COORDINATING IMAGINATION AND ACTION: A COUPLED NOISE PRIOR FOR WORLD ACTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

World Action Models (WAMs) combine visual imagination with action generation for robot control, yet imagined futures may not always align with the outcomes produced by generated actions. In our empirical analysis, failed rollouts exhibit larger discrepancies between imagined futures and actual execution than successful ones, highlighting the importance of imagination–execution consistency. Prior work has shown that stochastic initialization enables diffusion- and flow-based generative policies to express diverse plausible behaviors under the same conditioning, suggesting that source noise implicitly specifies choices among behavioral modes. In WAMs, such stochastic choices exist in both video imagination and action generation, but their source noises are commonly initialized independently. This leaves the two choices uncoordinated at initialization and requires subsequent cross-modal interaction to establish correspondence, potentially resulting in individually plausible yet incompatible imagination and action. To address this issue, we introduce **A**ction-**L**ed **I**magination via **G**aussian **N**oise Coupling (ALIGN), an action-anchored coupled Gaussian prior that coordinates imagination and action from the source of stochastic generation. ALIGN learns *where to couple* through a task-relevant stochastic subspace and *how strongly to couple* through a learnable coupling strength, providing a shared stochastic reference while preserving the original Gaussian marginals of both modalities. Across two representative WAM paradigms, ALIGN yields absolute success-rate gains of 3.4% and 1.4% for Fast-WAM-IDM and OpenWAM-Joint on LIBERO-Plus, respectively, and 2.7% and 1.1% on RoboCasa-GR1. Further analyses show improved imagination–execution consistency while maintaining behavioral diversity, supporting stochastic coordination between imagination and action as an effective design principle for WAMs.

1 INTRODUCTION

General-purpose embodied agents are expected to understand how the physical world evolves through interaction while mapping visual observations and language instructions to actions (Black et al., 2024; Intelligence et al., 2025). This has motivated growing interest in World Action Models (WAMs), which integrate future prediction with action generation (Zhu et al., 2025; Ye et al., 2026b; Yuan et al., 2026). Compared with Vision-Language-Action (VLA) policies that primarily predict actions, WAMs explicitly model future states to guide robot control (Hu et al., 2024; Wen et al., 2024).



Figure 1: Imagination–execution consistency is sensitive to stochastic initialization. With the observation and instruction fixed, different video–action noise pairings yield a broad range of errors.

Bringing future prediction into robot control makes imagination–execution consistency particularly important (Ruan et al., 2026; Qiu et al., 2026). Meanwhile, our analysis reveals that such consistency is influenced by stochastic initialization. As shown in Fig. 1, we measure imagination–execution discrepancy by comparing predicted future visual states with observations obtained after executing generated actions. Under identical conditioning, different video–action noise pairings produce different discrepancies, suggesting that stochastic choices at initialization affect the agreement between imagined futures and executed behaviors.

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

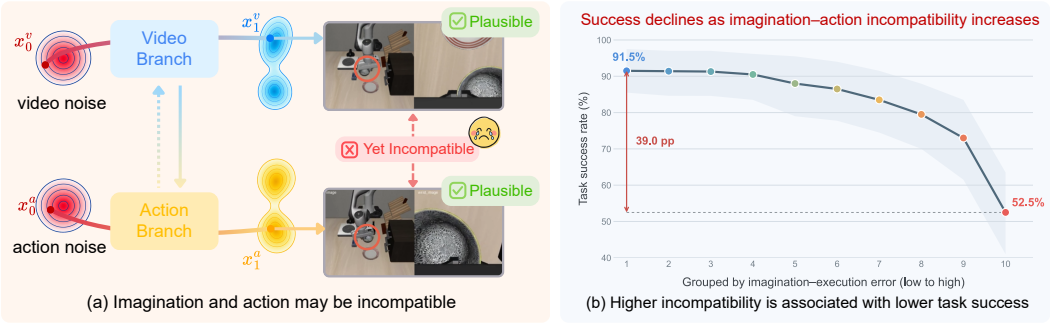


Figure 2: **Imagination–action incompatibility in World Action Models.** (a) Video and action branches can generate individually plausible outcomes from independently sampled noise, yet the imagined future and generated action may reflect incompatible behavioral choices. (b) Higher imagination–action incompatibility is associated with lower task success, highlighting the importance of coordinating stochastic choices between imagination and action.

This observation can be understood from the stochastic nature of generative policies. Prior studies have shown that different stochastic initializations can induce diverse plausible behavioral realizations under identical conditioning in diffusion- and flow-based policies (Chi et al., 2025; Black et al., 2024; Liu et al., 2022). Source noise can therefore be viewed as implicitly specifying a stochastic choice among multiple plausible behavioral modes. In WAMs, such stochastic choices arise in both video imagination and action generation (Zhu et al., 2025; Ye et al., 2026b). However, their source noises are commonly initialized independently, providing no shared stochastic reference for coordinating choices at initialization. The imagined future and generated action may therefore correspond to different behavioral modes despite being individually plausible, as illustrated in Fig. 2(a).

This motivates us to revisit the factorized Gaussian source priors commonly adopted in existing WAMs (Zhu et al., 2025; Ye et al., 2026b; Yuan et al., 2026; Wang et al., 2026), where video imagination and action generation are initialized from independently sampled Gaussian noise. Although subsequent cross-modal interaction or conditioning can facilitate information exchange between the two modalities, the source prior itself does not encode any shared stochastic reference for coordinating their initial choices. As shown in Fig. 2(b), higher imagination–action incompatibility is associated with lower task success, further motivating the coordination of stochastic choices at initialization. This naturally raises the question: *since source noise implicitly selects among plausible behavioral modes, can we coordinate these stochastic choices at initialization in WAMs to encourage imagination and action to follow the same behavioral mode?*

To address this question, we propose **A**ction-**L**ed **I**magination via **G**aussian **N**oise Coupling (ALIGN), an action-anchored coupled Gaussian prior that coordinates imagination and action from the source of stochastic generation. Since actions ultimately determine how the robot interacts with the environment, ALIGN uses the action source as the stochastic anchor and couples it with a learned subspace of the video source. In this way, the model learns *where to couple* through the stochastic subspace and *how strongly to couple* through a learnable coupling strength. The resulting joint prior establishes a shared stochastic reference while preserving the original Gaussian marginal of each modality. An information regularizer further calibrates the amount of shared stochastic information, balancing cross-modal coordination with modality-specific variation. We evaluate ALIGN across two representative WAM paradigms: Fast-WAM-IDM (Yuan et al., 2026), which generates actions conditioned on imagined futures, and OpenWAM-Joint (Wang et al., 2026), which jointly models future video and actions. On LIBERO-Plus (Fei et al., 2025), ALIGN yields absolute success-rate gains of 3.4% and 1.4% for Fast-WAM-IDM and OpenWAM-Joint, respectively; on RoboCasa-GR1 (Bjorck et al., 2025), the corresponding gains are 2.7% and 1.1%. Further analyses show improved imagination–execution consistency while maintaining behavioral diversity. Before delving into the details, we summarize our contributions as follows:

- We revisit stochastic initialization in generative WAMs and reveal that independent noise initialization overlooks the need for stochastic coordination between imagined futures and generated actions. Controlled noise-pairing experiments and rollout analyses further establish the connection between stochastic initialization, imagination–execution consistency, and task performance.

