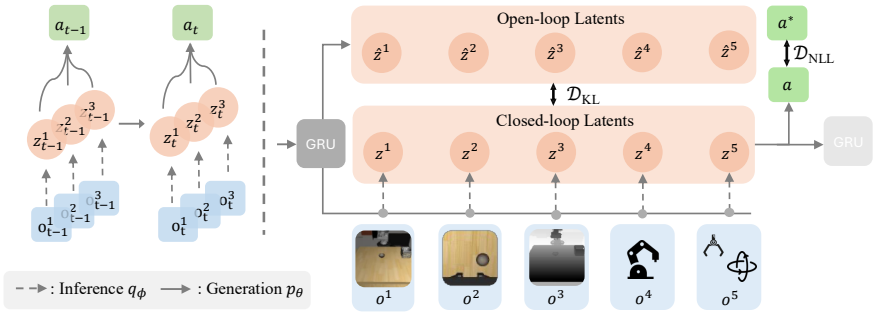


Figure 3: **Training BC-NOSTRA.** (Left) Graphical model showing the generation (p_θ, solid arrows) and inference (q_ϕ, broken arrows) procedures in our generative model. (Right) Training details for the modular latent space in NOSTRA. We use a GRU to for latent dynamics computation (open-loop latents), with separate networks for processing inputs (closed-loop latents) that share the GRU.

III. METHOD

We present NOSTRA, a multimodal state-space model that learns a modular latent space given inputs from multiple observation modalities. A modular latent space allows our downstream visuomotor policy, BC-NOSTRA, to attend to useful modalities during rollouts and ignore non-informative ones when training on heterogeneous datasets. BC-NOSTRA can also ignore individual noisy input modalities and use latent imagination to execute with only partial inputs. In this section, we describe our methodology for training BC-NOSTRA to fit a dataset of N expert demonstrations, $\mathfrak{D} = \{{}^{(i)}o_{1:T}^{1:M}, {}^{(i)}a_{1:T}^{*}\}_{i=1}^{N}$, containing M different input modalities that include multi-view RGB images, depth maps, and low-dimensional robot proprioception. We consider each input that is either captured by a separate physical sensor, occupies a different subspace, or requires a separate encoder, as its own modality, since it acts as an independent source of state information.

A. NOSTRA: A Multimodal State-Space Model

Joint distribution. BC-NOSTRA learns to maximize the likelihood of data under a joint model $p(a_{1:T}, o_{1:T}^{1:M})$, with learned latents $z_{0:T}^{1:M}$, that is factorized into two components: (1) a latent dynamics model that transitions a collection of latents $z_t^{1:M}$ given the latent history, and (2) an action and observation (not trained) generation model conditioned on the current multimodal latent. Concretely,

$$p_\theta(a_{1:T}, o_{1:T}^{1:M}) = \int p_\theta(a_{1:T}, o_{1:T}, z_{0:T}^{1:M}) dz_{0:T}^{1:M}$$
$$\triangleq \int p_\theta(z_0^{1:M}) \prod_{t=1}^{T} \underbrace{p_\theta(a_t|z_t^{1:M})}_{\text{action decoder}} p_\theta(o_t|z_t^{1:M}) \underbrace{p_\theta(z_t^{1:M}|z_{t-1}^{1:M})}_{\text{latent dynamics}} dz_{0:T}^{1:M}. \tag{1}$$

We model all distributions as diagonal multivariate Gaussians with learned means and variances for all latents, and learned means and fixed variance for decoders. Since the focus of this work is to learn visuomotor policies, we set the variance of the observation decoder $p_\theta(o_t|z_t^{1:M})$ to inf, resulting in a model that trains to generate actions only. Note that unlike BC-LSTM [1] or Diffusion Policy [2], which only learn $p(z_t|o_t)$, we explicitly learn a latent prior $p(z_t^{1:M}|z_{t-1}^{1:M})$ so the policy can operate when some or all observations are missing. Moreover, this joint formulation provides a principled variational inference view to per-modality encodings which define an approximate posterior $q_\phi(z_t^{1:M}|z_{t-1}^{1:M}, o_t)$ as we shall see below.

Inference. Similar to [25, 26], we introduce a learned posterior over latents amortized by current observations, $q(z_t^{1:M}|z_{t-1}^{1:M}, o_t^{1:M})$, to help integrate over the space of introduced latents. Since our latent space is modular, we can even sample a subset of latents in this *closed-loop* fashion, say, $z_t^{1:m} \sim q_\phi(z_t^{1:m}|z_{t-1}^{1:M}, o_t^{1:m})$ where $m \leq M$. During training, we default to inferring latents conditioned on all observations (i.e. $m = M$) to subsequently condition the action decoder and latent dynamics.

