

DeVA: Decoupled Video-Action Model with physical guidance for robot policy learning

Mengqi Zhang¹ Sahil Khose¹ Simar Kareer² Yuchen Song¹ Unnat Jain¹ Judy Hoffman¹
¹University of California, Irvine ²Georgia Institute of Technology

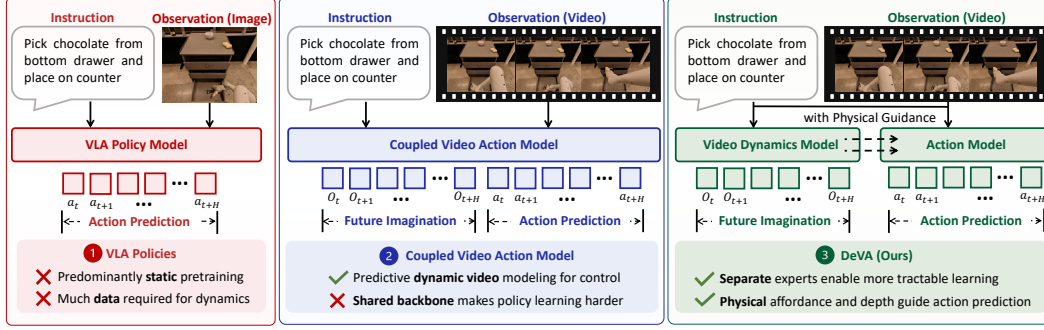


Figure 1: **Overview of DeVA compared with VLA and coupled VAM.** VLA directly maps image-language inputs to actions without explicit dynamics modeling, while coupled video-action models jointly process video and action tokens in a shared representation. DeVA decouples video dynamics modeling from action prediction with physical guidance, enabling efficient robot policy learning.

Abstract: Generalizable robot manipulation requires policies that can anticipate how visual scenes evolve while executing language instructions. While recent Vision-Language-Action models benefit from large-scale pretraining, their predominantly static pretraining objectives provide limited supervision for physical dynamics and temporal causality, leaving control-relevant knowledge to be learned from downstream robot demonstrations. Video generative models offer a promising foundation by encoding rich spatiotemporal priors through future predictions. However, existing Video-Action Models either couple video and action prediction in a shared backbone, making policy adaptation harder to optimize, or under-utilize video information when guiding the action branch. In this work, we introduce **DeVA**, a **Decoupled Video-Action** model with specialized video and action experts, multi-level feature transfer, and physically salient guidance. DeVA transfers representations from multiple video layers to the action expert, enabling rich information exchange while making policy learning more tractable. It further supervises intermediate video features and the action stream with physically salient guidance (affordance/depth). Experiments on both simulation benchmarks and real-world deployment demonstrate strong performance with limited data, faster convergence than a unified architecture, and clear performance gains from physical guidance. Additional videos, code, and checkpoints are available at: deva-model.github.io

Keywords: Video Action Model, Robot Manipulation, Video Generative Models

1 Introduction

Vision-Language-Action (VLA) models [1, 2, 3, 4] have emerged as a leading paradigm for language-conditioned robotic manipulation by adapting vision-language backbones pretrained on web-scale data. This initialization provides strong semantic priors for instruction following and scene understanding. However, the representations are learned predominantly through static image-text objectives with relatively limited temporal supervision and hence provide constrained priors for physical evolution, temporal causality, and the consequences of robot actions. Consequently, much of

this control-relevant knowledge is instead acquired during downstream imitation learning, placing substantial demands on task-specific demonstrations and policy optimization. Video generative models provide a complementary source of knowledge: learning to predict future observations requires modeling how scenes, objects, and interactions evolve over time. This has motivated Video-Action Models (VAMs), which adapt pretrained video backbones to jointly learn future visual dynamics alongside robot actions. Despite this promise, effectively transferring video dynamics to action prediction remains an open architectural challenge, particularly when balancing rich cross-modal interaction with tractable policy optimization.

Existing VAMs address this architectural trade-off through two broad directions. Unified formulations [5, 6, 7] model visual futures and robot actions within a shared backbone or latent representation. Although this design facilitates direct cross-modal exchange, it requires a single feature space to simultaneously support visual generation and control prediction, which restricts the ability to learn the modality-specific features necessary for solving the individual tasks. Earlier approaches [8] repurpose pretrained video models as predictive representation encoders for downstream policies, whereas concurrent dual-DiT architectures [9, 10] allow an action transformer to attend to intermediate video features. Nevertheless, these interfaces typically rely on representations extracted from a selected backbone layer at a fixed denoising stage, leaving the complementary abstractions distributed across the video backbone insufficiently explored. Moreover, whether a unified or dual architecture, prior work primarily acquired control-relevant features implicitly through video and action objectives and did not explore the integration of task-relevant signals such as scene geometry and interaction regions. An important open problem is therefore how to efficiently guide action learning from video pretraining while preserving modality-specific features and focusing on the information from videos most relevant for actions.

To address these challenges, we introduce DeVA, a decoupled Video-Action Model that combines specialized video and action experts with structured multi-level feature transfer and physically salient guidance. Rather than requiring a single representation to support both visual generation and control, DeVA assigns the two objectives to dedicated transformer branches, preserving modality-specific capacity and making downstream policy optimization more tractable. The decoupling, however, introduces the need for an effective interface between the two experts. DeVA therefore routes intermediate representations from multiple levels of the video backbone to corresponding action blocks through layer-wise cross-attention, augmented with learnable bridge tokens that compactly aggregate video dynamics into self-attention. This interface exposes the action expert to comprehensive spatiotemporal representations distributed across the video backbone, building on observations that diffusion features encode distinct abstractions across layers and denoising stages [11, 12]. To further emphasize physical information relevant to manipulation, DeVA applies affordance and relative-depth supervision to intermediate video features. These auxiliary objectives encourage the video backbone to encode task-relevant interaction regions and scene geometry, while the resulting decoder features are injected into the action stream as explicit physical guidance. Together, these components provide an effective framework that combines tractable expert specialization with rich and physically informed video-to-action learning.

We evaluate DeVA on three simulation benchmarks, including RoboCasa [13], LIBERO [14], LIBERO-plus [15], and on real-world bimanual manipulation tasks. Across these settings, DeVA achieves stronger or competitive performance against conventional diffusion policy, finetuned VLA models, and recent Video-Action Models, despite using limited task-specific demonstrations. Under a matched optimization budget, DeVA converges faster and outperforms its unified counterpart. Controlled comparisons further show improvements from affordance and depth guidance, supporting the benefits of the decoupled architecture and physically salient supervision.