- To address this limitation, we propose ALIGN, an action-anchored Gaussian noise coupling framework that explicitly coordinates stochastic choices between video imagination and action generation. ALIGN learns both where and how strongly to couple while preserving the Gaussian marginal distribution of each modality.
- We validate ALIGN across representative World Action Model paradigms, including both inverse-dynamics-based and joint-generation settings, on LIBERO-Plus, RoboCasa-GR1, and real-world manipulation tasks. ALIGN consistently improves closed-loop success while enhancing imagination–execution agreement without compromising behavioral diversity.

2 Related Work

2.1 Modeling Multimodal Action Distributions

Robot demonstrations can exhibit multiple valid behaviors for the same observation and task, motivating policies that capture multimodal action distributions. Prior approaches use energy-based models, discrete behavior modes, and trajectory diffusion to represent this variability (Florence et al., 2022; Shafiullah et al., 2022; Janner et al., 2022). Diffusion policies learn expressive conditional action distributions through iterative denoising (Chi et al., 2025; Ze et al., 2024; Liu et al., 2025; Team et al., 2024), and diffusion or flow-based action heads also underpin recent vision-language-action models (Black et al., 2024; Intelligence et al., 2025; Bjorck et al., 2025). Through the learned generative mapping, different random initializations can yield distinct behavioral realizations. Our work studies how to coordinate this action-side stochasticity with visual imagination through a shared stochastic reference.

2.2 World Action Models

World Action Models (WAMs) incorporate future world prediction into robot control by leveraging imagined dynamics for action generation. Existing approaches introduce future information through predicted visual observations (Du et al., 2023; Bharadhwaj et al., 2024; Zhang et al., 2026; Liang et al., 2025; Chen et al., 2025; Feng et al., 2025), predictive representations with inverse dynamics (Hu et al., 2024; Pai et al., 2025; Tian et al., 2025), and joint video–action generation (Zhu et al., 2025; Shen et al., 2026; Bi et al., 2026; Wang et al., 2026; Yan et al., 2026; Team et al., 2026; Chen et al., 2026). Recent studies further explore generalization, causal interaction, and long-horizon planning in WAMs (Ye et al., 2026b; Li et al., 2026; Kim et al., 2026; Ye et al., 2026a; Zhou et al., 2026). Fast-WAM separates the effects of video co-training and explicit test-time imagination. While its default policy does not generate future states during inference, its IDM variant first predicts future video representations and then derives actions from the imagined future (Yuan et al., 2026). We use this variant, denoted as Fast-WAM-IDM, throughout our experiments.

Alongside these developments, recent work examines the gap between visual prediction and reliable execution. Ruan et al. (2026) show that action-state consistency distinguishes successful and failed rollouts and use predictive consensus for candidate selection. AGRA identifies a mismatch between video representations and action decoding, motivating action-grounded representation alignment (Qiu et al., 2026). These findings highlight the importance of aligning predicted dynamics with generated actions. Our work complements these approaches by coupling video and action initialization to provide a shared stochastic reference for consistent imagination and execution.

2.3 Structured Noise Priors

PYoCo and VideoFusion introduce temporal correlation or shared components into video diffusion noise (Ge et al., 2023; Luo et al., 2023). Initialization refinement and noise warping further improve temporal coherence and motion control (Wu et al., 2024; Burgert et al., 2025; Chang et al., 2024). Multimodal diffusion coordinates outputs through joint modeling and cross-modal representations (Bao et al., 2023; Tang et al., 2023), while flow-based generation studies transports and noise–data couplings (Lipman et al., 2022; Liu et al., 2022; Tong et al., 2023). Our method instead learns where and how strongly to couple video noise to action noise across unequal-dimensional spaces, while preserving each modality's full standard Gaussian marginal.

3 PROBLEM FORMULATION

3.1 REPRESENTATIVE WORLD ACTION MODEL PARADIGMS

Let o_t denote the current observation, l denote the task instruction, and $a_{1:H}$ denote an action chunk of horizon H, where $a_h \in \mathbb{R}^{d_{\text{act}}}$ represents the action at relative step h. A conventional visuomotor policy predicts actions directly from the current context:

$$p(a_{1:H} \mid o_t, l). \tag{1}$$

In contrast, World Action Models (WAMs) explicitly introduce imagined future states into action generation. Let $v_{1:T}$ denote future visual states over a prediction horizon T, where video and action indices are defined relative to the current observation. Rather than relying solely on current context, WAMs leverage predicted future states to inform robot control. Existing WAMs mainly follow two representative paradigms, distinguished by how future imagination interacts with action generation.

Joint Future Video–Action Prediction. Joint-modeling WAMs generate future visual states and actions within a unified generative process. They model:

$$p(v_{1:T}, a_{1:H} \mid o_t, l), \tag{2}$$

where future visual and action variables are generated jointly and can exchange information throughout the generation process.

Future Video Prediction Followed by Action Prediction. Another class of WAMs first imagines future visual observations and then predicts actions conditioned on the imagined future:

$$p(v_{1:T}, a_{1:H} \mid o_t, l) = p(v_{1:T} \mid o_t, l)p(a_{1:H} \mid v_{1:T}, o_t, l). \tag{3}$$

This factorization emphasizes the generation order. In practice, the predicted future representation is provided to an inverse dynamics model or action prediction module for action generation.

Although these two paradigms differ in how imagination and action interact, both involve stochastic generation of future visual states and actions when their corresponding modules are generative. This makes the coordination between their stochastic choices a shared problem across WAM paradigms.

Policies that directly generate actions from current observations without explicitly modeling future visual states do not contain an explicit imagination process and therefore fall outside the imagination–action stochastic coordination problem studied here.

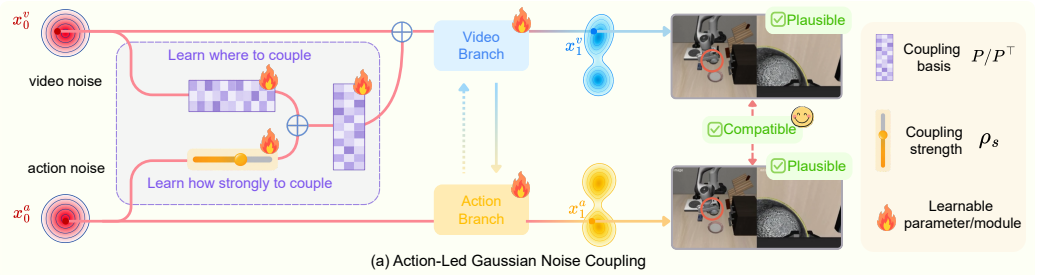
3.2 REVISITING INDEPENDENT NOISE INITIALIZATION IN WORLD ACTION MODELS

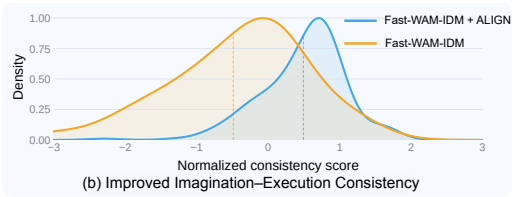
Diffusion- and flow-based policies map random source variables to trajectories, allowing different plausible behaviors under the same observation and instruction. Generative WAMs use this stochasticity for both future imagination and action generation. A noise draw has no predefined behavioral meaning, but different draws can lead to different outcomes through the learned generator.