Generation. To sample actions given a multimodal latent $z_{t-1}^{1:M}$, $p_\theta(z_t^{1:M}|z_{t-1}^{1:M})$ transitions the multimodal latent dynamics which is decoded into actions using $p(a_t|z_t^{1:M})$. This formulation allows us to sample actions conditioned solely on the latent state in an *open-loop* fashion without having to observe individual inputs. Since our latent space is modular, we can sample open-loop latents for certain modalities while being closed-loop for others. For example, for some $m < M$, we can sample $z_t^{1:m} \sim q_\phi(z_t^{1:m}|z_{t-1}^{1:M}, o_t^{1:m})$ closed-loop while $z_t^{m+1:M} \sim p_\theta(z_t^{m+1:M}|z_{t-1}^{1:M})$ is open-loop. When open-loop latents $z_t^{m+1:M}$ are used to transition latent dynamics, we call this mechanism "prior-forcing," which enables *latent imagination*. Overall, we train the following model components:

$$\begin{aligned}
\text{closed-loop latents} &: z_t^m \sim q_\phi(z_t^m|z_{t-1}^{1:M}, o_t^m) \quad \forall\, m \in [1, M] \\
\text{open-loop latents} &: \hat{z}_t^m \sim p_\theta(z_t^m|z_{t-1}^{1:M}) \quad \forall\, m \in [1, M] \\
\text{action decoder} &: a_t \sim p_\theta(a_t|z_t^{1:M})
\end{aligned}$$

Training objective. Since computing the log-likelihood of data in the assumed model (1) is intractable, we use ELBO as our training objective:

$$\ln p_\theta(a_{1:T}^*, o_{1:T}^{1:M}) \geq \sum_{t=1}^{T} \Big(\underbrace{\mathbb{E}_{q_\phi(z_t^{1:M}|z_{t-1}^{1:M}, o_t^{1:M})}\big[\ln p_\theta(a_t^*|z_t^{1:M})\big]}_{\mathcal{D}_{\text{LL}}}$$
$$- \underbrace{\mathbb{E}_{q_\phi(z_{t-1}^{1:M}|z_{<t-1}^{1:M}, o_{\leq t-1}^{1:M})}\Big[\text{KL}[q_\phi(z_t^{1:M}|z_{t-1}^{1:M}, o_t^{1:M})||p_\theta(z_t^{1:M}|z_{t-1}^{1:M})]\Big]}_{\mathcal{D}_{\text{KL}}} \Big). \tag{2}$$

We illustrate training in Figure 3. Note that we skip the observation reconstruction term since we set the decoder variance to inf, thus choosing the minimal loss needed for policy learning and latent imagination. Finally, our training loss is, $\mathcal{L}(\mathfrak{D}; \theta, \phi) := \mathcal{D}_{\text{NLL}} + \beta \mathcal{D}_{\text{KL}}$, where $\mathcal{D}_{\text{NLL}} := -\mathcal{D}_{\text{LL}}$ and β controls regularization during training.

Implementation details. We process image and depth inputs using a ResNet18 [27] convolutional network, and use low-dim inputs as-is, then pass them into separate MLP layers to compute individual closed-loop latents $q_\phi(z_t^m|z_{t-1}^{1:M}, o_t^m)$ for each modality m. Each MLP is followed by two small MLPs to output mean and variance. Similar to [5, 21], we

implement open-loop latent dynamics $p_\theta(z_t^m|z_{t-1}^{1:M})$ using a GRU [28]. The GRU network is shared between p_θ and q_ϕ. Finally, we fuse per-modality latent samples $z_t^{1:M}$ (computed using the re-parameterization trick), using a multi-head attention Modality Transformer followed by another small MLP. To aid training visual representations, we weigh gradients on p_θ and q_ϕ differently by setting $\hat{\mathcal{D}}_{\text{KL}}(q_\phi, p_\theta) := \gamma\mathcal{D}_{\text{KL}}(q_\phi, \text{stop\_grad}(p_\theta)), +(1-\gamma)\mathcal{D}_{\text{KL}}(\text{stop\_grad}(q_\phi), p_\theta)$ to get final loss as $\hat{\mathcal{L}}(\mathfrak{D}; \theta, \phi) := \mathcal{D}_{\text{NLL}} + \beta\hat{\mathcal{D}}_{\text{KL}}$, where $\gamma \in [0, 1]$ (lower values pushing p_θ toward q_ϕ more than vice-versa, relaxing regularization and preventing posterior collapse). We found β and γ easy to tune, and $\beta = 10^{-4}$ and $\gamma = 0.1$ to work well in our experiments.

B. Latent Imagination & Adapting to Informative Modalities

Latent imagination for multimodal robustness. The modular latent space in BC-NOSTRA helps mitigate single-modality corruptions during rollout, e.g., blocked camera or noisy depth. If a modality m is noisy, we sample open-loop $\hat{z}_t^m \sim p_\theta(z_t^m|z_{t-1}^{1:M})$ and closed-loop $z_t^{-m} \sim q_\phi(z_t^{-m}|z_{t-1}^{1:M}, o_t^k)$ where $-m$ is short for $\{1, \ldots, M\} \setminus \{m\}$. After fusing the open- and closed-loop latents in the Modality Transformer, the action decoder samples $a_t \sim p_\theta(a_t|\hat{z}_t^m, z_t^{-m})$ (policy output). This mechanism allows the robot to stay privy of all informative inputs while staying robust against noisy ones. We call this *multimodal latent imagination*, or MLI. MLI emerges from training on entirely noise-free multimodal inputs, i.e. training to decode closed-loop latents only. In Sec. IV-A, we show that MLI strictly strengthens the robustness of the visuomotor policy, compared to baselines that use noisy inputs as-is, or ablations that swap the joint latent state for an imagined one.

