MDF++: Visuo-Tactile World Action Model with End-to-end Representation Learning

Zixuan Huang **Jinhao Yu** **Zachary Charlick** **Dmitry Berenson**
Department of Robotics
University of Michigan, Ann Arbor, United States
{zixuanh,rickyu,zsc,dmitryb}@umich.edu

Abstract: Recently, world-action models (WAMs) have emerged as a powerful framework for robot manipulation. By jointly modeling action generation and future observation prediction, WAMs can act and anticipate the outcomes of the actions. However, existing WAMs typically rely on frozen pretrained encoders trained with reconstruction, which may not produce representations well suited for robot control. In this paper, we tackle the problem of learning physically-grounded, control-relevant representations for WAMs. We propose MDF++, a visuo-tactile world-action model that learns latent representations of multisensory observations end-to-end with a diffusion backbone. While naive joint training can lead to feature collapse, MDF++ introduces a multi-task learning strategy that leverages cross-modal correlations to learn task-relevant representations. We evaluate MDF++ on a suite of contact-rich manipulation tasks, including real-world bolt screwing and nut mating, and show that it outperforms WAMs built on frozen pretrained representations by up to 50%. Additional videos are available on our website: jointmdf.github.io.

Keywords: Multimodal World Model, Contact-rich Manipulation

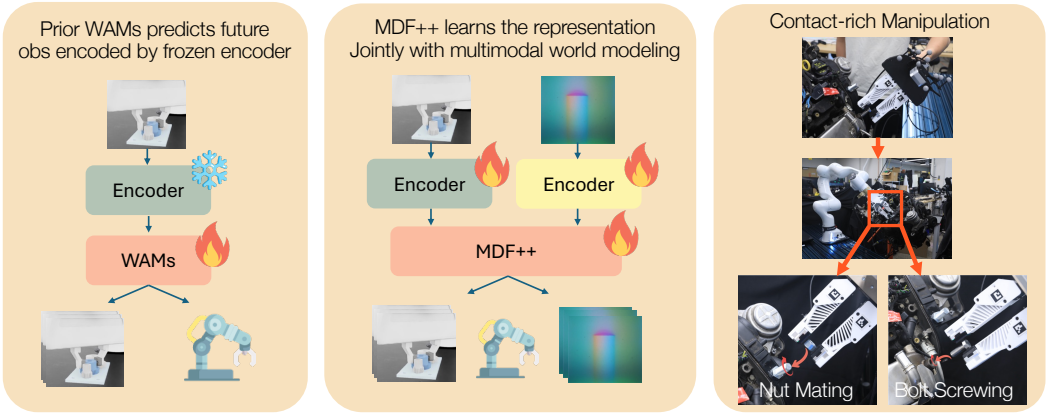


Figure 1: Existing WAMs typically predict future observations encoded by a frozen pretrained encoder optimized for visual generation or semantic understanding, which maybe suboptimal for robot manipulation. We propose MDF++, a visuo-tactile world-action model that learns task-aware representations end-to-end with cross-modality prediction.

1 Introduction

When tightening a screw or mating a nut, vision alone is often insufficient: successful execution depends on subtle contact cues, force feedback, and proprioceptive information. In order to perform such tasks, it would also be useful to anticipate how the robot's actions affect the world, such as how contact will form, how an object will move, and how force will propagate. This motivates robot

10th Conference on Robot Learning (CoRL 2026), Austin TX, USA.

models that jointly reason about actions and their consequences. World-action models (WAMs) [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] address this by jointly modeling future observations and actions within a single generative model. Recent work [2, 9, 10] shows that such auxiliary dynamics learning can improve model capability and generalization in robot control.

However, existing WAMs still have two key limitations for robot manipulation. First, most prior WAMs are built primarily around visual observations, leaving tactile and force information largely unexplored. Second, these models typically rely on frozen pretrained encoders, often trained with pixel reconstruction or self-supervised visual objectives. While such encoders provide useful features for image generation or semantic understanding, their representations are not necessarily aligned with robot control. This issue is even more pronounced for tactile, where broadly-applicable pretrained encoders are less established. A natural solution is to finetune the encoders end-to-end with the WAM objective, but we find that naive finetuning leads to feature collapse: the encoder and diffusion backbone co-adapt to produce latents that are easy to denoise but less useful for control.

To address the above limitations, we present MDF++, a visuo-tactile world-action model that learns task-relevant representations end-to-end with a latent diffusion backbone. We build on Multimodal Diffusion Forcing (MDF) [10], but instead of modeling dynamics in a frozen latent space, MDF++ finetunes observation encoders jointly with the world-action model under a mixture of task-oriented objectives, including policy learning, joint action–observation prediction, forward dynamics modeling, and inverse dynamics modeling. To avoid feature collapse during joint training, MDF++ introduces a cross-modal training strategy with selective gradient back-propagation, which encourages the latent space to capture the interplay between different modalities, yielding representations that are better aligned with robot control.

Our key contributions are three-fold. First, we propose MDF++, a diffusion-based visuo-tactile WAM that learns task-relevant multimodal representations end-to-end using cross-modal prediction. Second, we introduce a selective gradient-flow strategy that enables stable joint training of observation encoders and the world-action model while avoiding feature collapse. Third, we evaluate MDF++ on seven contact-rich manipulation tasks, including two real-world threading tasks requiring millimeter-level precision. In simulation, MDF++ outperforms frozen-encoder baselines by 28% and naive joint training by 22%. In real, it outperforms video-based WAMs [9] by 29%.

2 Related Work

2.1 World modeling for Robotics

Recently, world models have shown great promise for robotics. One line of work studies action-conditioned world models (WAs) to predict the future outcomes of actions, enabling data generation [11, 12], planning [13, 14, 15, 16], and policy evaluation [17, 12, 18]. Another line of work, world action models (WAMs), jointly predicts future observations and actions [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 19]. However, most prior work is limited to vision, whereas MDF++ captures the dynamics of multimodal sequences including tactile, force and robot proprioception.