For flow-based WAMs, let $x_0^v \in \mathbb{R}^{d_v}$ and $x_0^a \in \mathbb{R}^{d_a}$ denote the video and action source variables, where $d_a = Hd_{\text{act}}$ for an action chunk of horizon H. A common choice is to sample them independently:

$$x_0^v \sim \mathcal{N}(0, I_{d_v}), \qquad x_0^a \sim \mathcal{N}(0, I_{d_a}). \tag{4}$$

Thus $p(x_0^v, x_0^a) = p(x_0^v)p(x_0^a)$ and $I(x_0^v; x_0^a) = 0$: each source retains its variation, but the prior specifies no relationship between their random choices. Although joint denoising, cross-modal attention, or video-conditioned action prediction can facilitate information exchange between modalities during generation, they must establish such correspondence without guidance from the source prior. Source noise can be viewed as implicitly specifying a stochastic choice among multiple plausible behavioral modes (Chi et al., 2025; Black et al., 2024; Liu et al., 2022). Therefore, when video imagination and action generation are initialized from independent sources, their stochastic choices may become uncoordinated. This motivates a joint source prior that introduces controlled cross-modal dependence at initialization while preserving the standard Gaussian marginal distribution of each modality, as developed in Sec. 4.





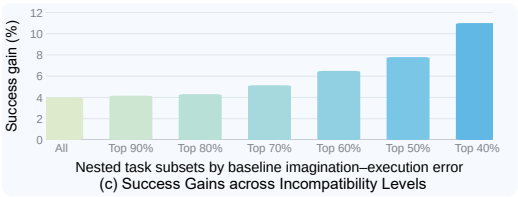


Figure 3: Action-led video–action noise coupling and its observed effects. (a) Action noise anchors video-noise sampling: the learned basis P selects the video subspace to couple, while ρ_s sets the amount of shared randomness. The remaining video component stays independent, and both source marginals remain standard Gaussian. (b) ALIGN shifts the distribution toward higher imagination–execution consistency. (c) In an exploratory 5k-step LIBERO comparison, task subsets with greater baseline mismatch show larger success-rate gains from ALIGN.

4 METHOD

To introduce the controlled dependence sought in Sec. 3.2, we propose **A**ction-**L**ed **I**magination via **G**aussian **N**oise Coupling (ALIGN), a coupled source prior that links action and video randomness while preserving each standard Gaussian marginal (Fig. 3(a)).

4.1 ALIGN: ACTION-LED IMAGINATION VIA GAUSSIAN NOISE COUPLING

ALIGN makes two design choices: *where to couple* and *how strongly to couple*. A learned video subspace specifies where action-side randomness enters; a learned strength controls the resulting cross-modal dependence.

ALIGN constructs a coupled Gaussian prior by modifying the joint distribution of the source states of video and action flows. We first independently sample

$$x_0^a \sim \mathcal{N}(0, I_{d_a}), \qquad x_0^v \sim \mathcal{N}(0, I_{d_v}), \tag{5}$$

where x_0^a and x_0^v denote the action and video source states, respectively.

Learn where to couple. To identify which stochastic directions in the video source should participate in cross-modal coordination, we learn a coupling basis

$$P = \mathrm{qf}(W) \in \mathbb{R}^{d_v \times k}, \qquad P^\top P = I_k. \tag{6}$$

Here $W \in \mathbb{R}^{d_v \times k}$ is a learnable full-column-rank matrix parameterizing the coupling subspace, and $\mathrm{qf}(\cdot)$ returns the orthonormal factor of its thin QR decomposition. The orthonormal constraint provides a valid subspace projection and preserves the standard Gaussian statistics of the projected component. We set $k = d_a \leq d_v$ to match the action source dimension. The corresponding stochastic component of the video source is

$$b = P^\top x_0^v. \tag{7}$$

Rather than coupling the entire video source space, ALIGN restricts cross-modal stochastic dependence to this learned subspace, leaving the orthogonal video component unaffected.

Algorithm 1 Optimization of ALIGN

Require: Paired data $\mathcal{D}$; parameters θ, W, ρ; weight λ
1: Initialize θ, full-column-rank W, and $0 \leq \rho < 1$
2: **for** each training iteration **do**
3: Sample a batch and independent Gaussian sources x_0^a, x_0^v
4: Set $P = \mathrm{qf}(W)$; couple sources by Eq. (9)
5: Update θ, W, ρ with Eq. (12), maintaining $0 \leq \rho < 1$
6: **end for**
Ensure: Trained WAM θ, basis $P = \mathrm{qf}(W)$, strength ρ

Learn how strongly to couple. Given the learned coupling subspace, we introduce a Gaussian-preserving coupling strength ρ to control the amount of shared stochastic information:

$$b' = \rho x_0^a + \sqrt{1-\rho^2}\, b, \tag{8}$$

where $0 \leq \rho < 1$. The notation ρ_s in Fig. 3(a) denotes the learned value of ρ at optimization step s. When $\rho = 0$, the two source variables remain independent; increasing ρ introduces stronger stochastic dependence between the action source and the selected video component.

The coupled source variables are then constructed as

$$\tilde{x}_0^a = x_0^a, \qquad \tilde{x}_0^v = x_0^v + P(b' - b). \tag{9}$$

This construction introduces a shared stochastic reference between imagination and action generation while retaining stochastic variation outside the coupled video subspace.

The amount of shared stochastic information is controlled through the mutual information between the coupled source variables:

$$I(\tilde{x}_0^v; \tilde{x}_0^a) = -\frac{k}{2}\log(1-\rho^2). \tag{10}$$

Motivated by information-bottleneck regularization (Alemi et al., 2016), we use this quantity as an information regularization term $\mathcal{L}_{\mathrm{cap}} = I(\tilde{x}_0^v; \tilde{x}_0^a)$. The paired video–action training objective identifies stochastic directions that are useful for coordinating the two modalities, while $\mathcal{L}_{\mathrm{cap}}$ limits the shared information capacity to encourage retaining only behavior-relevant stochastic dependencies. Importantly, $\mathcal{L}_{\mathrm{cap}}$ depends explicitly only on ρ and therefore provides direct gradients only to the coupling strength. The joint covariance and its determinant are derived in Appendix D.2.