Adaptive multimodal latent imagination. To deploy BC-NOSTRA with multimodal robustness, we require a metric to identify unseen noises in each modality. We use the per-modality KL divergence $\mathcal{D}_{\text{KL}}^m(t) := \text{KL}[q_\phi(z_t^m|z_{t-1}^{1:M}, o_t^m)||p_\theta(z_t^m|z_{t-1}^{1:M})]$ to identify such noises. We build a mechanism based on the intuitive observation that a sharp rise in $\mathcal{D}_{\text{KL}}^m(t)$ suggests modality m is noisy, using which we decide whether to use open- or closed-loop latent for each modality at each timestep. We call this mechanism AdaMLI. Concretely, at timestep T, AdaMLI maintains a running average of KL changes $r(T, k_c) = \frac{1}{k_c}\sum_{t=T-k}^{T}|\mathcal{D}_{\text{KL}}^m(t) - \mathcal{D}_{\text{KL}}^m(t-1)|$ up to k_c steps in the past. BC-NOSTRA switches from closed- to open-loop latents for a modality if $\mathcal{D}_{\text{KL}}^m(T) - \mathcal{D}_{\text{KL}}^m(T-1) > f_o * r(T, k)$, since this large change in KL would be attributed to a large unseen noise in that modality. To switch back to closed-loop latents, it resets the running average $r(T, k_o)$, and switches latents if $\mathcal{D}_{\text{KL}}^m(T-1) - \mathcal{D}_{\text{KL}}^m(T) > f_c * r(T, k)$, i.e. if the per-modality KL drops significantly. Note that we compute both latents at each timestep to compute the KL, i.e. we always observe the input but may or may not route extracted information to the action decoder. Thresholds and running window sizes were easy to tune; we found $f_o = 20, k_o = 10, f_c = 2, k_c = 2$ worked well for all experiments (sensitivity analysis in Figure 4).

Adapting to informative modalities. With robotic datasets, not all observations $(o^1, ..., o^M)$ are equally informative for different tasks or even different phases of completing a task (i.e. heterogeneous). For instance, during the grasping phase of the *coffee* task (see Figure 6), precise manipulation requires the model to focus on inputs from the wrist-view camera rather than the broader agent-view. With a single latent space representing all modalities, their individual influence would need to be learned and expressed in the latent. In contrast, BC-NOSTRA, with modular latents, allows the decoder to focus its learning on extracting the appropriate amount of information from each modality for the action decoding at hand, making learning more sample efficient. Moreover, as we show in Sec. IV-A, this enables our model to learn efficiently from large datasets containing non-informative inputs, in turn enabling better performance when co-training with diverse datasets.

IV. EXPERIMENTS

We seek to validate four key hypotheses. **H1**: Multimodal Latent Imagination (MLI) in NOSTRA enables robustness against noisy inputs, achieving higher success in manipulation tasks compared to baselines ranging from naive modality-dropout to latent imagination. **H2**: NOSTRA can identify when noises occur, and enable adaptively employing MLI over RGB, depth, and robot proprioception (AdaMLI). **H3**: Multimodal state-space enables NOSTRA to effectively pretrain from hetereogenous datasets that contain extra modalities uninformative to a downstream task. **H4**: BC-NOSTRA can be deployed on a real robot, with robustness capabilities useful in real-world manipulation. We evaluate on 12 simulated benchmark tasks and deploy NOSTRA to a real robot with 4 challenging manipulation tasks.

Tasks. We evaluate BC-NOSTRA on 12 MuJoCo-based tasks in simulation, including 4 MimicGen [29] tasks (*Stack D1*, *Stack Three D1*, *Square D2*, and *Coffee D2*) that contain RGB and robot proprioception, and 8 MimicLabs [30] tasks (see Table II) that additionally contain depth maps. We use 1000 demonstrations for each MimicGen task and 200 for MimicLabs tasks, made available by these works. We also show co-training experiments on 2 tasks from the MimicLabs benchmark. Our policies are trained on upto 6 input modalities: agent-view RGB, wrist-view RGB, agent-view depth, end-effector (EEF) position, EEF orientation, and gripper width. On a real Franka Emika Panda robot, we perform 4 tasks: *lift block*, *serve snack*, *marker in cup*, and *pour beans*. We collect 30 demonstrations for each task using a Meta Quest 2 headset and controller.

Baselines. We use BC-LSTM [1], Diffusion Policy (DP) [2], Diffusion Forcing [8] as baselines in our experiments. To ablate our method, we construct a baseline without modular latents but with latent imagination capability on the joint latent - we call this BC-RSSM since this largely shares model design choices with the recurrent state-space model (RSSM) in [21]. We trained *modality-dropout* variants of DP and BC-NOSTRA by dropping individual modalities with probability 0.5 and replacing with zero-tensors of same shape.