2 Related Work

Vision-Language-Action Models. Vision-Language-Action (VLA) models adapt pretrained vision-language representations for language-conditioned robot control. Early generalist policies established the benefits of scaling robot data and transferring web knowledge [16, 17, 4, 18]. Subsequent models expanded open-source pretraining, cross-embodiment learning, and continuous action generation [1,

2, 19, 20, 21, 22]. Recent work further improves action tokenization and adaptation [3, 23, 24], spatial and temporal grounding [25, 26, 27], and unified policy learning [28, 29]. These models inherit strong semantic priors, but their pretraining remains predominantly oriented toward static visual understanding. DeVA instead investigates video-generative pretraining as a complementary source of predictive dynamics for policy learning.

Video-based Action Models. Video prediction has been used for robot control through generated subgoals, inverse dynamics, and predictive representations [8, 30, 31, 32, 33]. Related approaches learn latent actions or use future prediction as an auxiliary policy objective [34, 35, 36]. More recent Video-Action Models (VAMs) model visual futures and actions more directly. Unified formulations place both modalities in a shared backbone or latent space [5, 6, 7, 37], whereas modular formulations connect separate video and action networks through predicted features or attention [9, 10, 38, 39, 40]. DeVA follows the modular direction but combines specialized experts with multi-level transfer across multiple video layers, providing richer video-to-action interaction without requiring a shared representation.

Physical Grounding for Manipulation. Affordance and geometry provide spatial structure for identifying where and how a robot should interact with its environment. Prior work learns actionable regions, object motion, or task-conditioned affordances from images, 3D observations, and human videos [41, 42, 43, 44, 45, 46]. Affordance representations have also been incorporated into hierarchical and diffusion-based policies to improve transfer and demonstration efficiency [47, 48, 49, 50]. Complementary approaches use depth, point clouds, or 3D feature fields for spatially grounded action prediction [51, 52, 53, 54, 55, 56, 57, 58, 59]. Rather than treating these cues only as policy inputs or external planning outputs, DeVA uses affordance and relative-depth supervision to shape intermediate video representations and directly guide the action expert.

3 Method

In this section, we present DeVA, a decoupled Video-Action Model that adapts the video generative backbone for robot policy learning through specialized video and action experts, multi-level feature interaction, and physically salient guidance. Given the current observation O_t and instruction T , the video expert predicts future observations $\{O_{t+1}, \dots, O_{t+h}\}$ that capture plausible scene evolution over a horizon h . In parallel, the action expert predicts the corresponding action sequence $\{A_t, \dots, A_{t+h-1}\}$. Together, the two experts model $P(\{O_{t+1}, \dots, O_{t+h}\}, \{A_t, \dots, A_{t+h-1}\} \mid O_t, T)$, linking predictive visual dynamics to executable control while maintaining specialized representations. DeVA injects intermediate representations from multiple levels of the video backbone to the action expert. It further applies affordance and depth supervision to the video features and uses the resulting physical features to guide action prediction. We describe the architecture in Sec. 3.1, the physical guidance mechanism in Sec. 3.2, and the training and inference pipeline in Sec. 3.3.

3.1 Decoupled Video-Action Architecture with Multi-Level Feature Interaction

Unlike some prior VAMs that model video and action prediction within a shared representation space [6, 7], DeVA in contrast assigns visual dynamics modeling and action prediction to specialized transformer experts. This separation maintains a dedicated pathway for predictive visual modeling while providing action-specific capacity for robot control. To remain connected, the two experts leverage multi-level feature interaction, implemented using layer-wise cross-attention and learnable bridge tokens.

Decoupled Architecture. DeVA consists of a video expert for future observation modeling and an action expert for robot action prediction (Fig. 2). The video expert is initialized from Cosmos-Predict2 [60], a latent video diffusion transformer. Its spatiotemporal VAE encodes video observations into compact latent tokens, and the video transformer models the evolution conditioned on the current frame and language instruction. The action expert follows a similar DiT-style architecture in the action domain and predicts the action trajectory aligned with the corresponding predicted future. The two experts maintain separate parameters and token representations for their respective objectives. This avoids requiring visual dynamics and robot actions to share a single feature space. At the same time, the experts are jointly optimized using separate generative objectives, allowing the action policy to benefit from the video dynamics while retaining its specialized control representation.

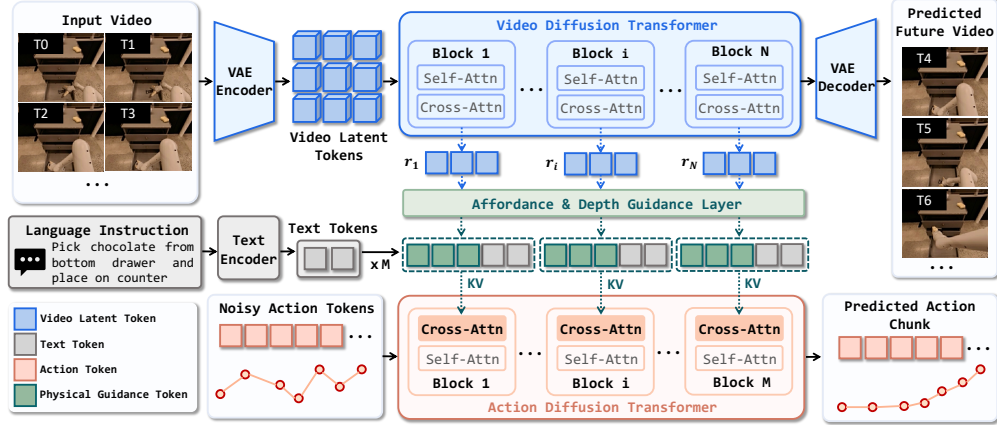


Figure 2: **Overview.** DeVA decouples video dynamics modeling and action prediction. The action stream attends to layer-wise video features through cross-attention to produce robot action chunks.

Multi-Level Feature Interaction. Although separate experts preserve modality-specific capacity, fully isolating them would prevent the action expert from accessing the predictive dynamics learned by the video backbone. Effective decoupling therefore requires structured interaction between the two experts. Moreover, diffusion models encode complementary representations across network layers and denoising stages [11, 12]. DeVA transfers intermediate video representations from multiple levels to corresponding action blocks through layer-wise cross-attention. Learnable bridge tokens provide an additional compact interface by aggregating video context and introducing it into the action stream through self-attention. Together, these mechanisms expose the action expert to complementary predictive representations distributed across the video backbone while preserving its specialized control capacity.