4.2 Gaussian Marginal Preservation

A key requirement when modifying the stochastic initialization of flow-based WAMs is to preserve the source distribution expected by the original generative model. ALIGN changes the dependency structure between video and action source variables while maintaining their Gaussian marginals.

Since the coupling modifies only the projected video component and $P^\top P = I_k$, the covariance of the coupled video source is

$$\begin{aligned} \mathrm{Cov}(\tilde{x}_0^v) &= (I_{d_v} - PP^\top) + (1-\rho^2)PP^\top + \rho^2 PP^\top \\ &= I_{d_v}. \end{aligned} \tag{11}$$

The coupled sources are zero-mean linear transformations of jointly Gaussian variables. Together with the covariance identity above, this gives $\tilde{x}_0^v \sim \mathcal{N}(0, I_{d_v})$ and $\tilde{x}_0^a \sim \mathcal{N}(0, I_{d_a})$. Thus, ALIGN introduces cross-modal stochastic dependency without changing the marginal source distribution of either modality. A detailed proof, including the independence of the coupled and residual video components, is given in Appendix D.1.

4.3 Optimization and Mechanistic Analysis

Training and Inference. ALIGN is optimized together with the original WAM training objective using paired video–action demonstrations. During training, we replace independent source noise

Table 1: LIBERO-Plus success rates (%) across seven perturbations. Avg. pools all episodes; ΔSR is the ALIGN gain.

Method	Camera	Robot	Lang.	Light	Backg.	Noise	Layout	Avg.
π_0	13.8	6.0	58.8	85.0	81.4	79.0	68.9	53.6
$\pi_{0.5}$	78.4	73.6	80.8	96.2	94.1	89.0	84.5	84.4
StarVLA	52.5	49.8	88.5	95.7	95.7	73.0	76.9	74.1
Being-H0.7	82.0	59.0	82.8	97.8	90.0	93.5	88.5	82.1
Joint Future Video-Action Prediction (Joint)								
OpenWAM-Joint	33.8	76.1	88.0	97.0	87.1	39.8	77.5	69.2
+ ALIGN	**35.9**	**77.5**	**89.4**	**97.2**	**89.4**	**41.3**	**78.8**	**70.6**
ΔSR (%)	+2.1%	+1.4%	+1.4%	+0.2%	+2.3%	+1.5%	+1.3%	+1.4%
Future Video Prediction Followed by Action Prediction (IDM)								
Fast-WAM-IDM	44.7	65.3	91.8	90.4	59.6	60.3	78.1	69.5
+ ALIGN	**48.5**	**71.7**	**92.1**	**95.0**	**65.2**	**62.6**	**80.2**	**72.9**
ΔSR (%)	+3.8%	+6.4%	+0.3%	+4.6%	+5.6%	+2.3%	+2.1%	+3.4%

Table 2: RoboCasa-GR1 task success rates (%).

Method	SR (%)
π_0	13.6
$\pi_{0.5}$	37.0
StarVLA	48.8
Being-H0.7	49.2
Joint	
OpenWAM-Joint	60.5
+ ALIGN	**61.6**
ΔSR	+1.1%
IDM	
Fast-WAM-IDM	68.4
+ ALIGN	**71.1**
ΔSR	+2.7%

with the coupled source in Eq. (9) and jointly optimize the WAM and coupling parameters. Algorithm 1 summarizes the optimization procedure; additional details are provided in Appendix C.1. The resulting objective for WAM parameters θ, coupling basis W, and strength ρ is

$$\mathcal{L}(\theta, W, \rho) = \mathcal{L}_{\mathrm{WAM}}(\theta; W, \rho) + \lambda \mathcal{L}_{\mathrm{cap}}(\rho), \qquad \lambda \geq 0. \tag{12}$$

The paired prediction loss updates W through $P = \mathrm{qf}(W)$; the capacity term directly regularizes only ρ. At inference, fresh Gaussian samples are coupled using the learned P and ρ, then passed through the unchanged backbone generation schedule.

Mechanistic Interpretation of ALIGN. Source-level correlation does not directly enforce imagination–execution consistency. However, ALIGN introduces a stochastic communication channel exploited by paired video–action supervision. This shared stochastic channel, together with paired video–action supervision during training, allows the model to associate the latent random factors underlying future imagination and action generation. Through this learned correspondence, ALIGN encourages both modalities to capture compatible behavioral modes, leading to improved imagination–execution consistency. Consequently, generated future states and actions are more likely to reflect consistent behavioral choices. The resulting improvement in imagination–execution agreement is evaluated in Fig. 3(b). The exploratory analysis in Fig. 3(c) further associates greater baseline mismatch with larger control gains; we examine this trend in Sec. 5.3.

5 Experiments

5.1 Experimental Setup

Benchmarks and baselines. We evaluate closed-loop control on LIBERO-Plus (Fei et al., 2025), which extends LIBERO (Liu et al., 2023), and on the RoboCasa-GR1 tabletop benchmark (Bjorck et al., 2025), alongside three real-world manipulation tasks on a Piper robot. LIBERO-Plus tests robustness to camera, robot, language, lighting, background, sensor-noise, and layout perturbations; RoboCasa-GR1 evaluates kitchen manipulation with the GR1 embodiment. We apply ALIGN to Fast-WAM-IDM (Yuan et al., 2026) and OpenWAM-Joint (Wang et al., 2026), representing inverse-dynamics and joint-generation WAMs, respectively. Reference baselines include π_0 (Black et al., 2024), $\pi_{0.5}$ (Intelligence et al., 2025), StarVLA (Community, 2026), and Being-H0.7 (Luo et al., 2026). Evaluation details are given in Appendix B.1. For each backbone, the ALIGN variant follows the corresponding Fast-WAM-IDM or OpenWAM-Joint training and inference settings, with the coupled noise prior and its capacity regularizer added. Models are fine-tuned on standard LIBERO for LIBERO-Plus evaluation and on the corresponding demonstrations for RoboCasa-GR1.

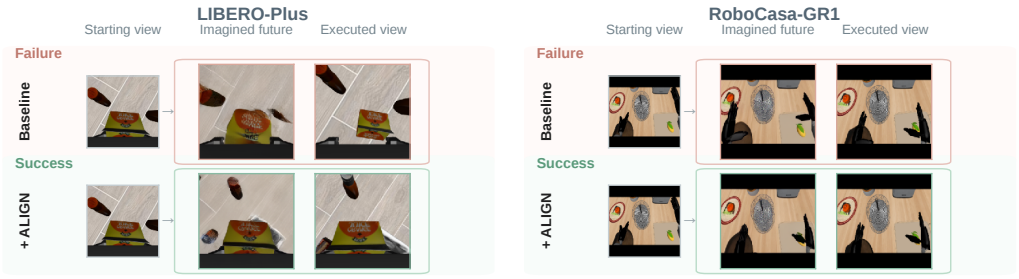


Figure 4: Selected imagination–execution examples on LIBERO-Plus (wrist view) and RoboCasa-GR1 (ego view). Failure and Success denote the outcomes of the full rollouts.