Training details. We train all models using the Adam [31] optimizer with learning rate of $1e$-4. Diffusion Policy is trained using AdamW [32] with a half-cosine decay schedule–all other configurations match with original work. All models are trained

Table I: **MLI leads to visual robustness.** Showing efficacy of multimodal latent imagination (MLI) on RoboMimic tasks each with 4 different test-time setups: (1) *no noise*, (2) *mask*: square black mask on camera observations, (3) *cam jitter*: random perturbation to camera pose, (4) *table tex*: table texture changed to an unseen one. Showing peak performance in 600 epochs of training, averaged over 50 rollouts. Best avg. performance across tasks within 5% is in bold.

Method	no noise					mask					cam jitter					table tex				
	Stack D1	*Stack 3 D1*	*Square D2*	*Coffee D2*	Avg.	*Stack D1*	*Stack 3 D1*	*Square D2*	*Coffee D2*	Avg.	*Stack D1*	*Stack 3 D1*	*Square D2*	*Coffee D2*	Avg.	*Stack D1*	*Stack 3 D1*	*Square D2*	*Coffee D2*	Avg.
BC-LSTM [1]	100	86	46	62	73.5	2	20	2	6	7.5	10	6	6	20	10.5	2	14	2	4	5.5
Diffusion Policy [2]	100	90	58	70	**79.5**	44	66	8	8	31.5	72	68	26	16	45.5	64	60	40	22	46.5
Diffusion Policy [2] + mod. drop.	90	58	10	66	56.0	24	30	2	12	17.0	66	36	2	38	35.5	56	42	2	30	32.5
Diffusion Forcing [8]	76	22	6	8	28.0	50	8	0	0	14.5	70	16	2	0	22.0	48	2	0	0	12.5
Diffusion Forcing [8] + prior forcing	-	-	-	-	-	72	18	4	4	24.5	72	18	4	4	24.5	72	18	4	4	24.5
BC-RSSM	100	86	48	66	75.0	28	38	14	0	20.0	42	66	12	20	35.0	0	0	20	8	7.0
BC-RSSM + LI	-	-	-	-	-	62	64	20	26	43.0	62	64	20	26	43.0	62	64	20	26	43.0
BC-NOSTRA + mod. drop. (ours)	98	86	54	72	**77.5**	34	76	48	50	52.0	48	58	46	44	49.0	48	72	28	24	43.0
BC-NOSTRA (ours)	100	90	54	74	**79.5**	22	68	18	0	27.0	24	66	22	16	32.0	10	34	0	2	11.5
BC-NOSTRA + MLI (ours)	-	-	-	-	-	84	80	38	38	**60.0**	84	80	38	38	**60.0**	84	80	38	38	**60.0**

for 300k gradient steps on a single NVIDIA A40 GPU. Models use ResNet18 vision encoder on 84x84 images with cropping augmentation; LSTM, RSSM, NOSTRA have latents of size 80 per modality, and use sequence length 16 for training.

A. Main Results

Multimodal latent imagination enables NOSTRA to be robust against unseen visual noises. We evaluate models on 16 distinct task instances created by adding 3 types of visual noise to 4 simulated tasks (Table I): black masks for camera occlusion, random camera position jitter, and changing table textures. Noise is added for multiple durations to hinder performance at critical stages of the task – $t \in [25, 75) \cup [100, 150)$ for *Stack D1* and $t \in [75, 100) \cup [125, 150)$ for all other tasks. On the base tasks, BC-NOSTRA performs comparably to or better than BC-LSTM and Diffusion Policy. Under noise, we find that latent imagination (LI) provides significant help for robustness. BC-NOSTRA with MLI surpasses Diffusion Policy by up to 30% on *Square D2* and *Coffee D2* and by an impressive 40% on *Stack D1* on masking noise. BC-NOSTRA +MLI outperforms BC-RSSM+LI by up to 22% (∼17% on average), demonstrating the advantage of a separable multimodal latent space that effectively filters away information from noisy inputs. Under noisy settings, the modality-dropout variant of BC-NOSTRA performs better than the base model, showing that it is a valid approach for robustness. However MLI can significantly outperform by replacing individual latents with imagined ones. We also test statistical significance of BC-NOSTRA performance across 5 seeds in Table IV obtaining <3% standard deviation.

Adaptive latent imagination enables robustness against multimodal noise. While our initial tests showed strong robustness to purely visual noise using MLI, they relied on an "oracle" to detect when noise occurred – an unscalable assumption. To address this, we developed AdaMLI, a mechanism that automatically identifies noise across any modality and adaptively switches to open-loop latents to maintain robustness. We tested BC-NOSTRA +AdaMLI on 8 MimicLabs tasks with noise added to RGB images, depth maps, and robot proprioception. Table II shows that BC-NOSTRA+AdaMLI significantly outperforms all other methods, beating Diffusion Policy by more than 20% on average and up to 60% on long-horizon tasks. This demonstrates its ability to handle noise irrespective of modality, offering a truly scalable solution. Moreover, Table III shows that AdaMLI does not hurt performance even in the absence of noise with <3% false positives switching to AdaMLI. We also perform sensitivity analysis of AdaMLI parameters in Figure 4.