Beyond the choice of modalities, a key design question is how to obtain the representation space in which world-action dynamics are learned. Most existing WA and WAMs rely on frozen pretrained encoders, typically trained with pixel reconstruction [20] or self-supervised learning objectives [21, 22]. For example, UWM [9] and UVA [2] use pretrained image VAEs [23, 20], Dino-WM [14] predicts future states in the DINO feature space [21], and MDF [10] builds on a diffusion-based point-cloud VAE [24]. While these encoders provide useful features for visual generation or semantic understanding, they are not necessarily aligned with robot control. As a result, the world model may learn dynamics in a representation space that is suboptimal for manipulation. In contrast, MDF++ starts from a pretrained encoder and finetunes it end-to-end with the latent world-action model backbone. Through cross-modal supervision from tactile, force, and actions, MDF++ learns a representation space that is better suited for robot manipulation. FLARE [25] finetunes the encoder using an EMA target, but it is not designed for unified WAMs with multiple query distributions and also does not support multimodal reasoning.

2.2 Multimodal Representation learning for Manipulation

The integration of multi-sensory information, such as audio [26, 27, 28], force [29, 30, 31], is critical for robot control. Particularly, there is growing interest in combining vision and tactile for adaptive and precise manipulation. Much of the prior works learn a visuo-tactile policy using either imitation learning [32, 33, 34, 35, 36] or reinforcement learning [37, 38, 39, 40, 41, 42, 43], where representations are primarily supervised by action prediction or task reward. Other works study self-supervised tactile [44, 45, 46, 35] or visuo-tactile representation learning [47, 36]. In contrast, our work studies representation learning for multimodal world-action models. While these models provide richer supervision through action and cross-modal prediction, naively training encoders with diffusion losses can cause feature collapse. We address this with selective gradient flow and show that the proposed method outperforms policy-only training and pretrained self-supervised encoders [45].

3 Preliminaries

3.1 Latent Diffusion Models

Diffusion models learn a data distribution by gradually corrupting clean samples with Gaussian noise and training a neural network to reverse this process [48]. Given a clean sample $\mathbf{x}_0$, the forward process produces a noisy sample $\mathbf{x}_k = \sqrt{\bar{\alpha}_k}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_k}\boldsymbol{\epsilon}$ at diffusion step k, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\bar{\alpha}_k$ is determined by a noise schedule. The reverse model is trained to denoise $\mathbf{x}_k$ by predicting the injected noise, yielding the standard objective $\mathbb{E}_{k,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_k, k)\|^2\right]$. At inference time, samples are generated by iterative denoising from Gaussian noise. Latent diffusion models perform this denoising process in a learned latent space rather than directly in the input space [23]. A sample $\mathbf{x}_0$ is first encoded into a latent representation $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$, where the encoder $\mathcal{E}$ is typically pretrained and kept frozen during diffusion training.

3.2 Unified World-Action Model

A unified world-action model (WAM) is a single generative model over observations and actions that can be queried under different conditioning patterns. Let $\mathbf{o}_{i:j}$ and $\mathbf{a}_{i:j}$ denote observation and action sequences from timestep i to j, respectively. Given an observation history $\mathbf{o}_{t-h:t}$ and a prediction horizon H, the same model can be used as a policy, a forward dynamics model, an inverse dynamics model, or a joint action–observation predictor by choosing which variables are conditioned on and which variables are generated: **policy** $p(\mathbf{a}_{t+1:t+H} \mid \mathbf{o}_{t-h:t})$, **world-action model** $p(\mathbf{o}_{t+1:t+H}, \mathbf{a}_{t+1:t+H} \mid \mathbf{o}_{t-h:t})$, **world model / forward dynamics** $p(\mathbf{o}_{t+1:t+H} \mid \mathbf{o}_{t-h:t}, \mathbf{a}_{t+1:t+H})$, and **inverse dynamics model** $p(\mathbf{a}_{t+1:t+H} \mid \mathbf{o}_{t-h:t+H})$.

3.3 Problem Formulation

We consider diffusion-based world-action modeling for contact-rich robot manipulation from an offline dataset of expert demonstrations. At each timestep t, the robot receives a multimodal observation $\mathbf{o}_t = \{\mathbf{o}_t^m\}_{m\in\mathcal{O}}$, where $\mathcal{O}$ denotes the set of observation modalities, such as vision, tactile sensing, force/torque sensing, and proprioception. The robot action is denoted by $\mathbf{a}_t$.

High-dimensional observations such as image and tactile are mapped into a latent representation by a modality-specific encoder ϕ_m: $\mathbf{z}_t^m = \phi_m(\mathbf{o}_t^m)$, where $m \in \mathcal{O}$. Together with the action stream, these latent observation tokens define a multimodal trajectory

$$\mathbf{x}_{1:T} = \{(\{\mathbf{z}_t^m\}_{m\in\mathcal{O}}, \mathbf{a}_t)\}_{t=1}^{T}.$$

We consider the problem of learning a unified world-action model from multimodal demonstration trajectories for contact-rich manipulation. Given observations from multiple sensory modalities, our goal is to train observation encoders jointly with a diffusion-based world-action model to capture task-relevant feature. At test time, we evaluate the model primarily through its ability to generate actions for robot manipulation tasks.