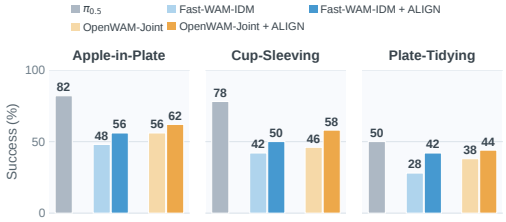


Figure 5: Real-world task success rates over 50 trials per method and task.

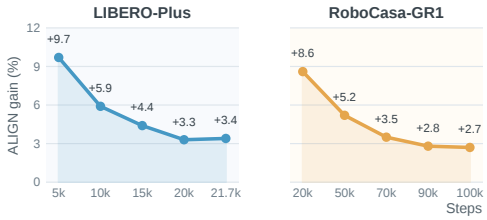


Figure 6: ALIGN improves Fast-WAM-IDM success throughout training, especially early.

5.2 Main Results

5.2.1 Simulation Results

LIBERO-Plus. ALIGN improves both backbones across all seven perturbation categories (Table 1). Overall success rises from 69.5% to 72.9% for Fast-WAM-IDM and from 69.2% to 70.6% for OpenWAM-Joint. Fast-WAM-IDM benefits most from robot (+6.4%), background (+5.6%), and lighting (+4.6%) perturbations. OpenWAM-Joint also improves across categories, with its largest gains under background (+2.3%) and camera (+2.1%) changes. The gains across both inverse-dynamics and joint-generation models support noise coupling as a way to improve control robustness in distinct WAM formulations.

RoboCasa-GR1. ALIGN also improves kitchen manipulation with the GR1 embodiment (Table 2). Fast-WAM-IDM success increases from 68.4% to 71.1%, while OpenWAM-Joint improves from 60.5% to 61.6%. Together with LIBERO-Plus, these results show consistent control improvements across the two evaluated backbones and benchmark settings.

5.2.2 Real-World Results

We conduct real-world experiments on the AgileX PiperX platform, equipped with a front-facing RealSense D455 and two wrist-mounted RealSense D435 cameras (Fig. 9). We evaluate Apple-in-Plate, Cup-Sleeving, and Plate-Tidying using 300 demonstrations per task and 50 evaluation trials per method. Figure 5 compares WAM backbones and ALIGN variants with $\pi_{0.5}$; task images and language instructions appear in Appendix B.3, with trial counts in Table 6. Across the three tasks, ALIGN yields absolute success-rate gains of 8%, 8%, and 14% for Fast-WAM-IDM, and 6%, 12%, and 6% for OpenWAM-Joint. The largest gain is on Plate-Tidying for Fast-WAM-IDM (14/50 to 21/50 successes). $\pi_{0.5}$ has the highest success rate on all three tasks in this comparison.

5.3 Mechanism Analysis

Improving imagination–execution consistency. We compare the imagined future with observations produced by executing the generated actions. In Fig. 3(b), the ALIGN distribution shifts toward higher consistency, indicating that its predicted dynamics more often match the realized rollout. This outcome-level analysis captures the overall agreement between imagined futures and executed behaviors. Figure 4 further illustrates this improvement through selected RoboCasa-GR1 and LIBERO-Plus rollouts. Additional Fast-WAM-IDM comparisons appear in Appendix A.

Table 3: Action diversity across successful trajectories on standard LIBERO.

Method	Spatial	Object	Goal	LIBERO-10	Mean
Fast-WAM-IDM	3.05	4.23	3.32	6.15	4.19
+ ALIGN	2.92	4.13	3.45	6.34	4.21

Table 4: **Learning the coupling strength.** Fast-WAM-IDM success rates (%).

Coupling strength	LIBERO-Plus	RoboCasa-GR1
Fixed ($\rho = 1$)	70.2	69.3
Learned, w/o $\mathcal{L}_{\text{cap}}$	70.5	69.7
Learned, w/ $\mathcal{L}_{\text{cap}}$	**72.9**	**71.1**

Table 5: **Coupling parameterization.** Fast-WAM-IDM success rates (%) on two benchmarks.

Reference	LIBERO-Plus	RoboCasa-GR1
Video-referenced	70.8	70.2
Shared-source coupling	70.7	70.0
Action-referenced (ours)	**72.9**	**71.1**

Larger gains for tasks with greater baseline mismatch. Figure 3(c) shows that task subsets with larger baseline prediction–execution errors achieve greater success-rate improvements with ALIGN. This pattern indicates that ALIGN provides larger benefits for tasks where independently initialized imagination and action are more likely to diverge. Together with the improved consistency in Fig. 3(b), these results suggest that coordinating stochastic choices between imagination and action is especially useful for reducing mismatch in challenging tasks.

Gains across training checkpoints. ALIGN improves Fast-WAM-IDM consistently throughout training, with larger gains observed at earlier checkpoints (Fig. 6). The consistent gains across training stages suggest that coupled initialization provides a useful prior for learning video–action coordination. The advantage remains at the final checkpoint, suggesting that the benefit of coupled initialization extends beyond early-stage training.

Preserving action diversity. ALIGN preserves action diversity while improving coordination. Table 3 shows similar diversity among successful Fast-WAM-IDM trajectories with and without ALIGN across all four LIBERO suites (4.21 versus 4.19 on average). The improved coordination therefore does not compromise the ability to represent multimodal action distributions. Appendix B.2 describes the measure.

5.4 Ablation Studies

Learning the coupling strength. Table 4 separates the benefit of learning coupling strength from constraining its capacity. Learning ρ without $\mathcal{L}_{\text{cap}}$ improves over fixed full coupling on both benchmarks. Adding the capacity term yields further gains, suggesting that useful coordination depends on regulating shared source information rather than maximizing dependence.

Choice of coupling reference. Table 5 compares action-referenced coupling with video-referenced and shared-source variants under the same Fast-WAM-IDM setting. The action-referenced variant performs best on both benchmarks, supporting the action source as the stochastic anchor for visual imagination. This suggests that action-guided coupling provides a more effective coordination signal between imagination and action. Appendix C.2 gives the constructions.

6 Conclusion

We introduced ALIGN, a coupled Gaussian prior that coordinates video imagination and action generation through a shared stochastic reference at initialization. Rather than replacing architectural mechanisms for video–action interaction, ALIGN provides a complementary source-level coordination mechanism that preserves standard Gaussian marginals. Its flexible design integrates with WAM paradigms, including inverse-dynamics-based and joint-generation models. Experiments on LIBERO-Plus and RoboCasa-GR1 demonstrate that ALIGN improves task performance and imagination–execution consistency across WAM paradigms while maintaining behavioral diversity. These results highlight stochastic initialization as a design dimension for world action models and suggest coordinating stochastic choices can complement approaches for video–action interaction.

AI USE STATEMENT

We used generative AI tools to polish the language and presentation of this manuscript. They also assisted with methodological and experimental planning, implementation, and interpretation of results. The authors reviewed all AI-assisted material and take responsibility for the final text, claims, and artifacts.