3.2 Physically Salient Guidance

The multi-level interaction interface provides the action expert with predictive video representations. However, these representations are optimized primarily for future prediction and may not explicitly emphasize the interaction regions and geometric structure needed for successful control. DeVA therefore applies task-conditioned affordance and relative-depth supervision to intermediate video features. The resulting decoder features are also injected into the action expert as physical guidance. This mechanism both shapes the video representations and directly conditions action prediction.

Affordance & Depth Decoding. We repurpose intermediate video features as dense representations for manipulation perception. Concretely, DeVA uniformly samples features from the video backbone layers and feeds them into lightweight DPT-style decoders with interleaved temporal-attention layers, as illustrated in Fig. 3. The decoders predict two complementary physical signals: task-conditioned affordance maps and relative-depth maps. We define the affordance map over the image plane as

$$\mathcal{A}_t(u, v) = P(p_t^{ee} = (u, v) \mid O_t, T),$$

where p_t^{ee} denotes the image-plane location of the end effector and each pixel represents its likelihood as an interaction location under instruction T . Language features are incorporated into the affordance decoder through FiLM layers. In simulation, we construct affordance targets from ground-truth contact locations or projected end-effector positions and convert them into smooth heatmaps using Gaussian kernels. For real-world data, where such annotations are unavailable, we generate pseudo-labels using an off-the-shelf affordance model [50]. We similarly obtain relative monocular depth estimations with a depth prediction model [57]. Additional details are provided in the Appendix.

Physical Guidance Injection. Beyond providing auxiliary supervision, the decoded physical features directly guide the action expert. We extract features from the final spatiotemporal blocks of the decoders, project them to the spatial resolution of the video features, and flatten them into tokens. These features are concatenated with the multi-level video representations along the channel dimension and supplied as additional keys and values to the action cross-attention layers. The action

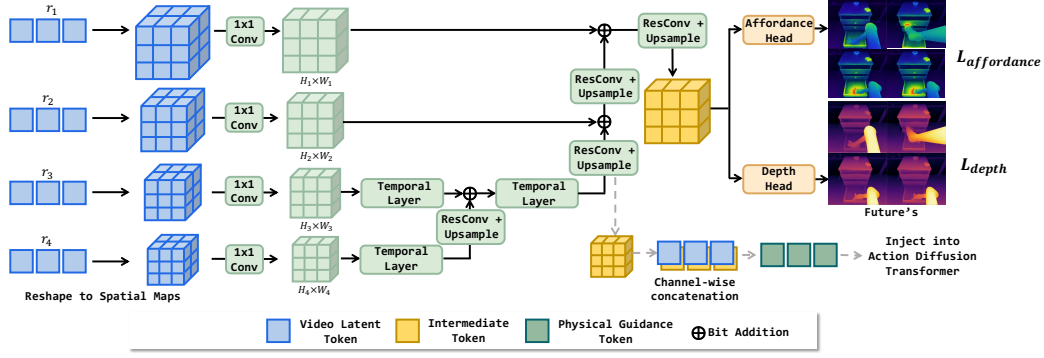


Figure 3: **Physical guidance module.** The module aggregates multi-level video features and decodes them into affordance and depth maps, which are then converted into physically grounded tokens and injected into the action transformer head to support action prediction.

expert can therefore attend to both predictive video dynamics and explicit cues about where and how the end effector should act. In this way, physically salient guidance both emphasizes control-relevant structure in the video backbone and makes this structure directly available to action prediction.

3.3 Training and Inference

The separation between the video and action experts supports a two-stage training procedure. In Stage 1, we warm up the video expert and physical decoders for 10K steps, jointly learning future visual dynamics, affordance, and depth. In Stage 2, we introduce the action expert and jointly optimize video and action prediction. The physical decoders are frozen during this stage, but continue to supervise the video features and provide guidance to the action expert.

Stage 1: Video and Decoder Warmup. The video expert is trained in latent space using the standard EDM denoising objective $\mathcal{L}_{\text{video}}$. We additionally supervise the affordance and depth decoders to encourage the intermediate video features to encode physically salient structure. The affordance decoder is trained using a noise-weighted pixel-wise binary cross-entropy loss:

$$\mathcal{L}_{\text{aff}} = \mathbb{E}_t \left[w(t) \ell_{\text{BCE}}(\hat{\mathcal{A}}_t, \mathcal{A}_t) \right], \quad w(t) = \frac{1}{(1 + \sigma_t)^2},$$

where t denotes the diffusion noise step and σ_t is the corresponding noise level. The loss is averaged over spatial locations and includes positive-class reweighting to address the foreground-background imbalance in the affordance maps. The depth decoder is trained using a noise-weighted masked reconstruction loss:

$$\mathcal{L}_{\text{depth}} = \mathbb{E}_t \left[w(t) \left(\ell_1^{\text{mask}}(\hat{D}_t, D_t) + \lambda_{\text{grad}} \ell_{\text{grad}}^{\text{mask}}(\hat{D}_t, D_t) \right) \right],$$

where ℓ_1^{mask} is computed only over valid depth pixels, while $\ell_{\text{grad}}^{\text{mask}}$ encourages local geometric consistency. The overall objective in Stage 1 is:

$$\mathcal{L}_v = \mathcal{L}_{\text{video}} + \lambda_{\text{aff}} \mathcal{L}_{\text{aff}} + \lambda_{\text{depth}} \mathcal{L}_{\text{depth}}.$$

Stage 2: Joint Video-Action Training. In Stage 2, we introduce the action expert and jointly optimize it with the video expert. The parameters of the physical decoders are frozen to maintain a stable physical prediction interface. However, the affordance and depth predictions remain active, allowing their gradients to pass through the fixed decoders and continue shaping the intermediate video features. The decoded physical features are also retained as direct guidance for the action expert. The action expert is trained using a flow-matching objective $\mathcal{L}_{\text{act}}$ conditioned on the current observation, language instruction, multi-level video representations, and decoded physical guidance. The complete Stage 2 objective is:

$$\mathcal{L} = \mathcal{L}_v + \mathcal{L}_{\text{act}}.$$

Inference. At inference time, the future video latents and action trajectory are initialized from Gaussian noise and sampled jointly through their respective generative processes. The video expert predicts a future visual rollout, while the action expert generates the corresponding action chunk. The lightweight physical decoders remain active to provide affordance- and depth-aware guidance.