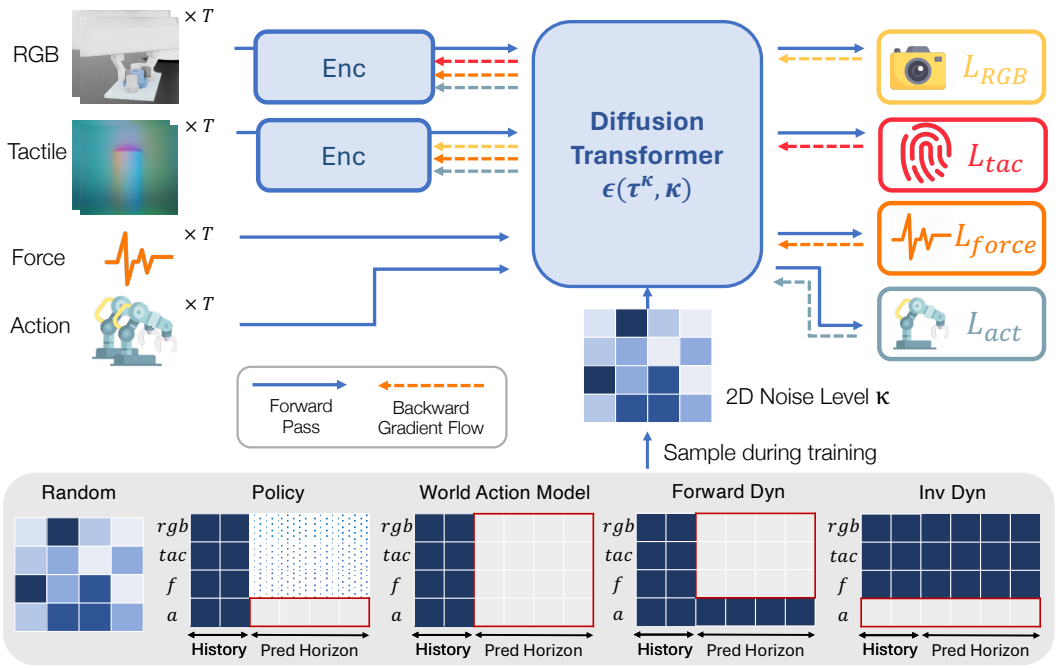


Figure 2: MDF++ training overview. MDF++ jointly trains observation encoders with a latent diffusion world-action model. During training, it samples structured 2D noise patterns for control-relevant tasks, including policy learning, world-action modeling, forward/inverse dynamics, and random masking, to capture temporal and cross-modal dependencies. To avoid collapsed representations that are easy to denoise but weak for control, MDF++ blocks each modality's denoising loss from updating its own encoder. Instead, each encoder is trained through cross-modal prediction losses, encouraging task-relevant cues for predicting other sensory modalities and actions.

4 Joint Representation Learning for Unified World-Action Model

4.1 Multi-task Training with Multimodal Diffusion Forcing

To train a single backbone that supports the diverse conditional queries of a unified WAM (Sec. 3.2), we adopt *Multimodal Diffusion Forcing* (MDF) [10] as our model architecture. Standard diffusion models apply a single scalar noise level k to the entire sample [48], imposing a fixed separation between conditioning inputs and prediction targets. This is restrictive for unified world-action modeling, where one model must support multiple prediction patterns.

MDF addresses this by treating noise as a continuous form of masking [49, 10]: zero noise corresponds to an observed token, intermediate noise to a partially corrupted token, and maximum noise to a fully masked token. By assigning noise levels independently across time and modality, MDF can express different conditional distributions over the same trajectory. Concretely, consider a trajectory of latent multimodal tokens and actions over a horizon T, $\tau = \{(z_t, a_t)\}_{t=1}^{T}$, where each timestep contains M modality streams (i.e., vision, tactile, force, proprioception, and action). MDF defines a 2D *Time–Modality Noise Level Matrix* $\mathbf{K} \in \{0, \dots, K\}^{T \times M}$, where each entry $k_{t,m}$ specifies the diffusion step applied to modality m at time t. The forward process corrupts each token according to its own noise level, $\tau_{t,m}^{k_{t,m}} = \sqrt{\bar{\alpha}_{k_{t,m}}}\,\tau_{t,m}^{0} + \sqrt{1-\bar{\alpha}_{k_{t,m}}}\,\epsilon_{t,m}, \quad \epsilon_{t,m} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. A diffusion transformer ϵ_θ conditioned on both the corrupted trajectory $\tau^{\mathbf{K}}$ and the matrix $\mathbf{K}$ is then trained to predict the per-token noise with the standard DDPM objective [48],

$$\mathcal{L}_{\text{MDF}}(\theta) = \mathbb{E}_{\tau^0, \mathbf{K}, \boldsymbol{\epsilon}} \left[\sum_{t=1}^{T} \sum_{m=1}^{M} \left\| \epsilon_{t,m} - \epsilon_\theta(\tau^{\mathbf{K}}, \mathbf{K})_{t,m} \right\|^2 \right]. \tag{1}$$

While the original MDF training scheme is highly flexible, it samples the noise-level matrix $\mathbf{K}$ from a purely random distribution (bottom left of Fig. 2). This covers a large family of prediction tasks,

since each masking pattern defines a different conditional distribution. However, many of these conditionals are only weakly related to the inference problems needed for robot manipulation.To bias training toward control-relevant queries, we train the model on five structured conditional distributions with equal sampling probability: **policy learning**, **world-action modeling**, **forward dynamics**, **inverse dynamics**, and **random masking**. As shown in Fig. 2, conditioning tokens are assigned noise level zero, while prediction tokens are assigned noise levels sampled uniformly from $[0, K]$. These structured patterns focus model capacity on conditional distributions directly useful for robot control. Although our final downstream objective is action generation for visuomotor control, prior work shows that auxiliary prediction tasks, such as forward and inverse dynamics modeling, can encourage representations to capture task-relevant features and temporal structure in demonstration trajectories [50, 51, 52].

However, **existing WAMs typically cannot fully leverage this multi-task supervision for representation learning**, because they rely on frozen encoders pretrained with reconstruction objectives [53, 2, 9, 10]. For example, MDF [10] uses a frozen point-cloud encoder pretrained with a reconstruction objective. Reconstruction encourages the encoder to preserve information needed to reproduce the input, including task-irrelevant details that may not be useful for downstream control. While this design provides stable geometric features, it decouples representation learning from the control and prediction objectives used by the world-action model. In the next section, we describe how MDF++ leverages this multi-task supervision to train observation encoders end-to-end with the diffusion backbone, yielding action-aware representations for robotic manipulation.

4.2 Cross-modal Training with Selective Gradient Flow

Given a diffusion-based WAM, the most direct way to fine-tune the observation encoders is to backpropagate the diffusion loss through the encoder. However, this naive end-to-end training scheme usually leads to collapsed feature space. Since the same latent tokens serve as both conditioning inputs and prediction targets, gradients from a modality's own denoising loss can encourage the encoder and diffusion backbone to co-adapt toward a latent space that is easy to denoise [54], but weak for control. In our experiments, this direct fine-tuning strategy often degrades task performance.