ETHICS STATEMENT

This work involves no human participants or personal data. Robot experiments were conducted in controlled settings. Policies for physical robots may fail in unfamiliar environments; deployment requires appropriate safeguards and validation.

REPRODUCIBILITY STATEMENT

We provide comprehensive details of the methodology, model configurations, experimental procedures, and evaluation metrics in the paper and appendix. To facilitate reproducibility, we will release the codebase and deployment instructions upon publication.

REFERENCES

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning*, pp. 1692–1717. PMLR, 2023.

Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. *arXiv preprint arXiv:2409.16283*, 2024.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, et al. Motus: A unified latent action world model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35101–35113, 2026.

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

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13–23. IEEE, 2025.

Pascal Chang, Jingwei Tang, Markus Gross, and Vinicius Da Costa De Azevedo. How i warped your noise: a temporally-correlated noise prior for diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 30015–30046, 2024.

Boyuan Chen, Tianyuan Zhang, Haoran Geng, Caiyi Zhang, Peihao Li, Kiwhan Song, William T Freeman, Jitendra Malik, Pieter Abbeel, Russ Tedrake, et al. Large video planner enables generalizable robot control. *arXiv preprint arXiv:2512.15840*, 2025.

Ronghan Chen, Yandan Yang, Zuojin Tang, Dongjie Huo, Tong Lin, Haoning Wu, Haoyun Liu, Yuzhi Chen, Lulu Zheng, Botai Yuan, et al. Abot-m0. 5: Unified mobility-and-manipulation world action model. *arXiv preprint arXiv:2607.00678*, 2026.

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

StarVLA Community. Starvla: A lego-like codebase for vision-language-action model developing. *arXiv preprint arXiv:2604.05014*, 2026.

Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in neural information processing systems*, 36:9156–9172, 2023.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, et al. Libero-plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025.

Yao Feng, Hengkai Tan, Xinyi Mao, Chendong Xiang, Guodong Liu, Shuhe Huang, Hang Su, and Jun Zhu. Vidar: Embodied video diffusion model for generalist manipulation. *arXiv preprint arXiv:2507.12898*, 2025.

Pete Florence, Corey Lynch, Andy Zeng, Oscar A Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. In *Conference on robot learning*, pp. 158–168. PMLR, 2022.

Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22873–22884. IEEE, 2023.

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.

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

Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*, 2022.

Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, et al. Cosmos policy: Fine-tuning video models for visuomotor control and planning. *arXiv preprint arXiv:2601.16163*, 2026.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026.

Junbang Liang, Pavel Tokmakov, Ruoshi Liu, Sruthi Sudhakar, Paarth Shah, Rares Ambrus, and Carl Vondrick. Video generators are robot policies. *arXiv preprint arXiv:2508.00795*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. In *International Conference on Learning Representations*, volume 2025, pp. 29982–30009, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Hao Luo, Wanpeng Zhang, Yicheng Feng, Sipeng Zheng, Haiweng Xu, Chaoyi Xu, Ziheng Xi, Yuhui Fu, and Zongqing Lu. Being-h0. 7: A latent world-action model from egocentric videos. *arXiv preprint arXiv:2605.00078*, 2026.

Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. *arXiv preprint arXiv:2303.08320*, 2023.

Jonas Pai, Liam Achenbach, Victoriano Montesinos, Benedek Forrai, Oier Mees, and Elvis Nava. mimic-video: Video-action models for generalizable robot control beyond vlas. *arXiv preprint arXiv:2512.15692*, 2025.

Lu Qiu, Yizhuo Li, Yi Chen, Yuying Ge, Yixiao Ge, and Xihui Liu. Making foresight actionable: Repurposing representation alignment in world action models. *arXiv preprint arXiv:2606.12217*, 2026.

Bo-Kai Ruan, Teng-Fang Hsiao, Ling Lo, and Hong-Han Shuai. Is the future compatible? diagnosing dynamic consistency in world action models. *arXiv preprint arXiv:2605.07514*, 2026.

Nur Muhammad Shafiullah, Zichen Cui, Ariuntuya Arty Altanzaya, and Lerrel Pinto. Behavior transformers: Cloning k modes with one stone. *Advances in neural information processing systems*, 35:22955–22968, 2022.

Yichao Shen, Fangyun Wei, Zhiying Du, Yaobo Liang, Yan Lu, Jiaolong Yang, Nanning Zheng, and Baining Guo. Videovla: Video generators can be generalizable robot manipulators. *Advances in neural information processing systems*, 38:95597–95621, 2026.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36:16083–16099, 2023.

MotuBrain Team, Chendong Xiang, Fan Bao, Haitian Liu, Hengkai Tan, Hongzhe Bi, James Li, Jiabao Liu, Jingrui Pang, Kiro Jing, et al. Motubrain: An advanced world action model for robot control. *arXiv preprint arXiv:2604.27792*, 2026.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, and Jiangmiao Pang. Predictive inverse dynamics models are scalable learners for robotic manipulation. In *International Conference on Learning Representations*, volume 2025, pp. 92033–92052, 2025.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Yuran Wang, Siqiao Huang, Mingleyang Li, Chenhao Zhang, Jiaqi Liang, Weiyang Jin, Yue Chen, Xuemin Chi, Donghao Zhou, Qize Yu, et al. Openwam: An open, modular exploration towards systematic world-action model pretraining. *arXiv preprint arXiv:2609.07398*, 2026.

Youpeng Wen, Junfan Lin, Yi Zhu, Jianhua Han, Hang Xu, Shen Zhao, and Xiaodan Liang. Vidman: Exploiting implicit dynamics from video diffusion model for effective robot manipulation. *Advances in Neural Information Processing Systems*, 37:41051–41075, 2024.

Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. Freeinit: Bridging initialization gap in video diffusion models. In *European conference on computer vision*, pp. 378–394. Springer, 2024.

Haodong Yan, Zhide Zhong, Jiaguan Zhu, Junjie He, Weilin Yuan, Wenxuan Song, Xin Gong, Yingjie Cai, Guanyi Zhao, Xu Yan, et al. S-vam: Shortcut video-action model by self-distilling geometric and semantic foresight. *arXiv preprint arXiv:2603.16195*, 2026.

Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Hao Li, Hengtao Li, Jie Li, Jindi Lv, Jingyu Liu, et al. Gigaworld-policy: An efficient action-centered world–action model. *arXiv preprint arXiv:2603.17240*, 2026a.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026b.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026.

Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. *arXiv preprint arXiv:2403.03954*, 2024.

Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, He Wang, Zhizheng Zhang, et al. Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge. *Advances in Neural Information Processing Systems*, 38: 24195–24228, 2026.

Pengfei Zhou, Shengcong Chen, Di Chen, Jiaxu Wang, Rongjun Jin, Bingwen Zhu, Yike Pan, Songen Gu, Kuanning Wang, Shufeng Nan, et al. τ_0-wm: A unified video-action world model for robotic manipulation. *arXiv preprint arXiv:2606.01027*, 2026.