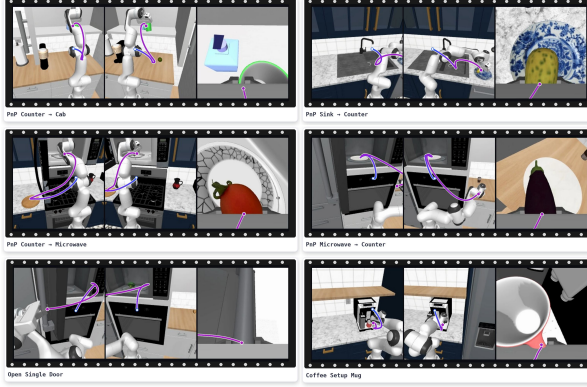


Figure 4: RoboCasa task rollout examples.

Success Rate	# Demos / Task	Avg.
GR00T-N1 [22]	300	49.6
DP-VLA [61]	3000	57.3
GR00T-N1 + DreamGen [33]	300 (+10K syn.)	57.6
GR00T-N1 + DUST [62]	300	58.5
π_0 [2]	300	62.5
GR00T-N1.5 [22]	300	64.1
FLARE [36]	300	66.4
GR00T-N1.5 + HAMLET [63]	300	66.4
Video-Action Models		
Video Policy [38]	300	66.0
UVA [37]	50	50.0
UWM [5]	1000	60.8
Cosmos-Policy [6]	50	67.1
Ours	50	72.0

Table 1: RoboCasa benchmark comparison.

4 Experiment

In this section, we evaluate and explore the potential of DeVA from three perspectives: benchmark comparison over state-of-the-art policy models; data efficiency and speedup based on our two-flow architecture; scalable explorations over different axes.

4.1 Benchmarks

Simulation Benchmarks. We evaluate DeVA on three complementary simulation benchmarks covering multitask manipulation, diverse environments, and robustness under controlled shifts:

- **RoboCasa [13].** A large-scale simulation benchmark featuring realistic kitchen environments and diverse everyday manipulation tasks, including pick-and-place and appliance operation.
- **LIBERO [14].** A tabletop manipulation benchmark organized into task suites that vary in spatial layouts, object identities, goal specifications, and task horizons.
- **LIBERO-Plus [15].** A robustness benchmark extending LIBERO with controlled perturbations across seven dimensions: object layout, camera viewpoint, robot initial state, language instruction, lighting condition, background texture, and sensor noise.

Together, these benchmarks evaluate policy performance across controlled tabletop tasks, visually diverse household environments, and systematic distribution shifts.

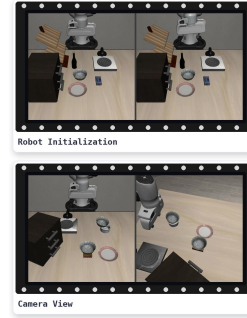
Real World Manipulation. Our real-world bimanual manipulation platform consists of two I2RT YAM Ultra robotic arms [68]. Each arm provides 6-DoF motion and a 1-DoF parallel gripper, resulting in a total of 14-DoF. The perception system uses three RGB-D cameras: one wrist-mounted camera on each arm for close-range observations and one external head-view camera for global scene context. Expert demonstrations are collected using a Meta Quest Pro teleoperation interface based on OpenTeleVision [69]. We evaluate DeVA on three bimanual tasks: handing over a marker between arms, lifting a pot, and picking up bottles of diverse appearances. These tasks test different forms of bimanual control, including inter-arm handoff, coordinated object handling, and object variations.

Success Rate	Spatial	Object	Goal	Long	Short-Avg	Avg.
Diffusion Policy [64]	78.3	92.5	68.3	50.5	79.7	72.4
π_0 -fast [23]	96.4	96.8	88.6	60.2	93.9	85.5
$\pi_{0.5}$ [3]	98.8	98.2	98.0	92.4	98.3	96.9
OpenVLA-OFT [24]	97.6	98.4	97.9	94.5	98.0	97.1
CogVLA [29]	98.6	98.8	96.6	95.4	98.0	97.4
Video-Action Models						
UVA [37]	-	-	-	90.0	-	-
UWM [5]	-	-	-	79.0	-	-
ViPRA [39]	-	-	-	79.0	-	-
Mimic-Video [9]	94.2	96.8	90.6	-	93.9	-
Cosmos-Policy [6]	98.1	100.0	98.2	<u>97.6</u>	98.8	98.5
DiT4DiT [10]	98.4	<u>99.6</u>	<u>98.6</u>	<u>97.6</u>	<u>98.9</u>	<u>98.6</u>
Ours	99.2	<u>99.6</u>	98.8	98.4	99.2	99.0

(a) LIBERO benchmark comparison.

Success Rate	Avg.
OpenVLA [1]	15.6
WorldVLA [65]	25.0
NORA [66]	39.0
UniVLA [28]	42.9
π_0 [2]	53.6
OpenVLA-OFT_w [24]	55.8
π_0 -fast [23]	61.6
OpenVLA-OFT_m [24]	67.9
RIPT-VLA [67]	68.4
OpenVLA-OFT [24]	<u>69.6</u>
Ours	80.8

(b) LIBERO-Plus.



(c) Perturbation.

Figure 5: Simulation benchmark comparisons on LIBERO and LIBERO-Plus. The robustness examples are on "Robot Init" and "Camera View" axes.

4.2 Baselines

Depending on the benchmark, we compare DeVA with baselines from three model families. First, conventional task-specific policies are represented by Diffusion Policy [64]. Second, VLA models include OpenVLA and its variants [1], CogVLA [29], the π -series [2, 3, 23], and GR00T-series [22]. Third, video-action baselines include Video Policy [38], UVA [37], UWM [5], ViPRA [39], Mimic-Video [9], Cosmos Policy [6], and DiT4DiT [10].