MDF++ avoids feature collapse with a simple selective gradient-flow rule inspired by REPA-E [54]: the encoder for each observation modality is updated only by diffusion losses from other streams, not by the loss of its own stream. Let $\mathcal{O}$ denote the set of observation modalities and $\mathcal{M} = \mathcal{O} \cup \{a\}$ denote all predicted streams, including actions. Let $\mathcal{L}_s$ be the MDF denoising loss for stream $s \in \mathcal{M}$. For each observation encoder ϕ_o, where $o \in \mathcal{O}$, MDF++ applies the masked gradient

$$\nabla_{\phi_o}^{\text{MDF++}} \mathcal{L} = \sum_{s \in \mathcal{M}} \mathbf{1}[s \neq o] \, \nabla_{\phi_o} \mathcal{L}_s. \tag{2}$$

Equivalently, when computing the denoising loss for modality m, gradients are stopped at the output of its own encoder ϕ_m, while gradients from cross-modal prediction losses are allowed to update it.

This rule trains each encoder to produce representations that are predictive of future signals in other modalities. For example, the visual encoder is not optimized to reconstruct visual appearance itself; instead, it is optimized through losses for predicting other sensory streams and actions. Since predicting tactile feedback, interaction forces, and successful actions requires information about the object, the end-effector, and their potential contact regions, this cross-modal supervision naturally encourages the visual representation to capture manipulation-relevant cues. Because no encoder is trained to denoise its own input, the learned representations are shaped by cross-modal predictive pressure rather than modality-specific reconstruction shortcuts.

Model Architecture. We instantiate τ as a sequence of modality-specific tokens. RGB and tactile images are encoded as CLS tokens from pretrained ViTs [55], while low-dimensional streams such as force, proprioception, and actions are projected with linear layers. At each timestep, all modality tokens are concatenated with their corresponding noise-level embeddings and processed by a diffusion transformer. **Training Recipe**. We identify several design choices that are important for training stability and performance. First, for each sampled trajectory, we sample multiple train-

ing modes to encourage robust learning across different conditioning and prediction settings and improve training efficiency. Second, MDF++ injects different noise levels into latent tokens during diffusion training, we detach the encoder outputs for highly corrupted tokens; gradients from these samples are often weakly correlated with the original observations and can destabilize representation learning. Third, we apply a cosine learning-rate schedule to the encoder and cool it down before the diffusion backbone.

5 Experiments

We evaluate MDF++ on a suite of simulated and real-world visuo-tactile manipulation tasks. Our experiments are designed to answer three questions: (1) Does the proposed selective gradient-flow strategy stabilize joint representation learning? (2) Are the learned task-aware representations more effective than pretrained semantic or reconstruction-based encoders for contact-rich manipulation? (3) Does multi-task training provides a better representation for contact-rich robot manipulation?

5.1 Simulation Experiments

Tasks: We consider five manipulation tasks that requires multimodal reasoning. Three precision-focused tasks are selected from Isaac Lab FACTORY [56]: PegInsert, GearMesh, and NutThread. We adapt the simulation environments from Forge [29] by adding a GelSight Mini tactile sensor. For each task, we collect 1,000 demonstrations using an RL expert policy. We additionally include two tactile-exploration tasks from MANIFEEL [57]. In ObjectSearch, the robot must locate, grasp, and lift a cube-shaped object randomly placed inside a closed container. In BallSorting, the robot must sort two visually similar balls using tactile sensing.

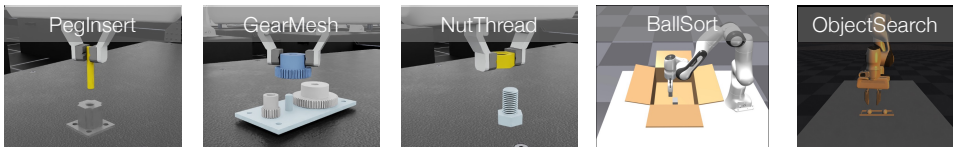
Baselines and ablations We additionally compare against **MDF (adapted)** [10], which uses frozen observation encoders and random masking in the same visuo-tactile setting, and **MDF + selective E2E**, which retains random masking but jointly trains the encoders with selective gradient flow. Their Forge runs use end-of-trajectory padding, unlike the MDF++ reference, so the comparison does not fully isolate the objectives. We compare MDF++ against several variants and representative baselines. **MDF++ (naive grad)** jointly trains the encoders and diffusion model by backpropagating denoising losses from all modalities through all encoders. **MDF++ (SIGReg)** adds SIGReg [58, 15] as a latent-space regularizer to mitigate representation collapse. SIGReg optimizes a normality test over random one-dimensional projections, encouraging the latent embeddings to match an isotropic Gaussian distribution. **MDF++ (frozen-ViT)** and **MDF++ (frozen-Sparsh)** use frozen pretrained encoders for visual and tactile observations. The pretrained ViT [59] provides a semantic representation, while Sparsh [45] is pretrained specifically for vision-based tactile sensing. These baselines test whether fixed pretrained representations are sufficient for visuo-tactile control. **UWM** [9] is a unified world-action model that encodes future observations using a frozen SDXL VAE [20], which is optimized for high-quality pixel reconstruction. This baseline evaluates a reconstruction-oriented latent space for world-action modeling. Finally, **DP w/ tactile** is a diffusion policy baseline that conditions on both visual and tactile observations. This baseline compares MDF++ against a policy-only model without the proposed multimodal world-action training objective.

5.1.1 Results

The success rate of all methods can be found at Table. 1. The success rates are computed over 150 trials for each method.

Structured queries complement selective representation learning. The adapted MDF baseline achieves 68.4% average success, and adding selective end-to-end representation learning under random masking increases this to 71.5%. Combining selective learning with structured policy, world-action, forward-dynamics, and inverse-dynamics queries yields 86.9%, an improvement of 18.5 percentage points over adapted MDF. These results support coupling structured queries with joint representation learning.