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv preprint arXiv:2504.02792*, 2025.

A Additional Visualizations

Figures 7 and 8 compare Fast-WAM-IDM with and without ALIGN on RoboCasa-GR1 and LIBERO-Plus.



Figure 7: Ten RoboCasa-GR1 examples. Each row shows the language instruction and starting view, followed by imagined and executed views for Fast-WAM-IDM and Fast-WAM-IDM + ALIGN.

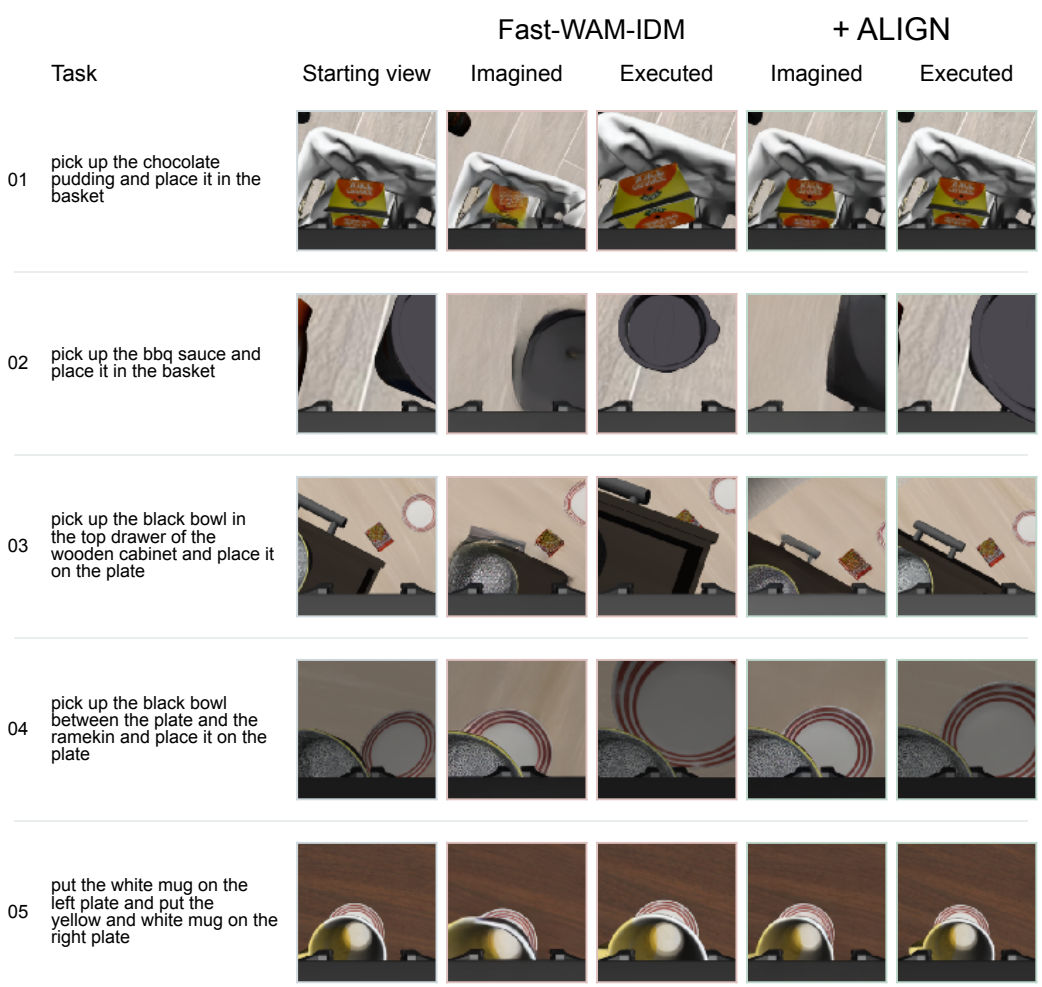


Figure 8: Five LIBERO-Plus examples. Each row shows the task description and starting wrist view, followed by imagined and executed views for Fast-WAM-IDM and Fast-WAM-IDM + ALIGN.

B Experimental Details

B.1 Evaluation Protocol

Apart from ALIGN's coupled noise prior, training and inference follow the respective Fast-WAM-IDM and OpenWAM-Joint configurations. Both backbones are fine-tuned on standard LIBERO for LIBERO-Plus evaluation; RoboCasa-GR1 models use the corresponding task demonstrations.

Within each backbone comparison, the baseline and coupled model use the same observations, action representation, training data, checkpoint budget, and denoising schedule. Evaluation uses matched task instances and initial states, identical episode limits, and the same action execution horizon. LIBERO-Plus Avg. denotes success over all evaluated episodes, rather than an unweighted mean of the seven category scores.

B.2 Action-Trajectory Diversity Measurement

We measure action-trajectory diversity using matched successful rollouts from final Fast-WAM-IDM checkpoints with and without ALIGN on standard LIBERO. For each task, we compute the mean pairwise distance between temporally aligned executed end-effector paths, then average across tasks and suites.



Figure 9: AgileX PiperX platform used for real-world evaluation, with a front-facing RealSense D455 and two wrist-mounted RealSense D435 cameras.



Apple-in-Plate



Cup-Sleeving



Plate-Tidying

Figure 10: The three real-world manipulation tasks. Language instructions are listed in Appendix B.3.

B.3 Real-World Tasks and Instructions

We collected 300 demonstrations for each of the three Piper manipulation tasks. Methods are trained on the same task data and evaluated in 50 trials per task. The language instructions were:

1. **Apple-in-Plate:** "Place the green apple into the plate."
2. **Cup-Sleeving:** "Insert the cup into the cup sleeve."
3. **Plate-Tidying:** "Place the plate into the storage area."

Table 6: Real-world task results. Each entry reports successful trials out of 50, followed by the success rate. All methods use the same task data.

Method	Apple-in-Plate	Cup-Sleeving	Plate-Tidying
$\pi_{0.5}$	41/50 (82%)	39/50 (78%)	25/50 (50%)
Fast-WAM-IDM	24/50 (48%)	21/50 (42%)	14/50 (28%)
Fast-WAM-IDM + ALIGN	**28/50 (56%)**	**25/50 (50%)**	**21/50 (42%)**
OpenWAM-Joint	28/50 (56%)	23/50 (46%)	19/50 (38%)
OpenWAM-Joint + ALIGN	**31/50 (62%)**	**29/50 (58%)**	**22/50 (44%)**

Algorithm 2 Coupled-source flow training and inference

Require: Paired data $\mathcal{D}$; WAM θ; coupling parameters W, ρ
1: **for** each training update **do**
2: Sample paired targets and conditions from $\mathcal{D}$
3: Sample Gaussian sources and couple them using Eq. (9)
4: Sample σ_v, σ_a and form the inputs in Eq. (13)
5: Update θ, W, ρ using Eq. (12)
6: **end for**
7: **for** each inference action chunk **do**
8: Sample and couple video–action sources using Eq. (9)
9: Generate the action chunk with the backbone sampler
10: **end for**

C Coupled-Source Optimization and Alternatives

C.1 Coupled Sources at Training and Inference Time

During training, ALIGN couples the video and action source variables using Eq. (9). Each branch then follows its backbone's flow-matching objective. For clean targets y^v, y^a and sampled noise levels σ_v, σ_a, the training inputs are

$$z^m_{\sigma_m} = (1 - \sigma_m)y^m + \sigma_m \tilde{x}^m_0, \qquad m \in \{v, a\}, \tag{13}$$

where each noise level follows the corresponding backbone training schedule. At inference, ALIGN initializes video and action generation with the learned coupled prior; the backbone then uses its usual sampler. Algorithm 2 summarizes both stages.