4.3 Simulation Results

RoboCasa. We evaluate DeVA on 24 kitchen manipulation tasks. For each task, we conduct 50 evaluation rollouts for three different seeds, resulting in 3,600 rollouts in total. The benchmark covers a range of manipulation challenges. Cabinet and microwave tasks require object retrieval in cluttered or constrained workspaces, while mug placement and stove manipulation demand precise object localization and control, as illustrated in Fig. 4. As shown in Table 1, DeVA achieves a success rate of 72.0%, outperforming different types of baselines. Importantly, DeVA uses only 50 demonstrations per task. Under the same data budget, it improves up to 22.0 points, while also outperforming several methods trained with 300–3,000 demonstrations per task. These results demonstrate strong data efficiency without relying on additional synthetic trajectories.

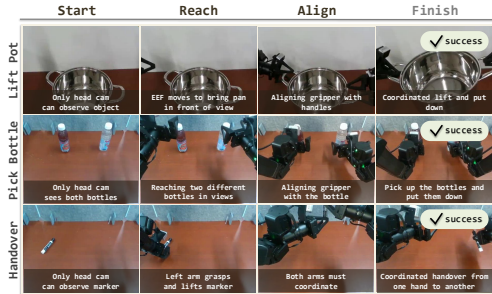
LIBERO & LIBERO-plus. We first evaluate DeVA on the four standard LIBERO suites: Spatial, Object, Goal, and Long, comprising 40 tasks in total. We evaluate every task over 50 rollouts using three seeds. We report the average success rate in Fig. 5. DeVA achieves the highest overall average success rate of 99.0%. For VAM baselines that report only LIBERO-Long, we restrict the comparison to that suite. Overall, DeVA maintains consistently high performance across spatial, object-centric, goal-conditioned, and long-horizon tasks. We further evaluate robustness on LIBERO-Plus, which contains 10,030 variants generated by applying controlled perturbations to the original LIBERO tasks. Following the official protocol, each variant is evaluated with one rollout. DeVA achieves an average success rate of 80.8%, exceeding the strongest baseline shown in the table, OpenVLA-OFT, by 11.2 percentage points. The 18.2-point drop from 99.0% on standard LIBERO also reveals a remaining gap between in-distribution and robustness.

4.4 Real-World Deployment.

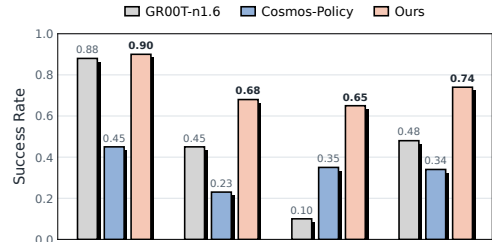
We evaluate DeVA on three bimanual tasks using the I2RT platform: Handover Marker, Lift Pot, and Pick Up Bottles. For each task, we collect 30 expert demonstrations and fine-tune DeVA, Cosmos Policy, and GR00T-N1.6 in a multi-task manner and using the same optimization budget. Each policy is evaluated over 10 trials per task. Figure 6(a) presents a representative successful rollout of Lift Pot, which requires precise grasping and coordinated motion between the two arms. As shown in Fig. 6(b), DeVA achieves an average success rate of 74%, compared with 48% for GR00T-N1.6 and 34% for Cosmos Policy. Additional detailed task-level analysis is provided in the Appendix.

4.5 Learning and Data Efficiency

We evaluate training efficiency from two perspectives: optimization convergence and data budget. First, we replace the decoupled experts and their feature-transfer interface with a unified backbone while keeping the remaining training setup unchanged. As shown in Fig. 7(b), both decoupled variants approach convergence within approximately 40-50K steps. At this point, the base and physically guided variants achieve success rates of approximately 66% and 71%, respectively, compared with 34% for the unified counterpart. Based on



(a) Real-world task keyframes.



(b) Real-world success rates comparison.

Figure 6: Real-robot keyframes for three tasks' completion and quantitative success rates comparison with the other two VLA/VAM.

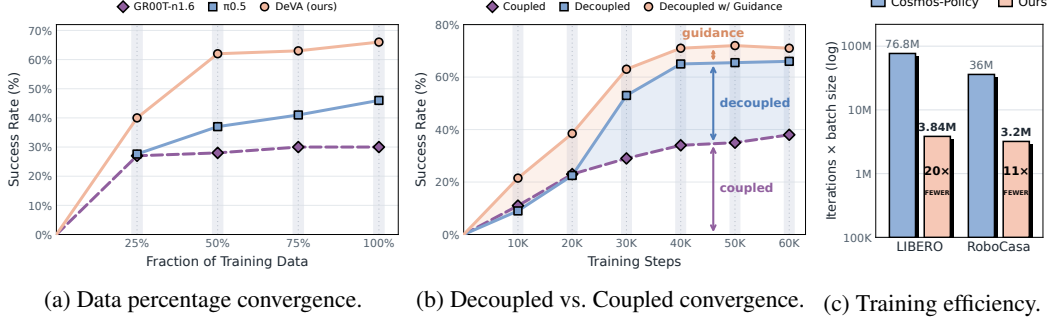


Figure 7: Comparison over training optimization and data efficiency.

the official reports in [6], DeVA processes up to 20 $\times$ fewer training examples than Cosmos-Policy as shown in Fig. 7(c). This comparison shows that specialized experts with structured feature transfer achieve both faster convergence and higher performance under the same optimization budget.

We further train DeVA and the VLA baselines using matched fractions of the RoboCasa demonstrations to reveal the learning curve. As shown in Fig. 7(a), DeVA achieves a higher success rate at every shared data budget and continues to improve as additional demonstrations are introduced. At the full data budget, DeVA (base variant) reaches approximately 66%, compared with 48% for $\pi_{0.5}$ and 29% for GR00T-N1.6. Together, these results show that the decoupled architecture improves both optimization tractability and the effective use of task-specific demonstrations.

5 Ablation

We ablate two aspects of DeVA on RoboCasa under a matched training budget: predictive visual modeling and physical guidance. As shown in Fig. 8(a), goal-image prediction improves the success rate from 19.8% to 25.8% over the action-only model, while future-video prediction further increases it to 36.8%. Incorporating the decoupled architecture with multi-level feature transfer raises performance to 66.0%, demonstrating the importance of both predictive video priors and their effective transfer to the action expert. Fig. 8(b) further shows that affordance and relative-depth guidance independently improve the base model. Combining both cues achieves the highest success rate of 72.0%, indicating their complementary contributions to physically grounded action prediction.