Table 1: Simulation tasks and success rates on five simulated visuo-tactile manipulation tasks. The two added MDF variants report per-task maxima across available evaluation seeds and policy/collocation modes, not seed means.



Model	PegInsert ↑	GearMesh ↑	NutThread ↑	BallSort ↑	ObjectSearch ↑	Avg. ↑
MDF++	85.3	78.7	92.0	93.3	85.3	**86.9**
MDF (adapted) [10]	89.3	78.7	92.7	42.0	39.3	68.4
MDF + selective E2E	86.7	54.7	92.0	79.3	44.7	71.5
MDF++ (naive grad)	56.7	50.0	90.7	67.3	56.0	64.1
MDF++(SIGReg [15])	50.7	32.0	84.7	48.0	75.3	58.1
MDF++(frozen-ViT)	86.7	66.0	89.3	36.7	14.7	58.7
MDF++(frozen-Sparsh)	90.0	64.7	81.3	40.0	18.7	58.9
MDF++(policy)	62.7	64.7	90.7	90.0	86.6	78.9
UWM [9] w/ tactile	88.6	72.0	89.3	53.3	68.0	74.2
DP [36] w/ tactile	88.7	81.3	91.3	28.0	54.0	68.7

Selective gradient flow stabilizes end-to-end representation learning. Table 1 shows that MDF++ achieves the best average success rate across the five simulated tasks, reaching 86.9%. MDF++ (naive grad), which backpropagates all denoising losses through all encoders, achieves 64.1%. This large gap supports our hypothesis that naive end-to-end diffusion training can produce shortcut latent representations that are easy to denoise but less useful for control. Adding SIGReg as a latent-space regularizer does not resolve this issue, achieving only 58.1% average success, suggesting that the regularization is insufficient for joint encoder learning of latent diffusion models.

Task-aware representations outperform frozen pretrained encoders. MDF++ substantially outperforms MDF++(frozen-ViT), MDF++(frozen-Sparsh) and UWM, which achieve average success rates of 58.7%, 58.9% and 74.2%, respectively. Although ViT and Sparsh provide strong semantic or tactile representations, their features are not optimized for the action, force, and contact dynamics required by these tasks. UWM performs better than the frozen-encoder ablations on average, but still lags behind MDF++. This suggests that predicting future observations at the pixel level can impose unnecessary reconstruction burden, forcing the model to allocate capacity to visual details that may be irrelevant for control. This limitation is especially evident on the tactile-exploration tasks: MDF++ achieves 93.3% on BallSort and 85.3% on ObjectSearch, while the frozen-encoder variants perform much worse. These results indicate that contact-rich manipulation benefits from representations shaped by task-specific multimodal prediction objectives.

World-action training improves average success over policy-only learning. MDF++ improves average success over MDF++(policy) from 78.9% to 86.9%. Its largest gains are on PegInsert and GearMesh (22.6 and 14.0 percentage points), with smaller gains on NutThread and BallSort (1.3 and 3.3 points); ObjectSearch decreases by 1.3 points. Compared with DP w/ tactile, MDF++ improves average success from 68.7% to 86.9%, with large gains on BallSort and ObjectSearch (65.3 and 31.3 points), although DP performs slightly better on PegInsert and GearMesh. These results suggest that auxiliary world-action objectives provide useful supervision beyond action prediction, while their benefit varies by task and baseline.

5.2 Representation Analysis

We examine feature variance and effective rank for the RGB and tactile encoders, and assess action information using a linear probe trained on frozen RGB features. Lower action-prediction mean squared error (MSE) indicates that actions are more linearly predictable from the representation.

Table 2 separates representation diversity from action predictability. The pretrained ViT has the highest RGB variance and effective rank, but its probe error is approximately $3.8\times$ that of MDF++. Self-denoising-only supervision produces low effective ranks in both modalities and a probe error

Table 2: Representation diagnostics. RGB and tactile features have dimension 192. The action probe uses frozen RGB features. High variance or effective rank alone does not imply low action-prediction error.

	RGB		Tactile		RGB probe
Representation	Variance	Eff. rank	Variance	Eff. rank	MSE ↓
Pretrained ViT	10.09	49.10	8.26	20.28	0.163
Self-denoising only	2.15	3.67	0.05	2.22	0.574
Naive E2E	7.51	23.51	2.90	5.99	0.520
MDF++	9.62	29.19	7.75	33.34	**0.043**

of 0.574, consistent with representation collapse. Naive end-to-end training retains more diversity, but its tactile effective rank remains low (5.99 versus 33.34 for MDF++), and its probe error is 0.520 versus 0.043. These diagnostics support the hypothesis that selective cross-modal supervision reduces representation collapse while making action information more linearly accessible.

5.3 Effect of Self-denoising Gradients

We scale each encoder's self-denoising gradients by λ_{self}, retaining cross-modal gradients and backbone denoising losses. Average simulation success rates at $\lambda_{\text{self}} = 0, 0.1, 0.5, 1$ are 86.9%, 80.0%, 74.9%, 64.1%, respectively. Zero corresponds to MDF++ and one to naive end-to-end training. Intermediate results use per-task maxima across available seeds and policy/collocation modes. The trend supports suppressing self-denoising encoder gradients, consistent with the representation diagnostics; differences in training recipes and evaluation selection limit a strictly controlled interpretation.

5.4 Physical Experiments

Hardware Platforms. We run our experiments on a Franka Research 3 robot with WSG-50 gripper. We use gelsight mini for tactile sensing and only feed the image of right finger to the policy. A Realsense 405 camera is mounted on the wrist to provide visual observation.

Tasks We evaluate our method on two challenging contact-rich manipulation tasks: Bolt Screwing and Nut Mating, on a real car engine. Both tasks require precise alignment under partial observability, where small pose errors can lead to failed insertion or jamming. In Bolt Screwing, the robot must align a bolt with a threaded hole and generate the appropriate downward force and rotational motion to initiate and maintain thread engagement. In Nut Mating, the robot must align a nut with a fixed bolt and guide it into the correct threading configuration. These tasks are particularly difficult because successful execution depends on millimeter-level accuracy; even slight misalignment can prevent the threads from engaging or cause cross-threading.