BC-NOSTRA outputs task-relevant actions in the absence of image inputs for extended durations. Figure 5 shows the (x, y, z) positions of the robot end-effector during rollout on the *Square* task with *mask* noise for $t \in [75, 110]$. The end-effector trajectory using BC-NOSTRA with latent imagination on the image inputs closely matches that when no noise was added. This shows that the open-loop latent maintains task-relevant information even in the absence of image inputs for extended durations (35 timesteps in this case) leading to successful task execution which would otherwise be hindered by visual noise.

Multimodal latents in NOSTRA ignore non-informative input modalities. We train BC-NOSTRA on 2 variants of 4 simulated tasks: one using a standard, informative agent-view camera, and another where the robot self-occludes the agent-view making it non-informative. Results in Table V show that BC-NOSTRA effectively identifies when the agent-view image is non-informative. It reduces the information stored in the corresponding latent space by up to 32%, measured by the per-modality KL divergence $\mathcal{D}_{\text{KL}}^{m}$ (in nats), while simultaneously increasing the information extracted from the remaining modalities. This demonstrates BC-NOSTRA's ability to adapt its latent representations to ignore non-informative inputs from a diverse set of modalities.

BC-NOSTRA enables better pre-training with non-informative inputs. We pre-trained visuomotor policies for 200 epochs on datasets of 1000 expert demos that included an occluded, non-informative agent-view camera. We then fine-tuned these models on 10, 20, 50 demos containing a useful agent-view image. As shown in Table VI, BC-NOSTRA consistently outperforms BC-RSSM (which lacks modular latents) by up to 30% when fine-tuning on just 10 demos. This strong performance demonstrates that the modular latent structure of BC-NOSTRA learns effective representations even from heterogeneous inputs, enabling superior performance on downstream tasks.

BC-NOSTRA learns effectively when co-training with diverse large-scale datasets. We evaluated our policies on two MimicLabs tasks: *bin bowl* and the longer-horizon *open drawer & put bowl*. For each task, we co-trained the policies with two different retrieved datasets: one large and diverse ($\mathcal{O}(10k)$ demonstrations) aligned by object/skill, and

Table II: **AdaMLI for multimodal noises.** We show success rates (SR) on 8 MimicLabs tasks with unseen noising in different sets of multimodal inputs: "RGB" (masking agent-view and wrist-view images), "RGBD" (additional masking depth maps), and "All" (additionally zero-ing out robot proprioception to emulate sensor failure). BC-NOSTRA measures noise in individual inputs and adaptively employs MLI to maintain high success for this suite of prehensile, non-prehensile, and long-horizon tasks.

			bin carrot			**bin bowl**			**open drawer**			**close drawer**		
Method	Avg. SR↑	Avg. Rank↓	RGB	RGBD	All	RGB	RGBD	All	RGB	RGBD	All	RGB	RGBD	All
BC-LSTM [1]	63.8	3.1	52	30	30	74	76	80	100	100	100	100	100	100
Diffusion Policy [2]	61.3	3.9	40	40	46	56	54	60	92	94	98	98	98	98
BC-RSSM	59.9	3.4	64	42	52	88	96	90	92	90	92	46	58	60
BC-NOSTRA (ours)	70.7	2.7	74	76	78	78	86	82	76	68	86	100	100	100
BC-NOSTRA + AdaMLI (ours)	**86.5**	**1.4**	86	82	86	92	96	90	92	100	96	100	100	100

	open microwave			**close microwave**			**open drawer & put bowl**			**put bowl & close drawer**		
Method	RGB	RGBD	All	RGB	RGBD	All	RGB	RGBD	All	RGB	RGBD	All
BC-LSTM [1]	60	52	46	46	46	56	44	56	56	40	40	46
Diffusion Policy [2]	44	48	50	68	62	64	22	18	14	68	70	70
BC-RSSM	88	68	82	52	34	20	40	54	40	22	36	32
BC-NOSTRA (ours)	52	70	58	88	80	84	64	60	48	28	26	34
BC-NOSTRA + AdaMLI (ours)	76	66	78	80	86	88	72	58	78	98	86	90

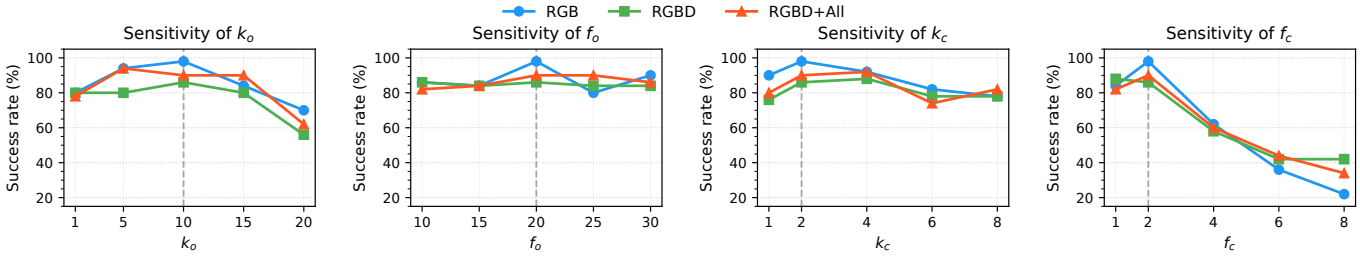


Figure 4: **Sensitivity analysis of AdaMLI parameters.** Success rates when noise was added to RGB, RGBD, or RGBD+All in the *put bowl & close drawer* task for different values of AdaMLI parameters around $k_o = 10, f_o = 20, k_c = 2, f_c = 2$. Selected values performed best, with success less sensitive to change in k_o, f_o, k_c compared to f_c.