C.2 Alternative Coupling Constructions

We describe two alternatives to the action-referenced construction. As in the main text, let $x^v_0 \sim \mathcal{N}(0, I_{d_v})$ and $x^a_0 \sim \mathcal{N}(0, I_{d_a})$ be independent, with $0 \le \rho \le 1$. We use $\tilde{x}^m_0$ for the resulting coupled source of modality m. Both alternatives retain standard Gaussian marginals for the full video and action noise vectors. The projection matrices are independent of the current Gaussian draws.

Video-referenced noise coupling. We keep video noise unchanged and use it to guide action noise through a column-orthonormal projection $P \in \mathbb{R}^{d_v \times d_a}$, $P^\top P = I_{d_a}$ and $d_v \ge d_a$:

$$\tilde{x}^v_0 = x^v_0, \qquad \tilde{x}^a_0 = \rho P^\top x^v_0 + \sqrt{1 - \rho^2}\, x^a_0. \tag{14}$$

Since $P^\top x^v_0$ is standard Gaussian and independent of x^a_0, $\mathrm{Cov}(\tilde{x}^a_0) = \rho^2 I_{d_a} + (1 - \rho^2) I_{d_a} = I_{d_a}$. The video marginal is unchanged.

Shared-source noise coupling. We sample a shared latent $u \sim \mathcal{N}(0, I_k)$ independently of x^v_0, x^a_0, and project it into both modalities using column-orthonormal matrices $U_m \in \mathbb{R}^{d_m \times k}$, $U_m^\top U_m = I_k$, for $m \in \{v, a\}$ and $k \le \min(d_v, d_a)$:

$$\tilde{x}^m_0 = x^m_0 + U_m \Big[\sqrt{\rho}\, u + \big(\sqrt{1 - \rho} - 1 \big) U_m^\top x^m_0 \Big]. \tag{15}$$

This replaces part of each modality's projected noise with shared randomness while preserving its orthogonal residual. Consequently, $\mathrm{Cov}(\tilde{x}^m_0) = (I_{d_m} - U_m U_m^\top) + \rho U_m U_m^\top + (1 - \rho) U_m U_m^\top = I_{d_m}$. All outputs are linear combinations of independent Gaussians, so these identities establish full-vector standard Gaussian marginals. The cross-covariance is $\rho U_v U_a^\top$, so the use of $\sqrt{\rho}$ on each branch matches the correlation strength ρ of the action- and video-referenced constructions. At the same P, ρ, video- and action-referenced sampling have identical joint distributions. Shared-source sampling is also equivalent when $k = d_a$, $U_v = P$, and $U_a = I_{d_a}$. Thus sampling direction alone does not define a distinct prior; differences between trained runs require differences in the learned covariance, optimization, or other implementation settings. At $\rho = 1$, the marginal guarantees still hold, although the joint Gaussian becomes singular and its mutual information diverges. The fixed-strength ablation uses this sampling boundary; the information-regularized learned variant requires $\rho < 1$.

D GAUSSIAN ANALYSIS

D.1 PROOF OF GAUSSIAN MARGINAL PRESERVATION

Fix the learned parameters P and ρ independently of the fresh source draws, with $P^\top P = I_k$ and $k = d_a$. Let $Q = PP^\top$ and $R = I_{d_v} - Q$. The orthogonal decomposition of the original video source is

$$x_0^v = Rx_0^v + Pb, \qquad b = P^\top x_0^v. \tag{16}$$

Since $x_0^v \sim \mathcal{N}(0, I_{d_v})$, we have $b \sim \mathcal{N}(0, I_k)$ and $r := Rx_0^v \sim \mathcal{N}(0, R)$. Their cross-covariance is $\mathrm{Cov}(b, r) = P^\top R = 0$. As they are jointly Gaussian, b and r are independent. The action draw $x_0^a \sim \mathcal{N}(0, I_k)$ is independent of both.

The coupled sources can therefore be written as

$$\tilde{x}_0^v = r + P\big(\sqrt{1-\rho^2}\, b + \rho x_0^a\big), \qquad \tilde{x}_0^a = x_0^a. \tag{17}$$

The vector in parentheses is zero-mean Gaussian, independent of r, with covariance $(1-\rho^2)I_k + \rho^2 I_k = I_k$. Hence

$$\mathrm{Cov}(\tilde{x}_0^v) = R + PP^\top = I_{d_v}, \qquad \mathrm{Cov}(\tilde{x}_0^a) = I_{d_a}. \tag{18}$$

Both outputs are Gaussian and zero-mean, proving that each full source vector retains its standard Gaussian marginal. Their cross-covariance is $\mathrm{Cov}(\tilde{x}_0^v, \tilde{x}_0^a) = \rho P$; thus the construction changes only their joint dependence. The proof holds conditionally at every fixed learned P, ρ, including $\rho = 1$ for sampling, although the joint distribution is then singular.

D.2 JOINT GAUSSIAN AND INFORMATION PENALTY

For fixed P, ρ with $P^\top P = I_{d_a}$, write the stacked coupled source in video–action order. Its covariance is

$$\Sigma = \begin{pmatrix} I_{d_v} & \rho P \\ \rho P^\top & I_{d_a} \end{pmatrix}, \qquad \det \Sigma = (1-\rho^2)^{d_a}. \tag{19}$$

The determinant follows from the Schur complement. For $\rho < 1$, Gaussian mutual information is $-\frac{1}{2}\log\det\Sigma$, giving Eq. (10) with $k = d_a$. It depends on ρ, not the orientation of P, and its derivative is $\partial\mathcal{L}_{\mathrm{cap}}/\partial\rho = k\rho/(1-\rho^2)$. The penalty therefore acts on source dependence rather than measuring semantic agreement between generated behaviors.

The main-text formulas describe one coupling block. If coupling is applied independently across disjoint time windows w, each window uses its own noise coordinates and an orthonormal basis P_w, with $k_w = d_{a,w} \le d_{v,w}$. The full marginal covariances remain identity, and mutual information adds across windows:

$$\mathcal{I}_{\mathrm{total}} = -\frac{1}{2}\sum_w k_w \log(1-\rho_w^2). \tag{20}$$

This sum assumes independent base draws across windows. Reusing the same noise across windows would require a different covariance calculation. The regularization convention (per block or summed over windows) and λ must be recorded when comparing implementations.