Data Collection and Processing Collecting high-quality demonstrations for Bolt Screwing and Nut Mating through direct teleoperation is challenging, as both tasks require precise alignment and axis-constrained rotation [60]. We therefore collect demonstrations using a customized version of UMI [61] equipped with GelSight sensors. In total, we collect 255 demonstrations for Nut Mating and 300 demonstrations for Bolt Screwing.

The original UMI formulation uses future end-effector states as actions to compensate for control delay. However, we find that this is insufficient for our tasks, since it does not account for controller tracking error. This issue is especially problematic for contact-rich manipulation, where millimeter-level accuracy is required, but blindly increasing controller gains may lead to excessive contact forces and potential hardware damage. To address this, we replay the collected trajectories in the robot frame and update the commanded actions using a residual correction proportional to the tracking error. Empirically, this reduces the average tracking error from approximately 4 mm to 1.2 mm, which is critical for successful execution. Please see appendix for details.

Evaluation protocol. At initialization, the nut or bolt is randomly placed between the robot fingers at different regions of the GelSight sensor, and each policy is evaluated under diverse initial end-effector poses. We evaluate three policies on two real-world threading tasks with 20 trials per policy-task pair. Each rollout is scored using a stage-based protocol: successful alignment receives 60 points, successful 180° rotation receives 75 points, and full credit is awarded only when the gripper releases and the nut or bolt remains fixed. The complete scoring rubric is provided in the appendix.

We additionally report success rates, defining success as releasing the manipulated object while it remains fixed to its threaded counterpart.

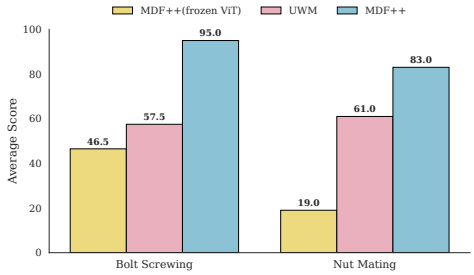


Figure 3: Real-world scores. MDF++ outperforms baselines on both threading tasks.

Result Analysis Figure 3 reports the stage-based scores for both threading tasks. MDF++ achieves success rates of 80% (16/20) on Bolt Screwing and 40% (8/20) on Nut Mating, compared with 50% (10/20) and 25% (5/20) for UWM, and 30% (6/20) and 15% (3/20) for MDF++(frozen-ViT), respectively. In our evaluations, MDF++ adapts to varied initial end-effector poses and object placements on the GelSight sensor, and recovers from small misalignments by making contact, refining alignment, and continuing rotation. UWM often fails at phase transitions, stopping before reaching the object or rotating before thread engagement. MDF++(frozen-ViT) frequently stops before contact or after coarse alignment. These observations suggest that task-aware representations help ground actions in the contact and alignment state.

6 Conclusion

We presented MDF++, a visuo-tactile world-action model that learns task-aware representations end-to-end with cross-modality prediction. By using multi-task world-action objectives together with selective gradient flow, MDF++ avoids pixel-level reconstruction and predicts future in a representation space that is helpful for manipulation. Experiments on simulated and real-world contact-rich manipulation tasks show that MDF++ outperforms frozen pretrained encoders, naive end-to-end finetuning, and policy-only training.

7 Limitations and future works

While MDF++ shows promising results for contact-rich manipulation, several limitations remain. First, MDF++ processes raw force signals at a relatively low frequency, which may miss high-frequency contact dynamics. Future work could learn force encoders over streaming high-frequency force measurements. Second, MDF++ requires manually tuning the loss weights for different modality. Automatically balancing these objectives would reduce task-specific tuning. Finally, although selective gradient flow helps prevent feature collapse, it also blocks gradients from each modality's own denoising loss, potentially limiting the diffusibility of the learned latent space. Future work could explore alternative latent-training strategies, such as Unified Latents [62], to improve diffusibility while preserving control-relevant features.

Acknowledgments

We thank Fan Yang and Luis Marques for helpful discussions of the ideas in this work, and Nima Fazeli for lending us a 3D printer and a WSG-50 gripper. This work was supported in part by the Office of Naval Research under grants N00014-24-1-2036 and N6833525C0329 and NSF grants IIS-2113401 and IIS-2220876.

References

[1] C.-L. Cheang, G. Chen, Y. Jing, T. Kong, H. Li, Y. Li, Y. Liu, H. Wu, J. Xu, Y. Yang, et al. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024.

[2] S. Li, Y. Gao, D. Sadigh, and S. Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025.

[3] J. Won, K. Lee, H. Jang, D. Kim, and J. Shin. Dual-stream diffusion for world-model augmented vision-language-action model. *arXiv preprint arXiv:2510.27607*, 2025.

[4] H. Wu, Y. Jing, C. Cheang, G. Chen, J. Xu, X. Li, M. Liu, H. Li, and T. Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *International Conference on Learning Representations*, volume 2024, pages 10641–10662, 2024.

[5] Y. Hu, Y. Guo, P. Wang, X. Chen, Y.-J. Wang, J. Zhang, K. Sreenath, C. Lu, and J. Chen. Video prediction policy: A generalist robot policy with predictive visual representations. *arXiv preprint arXiv:2412.14803*, 2024.

[6] M. J. Kim, Y. Gao, T.-Y. Lin, Y.-C. Lin, Y. Ge, G. Lam, P. Liang, S. Song, M.-Y. Liu, C. Finn, et al. Cosmos policy: Fine-tuning video models for visuomotor control and planning. *arXiv preprint arXiv:2601.16163*, 2026.

[7] J. Liang, P. Tokmakov, R. Liu, S. Sudhakar, P. Shah, R. Ambrus, and C. Vondrick. Video generators are robot policies. *arXiv preprint arXiv:2508.00795*, 2025.

[8] J. Pai, L. Achenbach, V. Montesinos, B. Forrai, O. Mees, and E. Nava. mimic-video: Video-action models for generalizable robot control beyond vlas. *arXiv preprint arXiv:2512.15692*, 2025.