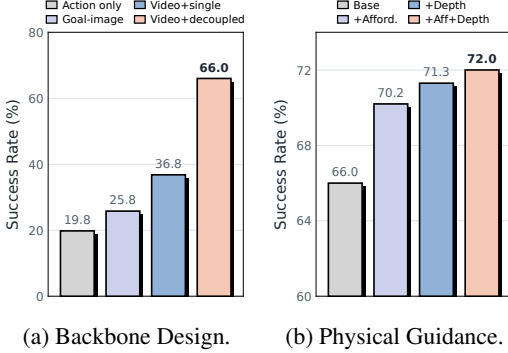


Figure 8: Ablations of the video-prediction architecture and physical-guidance components, evaluated using ~ 55 K-step checkpoints.

6 Limitation

While the results are encouraging, DeVA inherits some computational cost from its pretrained video backbone. Training spatiotemporal attention over future observations remains more expensive than action-only policy learning. Jointly denoising the video and action streams introduces additional inference overhead compared with models that predict only compact latent states and actions. Improving efficiency through latent-space prediction and accelerated sampling is an important direction for future work.

7 Conclusion

We present DeVA, a decoupled Video-Action Model that transfers predictive video representations to robot control through specialized experts, multi-level feature transfer, and physically salient affordance and depth guidance. Across simulation benchmarks and real-world bimanual manipulation, DeVA achieves strong performance with limited demonstrations, converges faster than a matched unified architecture, and benefits from physical guidance, demonstrating the promise of video generative priors for robot policy learning.

Acknowledgments

This work was supported in part by funding from NSF awards #2622839 and #2403297. All views and conclusions expressed in this work are those of the authors and not a reflection of these source

References

- [1] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, P. R. Sanketi, Q. Vuong, et al. Openvla: An open-source vision-language-action model. In *Conference on Robot Learning*, pages 2679–2713. PMLR, 2025.
- [2] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, et al. π_0 : A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.
- [3] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, et al. $\pi_0.5$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.
- [4] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.
- [5] 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. In *Proceedings of Robotics: Science and Systems (RSS)*, 2025.
- [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, and J. Gu. Cosmos policy: Fine-tuning video models for visuomotor control and planning. In *International Conference on Learning Representations (ICLR)*, 2026.
- [7] S. Ye, Y. Ge, K. Zheng, S. Gao, S. Yu, G. Kurian, S. Indupuru, Y. L. Tan, C. Zhu, J. Xiang, A. Malik, K. Lee, W. Liang, N. Ranawaka, J. Gu, Y. Xu, G. Wang, F. Hu, A. Narayan, J. Bjorck, J. Wang, G. Kim, D. Niu, R. Zheng, Y. Xie, J. Wu, Q. Wang, R. Julian, D. Xu, Y. Du, Y. Chebotar, S. Reed, J. Kautz, Y. Zhu, L. J. Fan, and J. Jang. World action models are zero-shot policies, 2026. URL <https://arxiv.org/abs/2602.15922>.
- [8] 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. In *International Conference on Machine Learning*, pages 24328–24346. PMLR, 2025.
- [9] 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 2512.15692*, 2025.
- [10] T. Ma, J. Zheng, Z. Wang, C. Jiang, A. Cui, J. Liang, and S. Yang. Dit4dit: Jointly modeling video dynamics and actions for generalizable robot control. *arXiv preprint arXiv:2603.10448*, 2026.
- [11] G. Luo, L. Dunlap, D. H. Park, A. Holynski, and T. Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36:47500–47510, 2023.
- [12] D. Kim, X. Thomas, and D. Ghadiyaram. Revelio: Interpreting and leveraging semantic information in diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4659–4669, 2025.
- [13] S. Nasiriany, A. Maddukuri, L. Zhang, A. Parikh, A. Lo, A. Joshi, A. Mandlekar, and Y. Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. In *Robotics: Science and Systems (RSS)*, 2024.

- [14] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36: 44776–44791, 2023.
- [15] S. Fei, S. Wang, J. Shi, Z. Dai, J. Cai, P. Qian, L. Ji, X. He, S. Zhang, Z. Fei, et al. Libero-plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025.
- [16] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.
- [17] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, et al. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*, pages 8469–8488. PMLR, 2023.
- [18] K. Bousmalis, G. Vezzani, D. Rao, C. Devin, A. X. Lee, M. Bauzá, T. Davchev, Y. Zhou, A. Gupta, A. Raju, et al. Robocat: A self-improving generalist agent for robotic manipulation. *arXiv preprint arXiv:2306.11706*, 2023.
- [19] Octo Model Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, C. Xu, J. Luo, T. Kreiman, Y. Tan, L. Y. Chen, P. Sanketi, Q. Vuong, T. Xiao, D. Sadigh, C. Finn, and S. Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.
- [20] Q. Li, Y. Liang, Z. Wang, L. Luo, X. Chen, M. Liao, F. Wei, Y. Deng, S. Xu, Y. Zhang, et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024.
- [21] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. In *International Conference on Learning Representations*, volume 2025, pages 29982–30009, 2025.
- [22] J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.
- [23] K. Pertsch, K. Stachowicz, B. Ichter, D. Driess, S. Nair, Q. Vuong, O. Mees, C. Finn, and S. Levine. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025.
- [24] M. J. Kim, C. Finn, and P. Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025.
- [25] H. Zhen, X. Qiu, P. Chen, J. Yang, X. Yan, Y. Du, Y. Hong, and C. Gan. 3d-vla: A 3d vision-language-action generative world model. In *International Conference on Machine Learning*, pages 61229–61245. PMLR, 2024.
- [26] D. Qu, H. Song, Q. Chen, Y. Yao, X. Ye, Y. Ding, Z. Wang, J. Gu, B. Zhao, D. Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025.
- [27] R. Zheng, Y. Liang, S. Huang, J. Gao, H. Daumé III, A. Kolobov, F. Huang, and J. Yang. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. In *International Conference on Learning Representations*, volume 2025, pages 54277–54296, 2025.
- [28] Q. Bu, Y. Yang, J. Cai, S. Gao, G. Ren, M. Yao, P. Luo, and H. Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025.