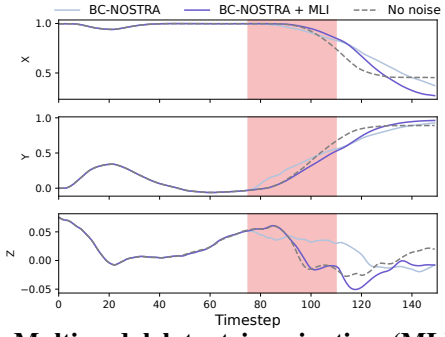


Figure 5: **Multimodal latent imagination (MLI) predicts accurate actions over extended durations.** We noise inputs for $t \in [75, 110]$ (red region) in the *Square* task, and show end-effector trajectories output by BC-NOSTRA, with MLI (—) and without (—), as well as BC-NOSTRA on clean inputs (- -). MLI closely matches robot's trajectory to what it could have taken had there been no noise.

Table III: Success rate drops minimally when AdaMLI is enabled in noise-free settings, attributed to very low false-positive rates (FPR) for detecting noise using our per-modality KL-based mechanism.

			FPR↓	
Task	**SR↑**	**SR↑ w/ AdaMLI**	**Agent-view RGB**	**Wrist-view RGB**
Stack D1	100%	94%	0.4%	2.4%
Stack Three D1	90%	84%	1.3%	2.0%

Table IV: BC-NOSTRA performance across multiple seeds. Selected checkpoint had max success in 600 epochs of training.

Task	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Mean ± Std
Stack D1	100	100	100	100	100	100.0 ± 0.0
Stack Three D1	90	96	88	92	94	92.0 ± 2.8
Square D2	54	58	60	58	60	58.0 ± 2.2
Coffee D2	74	72	70	72	78	73.2 ± 2.7

a smaller, more specific one ($\mathcal{O}(1k)$ demonstrations) aligned by camera poses and spatial arrangements. Diffusion Policy experienced a significant drop in performance when co-trained with the larger, more diverse dataset on the long-horizon *open drawer & put bowl* task. In contrast, BC-NOSTRA showed a consistent performance boost, achieving a 14% improvement and outperforming a joint latent space model, BC-RSSM, by 8%. Similar gains were observed on the shorter-horizon *bin bowl* task. This consistent improvement highlights the benefits of our modular latent space design in effectively leveraging diverse, heterogeneous data.

Input type	Stack D1		Stack Three D1		Square D2		Coffee D2	
	good view	*occl. view*	*good view*	*occl. view*	*good view*	*occl. view*	*good view*	*occl. view*
Wrist-view RGB	4.066	4.847 (+19.2%)	4.446	5.794 (+30.3%)	4.374	4.909 (+12.2%)	4.349	4.684 (+7.7%)
Agent-view RGB	3.033	2.401 (-20.8%)	3.981	2.707 (-32.0%)	2.876	2.580 (-10.3%)	2.954	2.476 (-16.2%)
End-effector ori.	0.740	1.039 (+40.4%)	1.025	1.820 (+77.5%)	0.919	1.259 (+36.9%)	0.452	0.643 (+42.3%)
End-effector pos.	0.347	0.673 (+94.0%)	0.401	1.083 (+169.9%)	0.326	0.598 (+83.5%)	0.278	0.335 (+20.7%)
Gripper width	0.270	0.302 (+11.9%)	0.324	0.462 (+42.7%)	0.304	0.348 (+14.2%)	0.283	0.312 (+10.0%)
Total	8.456	9.262 (+9.5%)	10.177	11.865 (+16.6%)	8.800	9.693 (+10.1%)	8.316	8.451 (+1.6%)

Table V: **Adapting to informative modalities.** We compare per-modality latent information in two task variants–with an informative "good view" camera, vs. an "occluded view" where the robot self-occludes the camera–as KL divergence (in nats; 200 training epochs). BC-NOSTRA's multimodal latent space adaptively reduces the information extracted from the non-informative, occluded view, while increasing focus on other useful inputs.

Table VI: Success rates when fine-tuning behavior cloning policies using pre-trained checkpoints, trained for 200 epochs on a dataset containing a non-informative agent-view image. Results show that a modular latent space almost always leads to more successful fine-tuning.