[9] C. Zhu, R. Yu, S. Feng, B. Burchfiel, P. Shah, and A. Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv preprint arXiv:2504.02792*, 2025.

[10] Z. Huang, H. Hou, and D. Berenson. Unified multimodal diffusion forcing for forceful manipulation. *arXiv preprint arXiv:2511.04812*, 2025.

[11] G. Team, A. Ye, B. Wang, C. Ni, G. Huang, G. Zhao, H. Li, J. Li, J. Zhu, L. Feng, et al. Gigabrain-0: A world model-powered vision-language-action model. *arXiv preprint arXiv:2510.19430*, 2025.

[12] Y. Wang, R. Syed, F. Wu, M. Zhang, A. Onol, J. Barreiros, H. Nayyeri, T. Dear, H. Zhang, and Y. Li. Interactive world simulator for robot policy training and evaluation. *arXiv preprint arXiv:2603.08546*, 2026.

[13] Y. Du, S. Yang, P. Florence, F. Xia, A. Wahid, P. Sermanet, T. Yu, P. Abbeel, J. B. Tenenbaum, L. Kaelbling, et al. Video language planning. In *International Conference on Learning Representations*, volume 2024, pages 31138–31155, 2024.

[14] G. Zhou, H. Pan, Y. LeCun, and L. Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024.

[15] L. Maes, Q. L. Lidec, D. Scieur, Y. LeCun, and R. Balestriero. Leworldmodel: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

[16] 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.

[17] G. R. Team, K. Choromanski, C. Devin, Y. Du, D. Dwibedi, R. Gao, A. Jindal, T. Kipf, S. Kirmani, I. Leal, et al. Evaluating gemini robotics policies in a veo world simulator. *arXiv preprint arXiv:2512.10675*, 2025.

[18] K. Zhang, S. Sha, H. Jiang, M. Loper, H. Song, G. Cai, Z. Xu, X. Hu, C. Zheng, and Y. Li. Real-to-sim robot policy evaluation with gaussian splatting simulation of soft-body interactions. *arXiv preprint arXiv:2511.04665*, 2025.

[19] Y. Zheng, S. Gu, W. Li, Y. Zheng, Y. Zang, S. Tian, X. Li, C. Hao, C. Gao, S. Liu, et al. Omnivta: Visuo-tactile world modeling for contact-rich robotic manipulation. *arXiv preprint arXiv:2603.19201*, 2026.

[20] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pages 1862–1874, 2024.

[21] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[22] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15619–15629, 2023.

[23] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[24] S. Luo and W. Hu. Diffusion probabilistic models for 3d point cloud generation. In *CVPR*, pages 2837–2845, 2021.

[25] R. Zheng, J. Wang, S. Reed, J. Bjorck, Y. Fang, F. Hu, J. Jang, K. Kundalia, Z. Lin, L. Magne, et al. Flare: Robot learning with implicit world modeling. *arXiv preprint arXiv:2505.15659*, 2025.

[26] M. Du, O. Y. Lee, S. Nair, and C. Finn. Play it by ear: Learning skills amidst occlusion through audio-visual imitation learning. *arXiv preprint arXiv:2205.14850*, 2022.

[27] H. Li, Y. Zhang, J. Zhu, S. Wang, M. A. Lee, H. Xu, E. Adelson, L. Fei-Fei, R. Gao, and J. Wu. See, hear, and feel: Smart sensory fusion for robotic manipulation. *CoRL*, 2022.

[28] Z. Liu, C. Chi, E. Cousineau, N. Kuppuswamy, B. Burchfiel, and S. Song. Maniwav: Learning robot manipulation from in-the-wild audio-visual data. In *8th Annual Conference on Robot Learning*, 2024.

[29] M. Noseworthy, B. Tang, B. Wen, A. Handa, C. Kessens, N. Roy, D. Fox, F. Ramos, Y. Narang, and I. Akinola. Forge: Force-guided exploration for robust contact-rich manipulation under uncertainty. *IEEE Robotics and Automation Letters*, 2025.

[30] J. H. Kang, S. Joshi, R. Huang, and S. K. Gupta. Robotic compliant object prying using diffusion policy guided by vision and force observations. *IEEE Robotics and Automation Letters*, 2025.

[31] Y. Wu, Z. Chen, F. Wu, L. Chen, L. Zhang, Z. Bing, A. Swikir, S. Haddadin, and A. Knoll. Tacdiffusion: Force-domain diffusion policy for precise tactile manipulation. *arXiv preprint arXiv:2409.11047*, 2024.

[32] B. Huang, Y. Wang, X. Yang, Y. Luo, and Y. Li. 3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing. *arXiv preprint arXiv:2410.24091*, 2024.

[33] A. George, S. Gano, P. Katragadda, and A. B. Farimani. Vital pretraining: Visuo-tactile pretraining for tactile and non-tactile manipulation policies. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 258–264. IEEE, 2025.

[34] F. Liu, C. Li, Y. Qin, J. Xu, P. Abbeel, and R. Chen. Vitamin: Learning contact-rich tasks through robot-free visuo-tactile manipulation interface. *arXiv preprint arXiv:2504.06156*, 2025.

[35] Z. Xu, R. Uppuluri, X. Zhang, C. Fitch, P. G. Crandall, W. Shou, D. Wang, and Y. She. Unit: Data efficient tactile representation with generalization to unseen objects. *IEEE Robotics and Automation Letters*, 2025.

[36] X. Zhu, B. Huang, and Y. Li. Touch in the wild: Learning fine-grained manipulation with a portable visuo-tactile gripper. *Advances in Neural Information Processing Systems*, 38: 153783–153812, 2026.

[37] H. Van Hoof, N. Chen, M. Karl, P. Van Der Smagt, and J. Peters. Stable reinforcement learning with autoencoders for tactile and visual data. In *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 3928–3934. IEEE, 2016.

[38] S. Suresh, H. Qi, T. Wu, T. Fan, L. Pineda, M. Lambeta, J. Malik, M. Kalakrishnan, R. Calandra, M. Kaess, et al. Neuralfeels with neural fields: Visuotactile perception for in-hand manipulation. *Science Robotics*, 9(96):eadl0628, 2024.