- [29] W. Li, R. Zhang, R. Shao, J. He, and L. Nie. Cogvla: Cognition-aligned vision-language-action model via instruction-driven routing & sparsification. *Advances in neural information processing systems*, 2025.
- [30] Y. Du, S. Yang, B. Dai, H. Dai, O. Nachum, J. Tenenbaum, D. Schuurmans, and P. Abbeel. Learning universal policies via text-guided video generation. *Advances in neural information processing systems*, 36:9156–9172, 2023.
- [31] K. Black, M. Nakamoto, P. Atreya, H. Walke, C. Finn, A. Kumar, and S. Levine. Zero-shot robotic manipulation with pre-trained image-editing diffusion models. In *International Conference on Learning Representations*, volume 2024, pages 33431–33452, 2024.
- [32] H. Bharadhwaj, D. Dwibedi, A. Gupta, S. Tulsiani, C. Doersch, T. Xiao, D. Shah, F. Xia, D. Sadigh, and S. Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. In *Conference on Robot Learning*, pages 3936–3951. PMLR, 2025.
- [33] J. Jang, S. Ye, Z. Lin, J. Xiang, J. Bjorck, Y. Fang, F. Hu, S. Huang, K. Kundalia, Y.-C. Lin, et al. Dreamgen: Unlocking generalization in robot learning through video world models. In *Conference on Robot Learning*, pages 5170–5194. PMLR, 2025.
- [34] S. Ye, J. Jang, B. Jeon, S. J. Joo, J. Yang, B. Peng, A. Mandlekar, R. Tan, Y.-W. Chao, B. Y. Lin, et al. Latent action pretraining from videos. In *International Conference on Learning Representations*, volume 2025, pages 28213–28239, 2025.
- [35] 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.
- [36] 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. In *Conference on Robot Learning*, pages 3952–3971. PMLR, 2025.
- [37] S. Li, Y. Gao, D. Sadigh, and S. Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025.
- [38] 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.
- [39] S. Routray, H. Pan, U. Jain, S. Bahl, and D. Pathak. Vipra: Video prediction for robot actions. *arXiv preprint arXiv:2511.07732*, 2025.
- [40] T. Yuan, Z. Dong, Y. Liu, and H. Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026.
- [41] K. Mo, L. J. Guibas, M. Mukadam, A. Gupta, and S. Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6813–6823, 2021.
- [42] S. Deng, X. Xu, C. Wu, K. Chen, and K. Jia. 3d affordancenet: A benchmark for visual object affordance understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1778–1787, 2021.
- [43] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13778–13790, 2023.
- [44] C. Yuan, C. Wen, T. Zhang, and Y. Gao. General flow as foundation affordance for scalable robot learning. In *Conference on Robot Learning*, pages 1541–1566. PMLR, 2025.

- [45] W. Yuan, J. Duan, V. Blukis, W. Pumacay, R. Krishna, A. Murali, A. Mousavian, and D. Fox. Robopoint: A vision-language model for spatial affordance prediction in robotics. In *Conference on Robot Learning*, pages 4005–4020. PMLR, 2025.
- [46] W. Huang, C. Wang, Y. Li, R. Zhang, and L. Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. In *Conference on Robot Learning*, pages 4573–4602. PMLR, 2025.
- [47] S. Nasiriany, S. Kirmani, T. Ding, L. Smith, Y. Zhu, D. Driess, D. Sadigh, and T. Xiao. Rt-affordance: Affordances are versatile intermediate representations for robot manipulation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8249–8257. IEEE, 2025.
- [48] K. Rana, J. Abou-Chakra, S. Garg, R. Lee, I. Reid, and N. Suenderhauf. Learning from 10 demos: Generalisable and sample-efficient policy learning with oriented affordance frames. In *Conference on Robot Learning*, pages 5464–5482. PMLR, 2025.
- [49] S. Wu, Y. Zhu, Y. Huang, K. Zhu, J. Gu, J. Yu, Y. Shi, and J. Wang. Affordddp: Generalizable diffusion policy with transferable affordance. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6971–6980, 2025.
- [50] Y. Tang, W. Huang, Y. Wang, C. Li, R. Yuan, R. Zhang, J. Wu, and L. Fei-Fei. Uad: Un-supervised affordance distillation for generalization in robotic manipulation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3822–3831. IEEE, 2025.
- [51] M. Shridhar, L. Manuelli, and D. Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR, 2023.
- [52] A. Goyal, J. Xu, Y. Guo, V. Blukis, Y.-W. Chao, and D. Fox. Rvt: Robotic view transformer for 3d object manipulation. In *Conference on Robot Learning*, pages 694–710. PMLR, 2023.
- [53] T. Gervet, Z. Xian, N. Gkanatsios, and K. Fragkiadaki. Act3d: 3d feature field transformers for multi-task robotic manipulation. In *Conference on Robot Learning*, pages 3949–3965. PMLR, 2023.
- [54] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.
- [55] Y. Jia, J. Liu, S. Chen, C. Gu, Z. Wang, L. Luo, L. Lee, P. Wang, Z. Wang, R. Zhang, et al. Lift3d foundation policy: Lifting 2d large-scale pretrained models for robust 3d robotic manipulation. *arXiv preprint arXiv:2411.18623*, 2024.
- [56] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.
- [57] S. Chen, H. Guo, S. Zhu, F. Zhang, Z. Huang, J. Feng, and B. Kang. Video depth anything: Consistent depth estimation for super-long videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22831–22840. IEEE, 2025.
- [58] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024.
- [59] L. Fan, Z. Xu, C. Cao, W. Zhang, M. Yuan, and J. Chen. Aim: Intent-aware unified world action modeling with spatial value maps. *arXiv preprint arXiv:2604.11135*, 2026.
- [60] NVIDIA. Cosmos-predict2: World simulation model for physical ai, 2025. URL <https://github.com/nvidia-cosmos/cosmos-predict2>.

- [61] B. Han, J. Kim, and J. Jang. A dual process vla: Efficient robotic manipulation leveraging vlm. *arXiv preprint arXiv:2410.15549*, 2024.
- [62] 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.
- [63] M. Koo, D. Choi, T. Kim, K. Lee, C. Kim, Y. Seo, and J. Shin. Hamlet: Switch your vision-language-action model into a history-aware policy. *arXiv preprint arXiv:2510.00695*, 2025.
- [64] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.
- [65] 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.
- [66] C.-Y. Hung, Q. Sun, P. Hong, A. Zadeh, C. Li, U. Tan, N. Majumder, S. Poria, et al. Nora: A small open-sourced generalist vision language action model for embodied tasks. *arXiv preprint arXiv:2504.19854*, 2025.
- [67] S. Tan, K. Dou, Y. Zhao, and P. Krähenbühl. Interactive post-training for vision-language-action models. *arXiv preprint arXiv:2505.17016*, 2025.
- [68] I2RT Robotics. YAM Arm Series. <https://doc.i2rt.com/products/yam>, 2024.
- [69] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang. Open-television: Teleoperation with immersive active visual feedback. In *Conference on Robot Learning*, pages 2729–2749. PMLR, 2025.
- [70] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2955–2966, 2023.