	Stack D1			Stack Three D1			Square D2			Coffee D2		
Num. demos →	**10**	**20**	**50**	**10**	**20**	**50**	**10**	**20**	**50**	**10**	**20**	**50**
BC-LSTM [1]	28	62	72	2	2	20	**2**	4	**12**	6	18	24
BC-RSSM	14	16	6	2	2	4	0	0	0	0	4	8
BC-NOSTRA (ours)	**44**	**74**	**90**	**4**	**14**	**22**	**2**	**12**	**12**	**24**	**34**	**46**

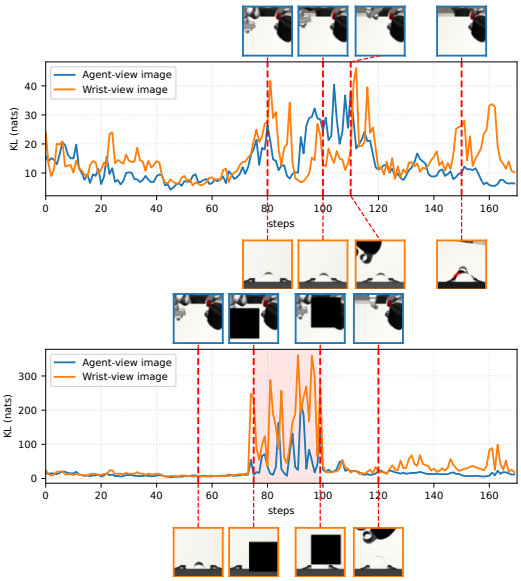
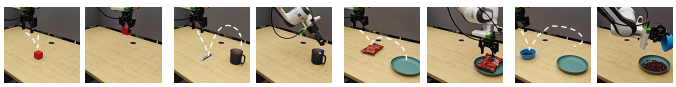


Figure 6: KL divergence (in nats) as a measure of information extracted from input images at different stages of the *Coffee* task. **Top:** *t=80*: picking up pod; *t=100*: moving towards coffee machine, agent-view more informative; *t=110*: coffee machine visible in wrist-view, both views useful; *t=150* - fine-tuned placing, wrist-view useful but agent-view not. **Bottom:** masking noise in $t \in [75, 100)$.



1) Lift 2) Marker in Cup 3) Serve Snack 4) Pour Beans

Figure 7: **Real-world tasks.** We evaluate our method and baselines on 4 tasks using a Franka Emika Panda robot. We show the initial state (left) and final state (right) per task.

Table VII: **Co-training with the MimicLabs Dataset.** We used 10 target demos for the *bin bowl* task and 20 for the *clear table* task. Showing peak performance over 600 epochs of training, averaged over 50 rollouts. *Obj/Skill* refers to retrieving demos with matching objects and skills, *+all* refers to additional retrieval to align camera pose and placement arrangements.

	bin bowl			open drawer & place bowl		
Method	*Target only*	*Obj/Skill*	*+all*	*Target only*	*Obj/Skill*	*+all*
BC-LSTM [1]	**38**	**52** (+14)	60 (+22)	42	40 (-2)	**58** (+16)
Diffusion Policy [2]	10	30 (+20)	30 (+20)	**48**	28 (-20)	26 (-22)
BC-RSSM	**40**	38 (-2)	50 (+10)	40	46 (+6)	38 (-2)
BC-NOSTRA (ours)	**38**	44 (+6)	**66** (+28)	**52**	**66** (+14)	**54** (+2)

B. Real-Robot Experiments

We trained BC-NOSTRA, BC-RSSM, and Diffusion Policy on 4 real-world tasks requiring a variety of motion skills (lifting, picking and placing, pouring) (see Figure 7). On 3 tasks, we show model performance when the camera was occluded by hand, and BC-NOSTRA used AdaMLI to stay robust. Results in Table VIII are averaged over 10 trials.

BC-NOSTRA learns stable and sample efficient policies in the real world. Using just 30 human demos, BC-NOSTRA is stable to train and learns policies that outperform Diffusion Policy on 2 out of 4 tasks and is at-par with one other. The modular latent space is crucial since we found BC-RSSM with the joint latent space to undergo unstable training, resulting in no success even on the easy *lift* task. We believe BC-NOSTRA introduces structure to the latent space which acts as regularization to aid stable training.

BC-NOSTRA enables robustness against camera occlusion in the real-world We evaluated all policies in a real-world setting with an external occlusion of the wrist-view camera by a human hand. As shown in Table VIII, BC-NOSTRA with AdaMLI successfully detected and adapted to this unforeseen noise, achieving a success rate that nearly matched its performance without occlusion. In contrast, Diffusion Policy failed to complete two out of three tasks – under occluded camera, the robot made jerky or non-motions causing errors to accumulate. BC-NOSTRA avoided this failure mode by switching to open-loop latent imagination, effectively ignoring the noisy camera

Table VIII: Success rates (avg. over 10 trials) on 4 real-robot tasks; *base* (no noise) and with test-time *camera occlusion*.

Task	**lift**	**marker in cup**		**serve snack**		**pour beans**	
		base	*cam occl.*	*base*	*cam occl.*	*base*	*cam occl.*
Diffusion Policy [2]	60	30	0	50	0	**70**	40
BC-RSSM	0	0	0	20	0	40	40
BC-NOSTRA (ours)	**60**	**50**	**40**	**60**	**40**	50	**50**

input and maintaining high task success.