[39] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *2019 International conference on robotics and automation (ICRA)*, pages 8943–8950. IEEE, 2019.

[40] Y. Chen, A. Sipos, M. Van der Merwe, and N. Fazeli. Visuo-tactile transformers for manipulation. *CoRL*, 2022.

[41] J. Hansen, F. Hogan, D. Rivkin, D. Meger, M. Jenkin, and G. Dudek. Visuotactile-rl: Learning multimodal manipulation policies with deep reinforcement learning. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8298–8304. IEEE, 2022.

[42] Y. Yuan, H. Che, Y. Qin, B. Huang, Z.-H. Yin, K.-W. Lee, Y. Wu, S.-C. Lim, and X. Wang. Robot synesthesia: In-hand manipulation with visuotactile sensing. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6558–6565. IEEE, 2024.

[43] C. Sferrazza, Y. Seo, H. Liu, Y. Lee, and P. Abbeel. The power of the senses: Generalizable manipulation from vision and touch through masked multimodal learning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9698–9705. IEEE, 2024.

[44] G. Cao, J. Jiang, D. Bollegala, and S. Luo. Learn from incomplete tactile data: Tactile representation learning with masked autoencoders. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10800–10805. IEEE, 2023.

[45] C. Higuera, A. Sharma, C. K. Bodduluri, T. Fan, P. Lancaster, M. Kalakrishnan, M. Kaess, B. Boots, M. Lambeta, T. Wu, et al. Sparsh: Self-supervised touch representations for vision-based tactile sensing. *arXiv preprint arXiv:2410.24090*, 2024.

[46] J. Zhao, Y. Ma, L. Wang, and E. H. Adelson. Transferable tactile transformers for representation learning across diverse sensors and tasks. *arXiv preprint arXiv:2406.13640*, 2024.

[47] V. Dave, F. Lygerakis, and E. Rueckert. Multimodal visual-tactile representation learning through self-supervised contrastive pre-training. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8013–8020. IEEE, 2024.

[48] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[49] B. Chen, D. Martí Monsó, Y. Du, M. Simchowitz, R. Tedrake, and V. Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

[50] Z. J. Cui, H. Pan, A. Iyer, S. Haldar, and L. Pinto. Dynamo: In-domain dynamics pretraining for visuo-motor control. *Advances in Neural Information Processing Systems*, 37:33933–33961, 2024.

[51] D. Brandfonbrener, O. Nachum, and J. Bruna. Inverse dynamics pretraining learns good representations for multitask imitation. *Advances in Neural Information Processing Systems*, 36: 66953–66978, 2023.

[52] Y. Guo, Y. Hu, J. Zhang, Y.-J. Wang, X. Chen, C. Lu, and J. Chen. Prediction with action: Visual policy learning via joint denoising process. *Advances in Neural Information Processing Systems*, 37:112386–112410, 2024.

[53] J. Cen, C. Yu, H. Yuan, Y. Jiang, S. Huang, J. Guo, X. Li, Y. Song, H. Luo, F. Wang, et al. Worldvla: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025.

[54] X. Leng, J. Singh, Y. Hou, Z. Xing, S. Xie, and L. Zheng. Repa-e: Unlocking vae for end-to-end tuning of latent diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18262–18272, 2025.

[55] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[56] M. Mittal, P. Roth, J. Tigue, A. Richard, O. Zhang, P. Du, A. Serrano-Muñoz, X. Yao, R. Zurbrügg, N. Rudin, L. Wawrzyniak, M. Rakhsha, A. Denzler, E. Heiden, A. Borovicka, O. Ahmed, I. Akinola, A. Anwar, M. T. Carlson, J. Y. Feng, A. Garg, R. Gasoto, L. Gulich, Y. Guo, M. Gussert, A. Hansen, M. Kulkarni, C. Li, W. Liu, V. Makoviychuk, G. Malczyk, H. Mazhar, M. Moghani, A. Murali, M. Noseworthy, A. Poddubny, N. Ratliff, W. Rehberg, C. Schwarke, R. Singh, J. L. Smith, B. Tang, R. Thaker, M. Trepte, K. V. Wyk, F. Yu, A. Millane, V. Ramasamy, R. Steiner, S. Subramanian, C. Volk, C. Chen, N. Jawale, A. V. Kuruttukulam, M. A. Lin, A. Mandlekar, K. Patzwaldt, J. Welsh, H. Zhao, F. Anes, J.-F. Lafleche, N. Moënne-Loccoz, S. Park, R. Stepinski, D. V. Gelder, C. Amevor, J. Carius, J. Chang, A. H. Chen, P. de Heras Ciechomski, G. Daviet, M. Mohajerani, J. von Muralt, V. Reutskyy, M. Sauter, S. Schirm, E. L. Shi, P. Terdiman, K. Vilella, T. Widmer, G. Yeoman, T. Chen, S. Grizan, C. Li, L. Li, C. Smith, R. Wiltz, K. Alexis, Y. Chang, D. Chu, L. J. Fan, F. Farshidian, A. Handa, S. Huang, M. Hutter, Y. Narang, S. Pouya, S. Sheng, Y. Zhu, M. Macklin, A. Moravanszky, P. Reist, Y. Guo, D. Hoeller, and G. State. Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning. *arXiv preprint arXiv:2511.04831*, 2025. URL https://arxiv.org/abs/2511.04831.

[57] Q. K. Luu, P. Zhou, Z. Xu, Z. Zhang, Q. Qiu, and Y. She. Manifeel: Benchmarking and understanding visuotactile manipulation policy learning. *arXiv preprint arXiv:2505.18472*, 2025.

[58] R. Balestriero and Y. LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

[59] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

[60] S. Sha, Y. Wang, B. Huang, A. Loquercio, and Y. Li. Efficient and reliable teleoperation through real-to-sim-to-real shared autonomy. *arXiv preprint arXiv:2603.17016*, 2026.

[61] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*, 2024.

[62] J. Heek, E. Hoogeboom, T. Mensink, and T. Salimans. Unified latents (ul): How to train your latents. *arXiv preprint arXiv:2602.17270*, 2026.