V. Conclusion

We presented BC-Nostra, a visuomotor policy that leverages a multi-modal state-space model to enable modality-level robustness to noise, and learns latent embeddings that adapt to the heterogeneity in a manipulation task. We proposed a novel inference strategy called Adaptive Multimodal Latent Imagination (AdaMLI) that allows our visuomotor policy to identify noisy inputs and be robust against them, without any explicit training with noisy or missing modalities. Leveraging its modular latent space, BC-Nostra offered an effective methodology for fine-tuning visuomotor policies pre-trained or co-trained with diverse robotic datasets. Our experiments strongly indicated that the ideas proposed in this paper can result in robust visuomotor policies, and while proven extensively in single-task settings, our design decisions are general, and will pave the way for the development of robust large-scale multi-task behavior models.

References

[1] Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *Conference on Robot Learning (CoRL)*, 2021. 1, 2, 3, 4, 5, 6, 7

[2] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*, 2023. 1, 2, 3, 4, 5, 6, 7

[3] Zihao He, Hongjie Fang, Jingjing Chen, Hao-Shu Fang, and Cewu Lu. Foar: Force-aware reactive policy for contact-rich robotic manipulation, 2024. URL https://arxiv.org/abs/2411.15753. 1

[4] Yuqing Du, Daniel Ho, Alexander A. Alemi, Eric Jang, and Mohi Khansari. Bayesian imitation learning for end-to-end mobile manipulation, 2022. URL https://arxiv.org/abs/2202.07600. 1

[5] Vaibhav Saxena, Jimmy Ba, and Danijar Hafner. Clockwork variational autoencoders. *CoRR*, abs/2102.09532, 2021. URL https://arxiv.org/abs/2102.09532. 1, 2, 3

[6] Jurgis Pasukonis, Timothy Lillicrap, and Danijar Hafner. Evaluating long-term memory in 3d mazes, 2022. 1

[7] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models, 2022. 1, 2

[8] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024. 2, 4, 5

[9] Vaibhav Saxena, Yotto Koga, and Danfei Xu. C3DM: Constrained-Context Conditional Diffusion Models for Imitation Learning. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=jcleXdnRA1.

[10] Eugenio Chisari, Nick Heppert, Max Argus, Tim Welschehold, Thomas Brox, and Abhinav Valada. Learning robotic manipulation policies from point clouds with conditional flow matching. *arXiv preprint arXiv:2409.07343*, 2024.

[11] Qinglun Zhang, Zhen Liu, Haoqiang Fan, Guanghui Liu, Bing Zeng, and Shuaicheng Liu. Flowpolicy: Enabling fast and robust 3d flow-based policy via consistency flow matching for robot manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 14754–14762, 2025. 2

[12] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016. 2

[13] Junbang Liang, Ruoshi Liu, Ege Ozguroglu, Sruthi Sudhakar, Achal Dave, Pavel Tokmakov, Shuran Song, and Carl Vondrick. Dreamitate: Real-world visuomotor policy learning via video generation, 2024. URL https://arxiv.org/abs/2406.16862. 2

[14] Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. *arXiv preprint arXiv:2412.14803*, 2024.

[15] Yunhao Luo and Yilun Du. Grounding video models to actions through goal conditioned exploration. *arXiv preprint arXiv:2411.07223*, 2024. 2

[16] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation, 2022. URL https://arxiv.org/abs/2209.05451. 2

[17] I Liu, Chun Arthur, Sicheng He, Daniel Seita, and Gaurav Sukhatme. Voxact-b: Voxel-based acting and stabilizing policy for bimanual manipulation. *arXiv preprint arXiv:2407.04152*, 2024. 2

[18] Haoyi Zhu, Yating Wang, Di Huang, Weicai Ye, Wanli Ouyang, and Tong He. Point cloud matters: Rethinking the impact of different observation spaces on robot learning. *Advances in Neural Information Processing Systems*, 37:77799–77830, 2024. 2

[19] Skand Peri, Iain Lee, Chanho Kim, Li Fuxin, Tucker Hermans, and Stefan Lee. Point cloud models improve visual robustness in robotic learners. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 17529–17536. IEEE, 2024. 2

[20] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model, 2024. URL https://arxiv.org/abs/2406.09246. 2

[21] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pages 2555–2565. PMLR, 2019. 2, 3, 4

[22] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. *NeurIPS*, 2018. 2

[23] Yuge Shi, N. Siddharth, Brooks Paige, and Philip H. S. Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models, 2019. URL https://arxiv.org/abs/1911.03393.

[24] Thomas M. Sutter, Imant Daunhawer, and Julia E Vogt. Generalized multimodal ELBO. In *ICLR*, 2021. URL https://openreview.net/forum?id=5Y21V0RDBV. 2

[25] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Bejing, China, 22–24 Jun 2014. PMLR. 3

[26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114. 3

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[28] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014. URL https://arxiv.org/abs/1406.1078. 4

[29] Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations, 2023. URL https://arxiv.org/abs/2310.17596. 4

[30] Vaibhav Saxena, Matthew Bronars, Nadun Ranawaka Arachchige, Kuancheng Wang, Woo Chul Shin, Soroush Nasiriany, Ajay Mandlekar, and Danfei Xu. What matters in learning from large-scale datasets for robot manipulation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LqhorpRLIm. 4

[31] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 4

[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101. 4