Appendix

A Implementation Details

A.1 Video Expert Architecture

In DeVA, we adopt Cosmos-Predict2 [60] as the video generative backbone and initialize it from the pretrained 480p/16-fps checkpoint. During training, the video stream is visually conditioned only on the initial observation frame, while the remaining frames serve as future prediction targets. We retain the pretrained T5 text encoder for language conditioning. Multi-view observations are spatially tiled into a single composite frame. For example, RoboCasa provides three views, `agentview_left`, `agentview_right`, and `end_effector_view`, which are arranged in either a 1×3 or 2×2 layout, with the unused tile in the latter left blank. We observe negligible performance differences between the two layouts. Tiling allows all views to be processed within a single video stream, reducing computation compared with processing them separately. The video backbone comprises 28 DiT blocks and approximately 2B parameters. It is trained using the standard EDM denoising objective in latent space:

$$\mathcal{L}_{\text{video}} = \mathbb{E}_t \left[\lambda(t) \ell_{\text{EDM}}(\hat{V}_t, V_t) \right],$$

where $\hat{V}_t$ denotes the denoised video latent prediction at diffusion time t , V_t denotes the ground-truth target latent, and $\lambda(t)$ is the standard EDM weighting term.

A.2 Action Expert Architecture

The action expert follows a standard action-DiT architecture based on the action head of GR00T-N1.5 [22]. To avoid redundant visual encoding, it is conditioned directly on the intermediate video representations, T5 language embeddings, and physical guidance, without a separate visual encoder. Because the initial observation is already encoded by the video backbone, the transferred features provide both the current visual context and predictive dynamics required for action generation. An input MLP projects noisy action trajectories into the model dimension, while an output MLP maps the denoised tokens back to the action space.

Alongside the action tokens, 12 learnable bridge tokens of dimension 1024 are provided as input and expanded across the batch to a shape of $(n, 12, 1024)$. Within the self-attention layers, these tokens aggregate transferred video information and propagate global context across the action stream. The action expert contains 16 layers and approximately 500M parameters, with self-attention and cross-attention interleaved across the network. Accordingly, eight layers are uniformly sampled from early to late stages of the video backbone and routed to the corresponding cross-attention layers. The action expert is trained using a flow-matching objective in the action space:

$$\mathcal{L}_{\text{act}} = \mathbb{E}_t \left[\left\| v_{\theta}(\tilde{A}_t, t, c) - u_t \right\|_2^2 \right],$$

where t denotes the flow-matching timestep, $\tilde{A}_t$ denotes the interpolated noisy action trajectory at time t , u_t is the corresponding target flow, and c denotes the conditioning inputs, including observations, language instructions, and video features and physical guidance.

A.3 Affordance and Depth Guidance Decoder

Intermediate features from video generative backbones provide useful perceptual representations for downstream tasks. Prior work [70] has explored transferring diffusion-model features to classification, segmentation, and other dense prediction tasks. In DeVA, we attach two DPT-style decoders to the video backbone for affordance and relative-depth prediction. To capture temporal structure, temporal

transformer blocks are inserted before the residual convolutional blocks, enabling information exchange across frames and encouraging temporally consistent predictions.

We uniformly sample eight intermediate layers across the video backbone to distribute physical supervision across its depth. Four layers provide multi-scale features to the affordance decoder, while the remaining four provide features to the relative-depth decoder. Because a scene may contain multiple plausible interaction regions depending on the task, the affordance decoder is conditioned on the T5 language embeddings through FiLM layers. Each decoder contains four decoding stages, and the two decoders add approximately 40M parameters in total, which is small relative to the combined video and action experts.

Features from the third stage of both decoders are further used as physical guidance for the action expert. These features are resampled to the spatiotemporal resolution of the video tokens and concatenated with the corresponding video features along the channel dimension. The fused features are then projected to the hidden dimension of the text tokens and appended to them along the token dimension. The resulting conditioning sequence provides the keys and values for the cross-attention layers in the action expert.

A.4 Parameter Overview

DeVA combines a pretrained 2B-parameter Video2World diffusion transformer (DiT) with a specialized action expert and two lightweight physical-guidance decoders. Table 2 reports the parameter count of each component. The T5-XXL text encoder (4.86B parameters) and Cosmos continuous VAE tokenizer remain frozen throughout training, providing language-conditioning embeddings and mapping videos to and from the latent space, respectively. The components optimized across our two-stage pipeline include the Video2World DiT (1.96B parameters) and action expert (564M parameters), comprising the state/action embedders, action transformer, and cross-attention modules for video feature transfer. The affordance decoder (17.6M), relative-depth decoder (13.4M), and fusion projection (5.2M) add 36.3M parameters, corresponding to only 1.4% of the parameters optimized during training. The decoders are trained during Stage 1 and frozen during Stage 2, while the fusion projection injects their features into the action-conditioning pathway. Overall, DeVA contains 7.55B parameters, of which 2.56B (33.9%) are updated during the two-stage training pipeline.

Table 2: Parameter overview of DeVA with information about frozen and active modules.

Component	Parameters	Status
T5-XXL	4,864,791,552	frozen
VAE tokenizer	126,892,531	frozen
Video2world DiT	1,956,413,440	trained
Action head	563,865,614	trained
Affordance head	17,630,465	trained
Depth head	13,432,065	trained
Joint-fusion projection	5,244,928	trained
Trainable total	2,556,586,512	
Frozen total	4,991,684,083	
Total	7,548,270,595	

A.5 Training Scheme

Following the two-stage pipeline, we fine-tune the 2B-parameter Video2World diffusion transformer (DiT) for future-video prediction and train the action expert using action-space flow matching. Two auxiliary dense-prediction objectives supervise the affordance-heatmap and relative-depth decoders. Training is performed on eight GPUs using Fully Sharded Data Parallelism, with a per-GPU batch size of 8 (global batch size 64) and bfloat16 mixed precision. The optimization, data, and auxiliary-decoder settings are summarized in Tables 3, 4, and 5, respectively.

We use FusedAdamW with a piecewise learning-rate schedule. The learning rate is linearly increased over the first 1,000 steps to a peak value of 1×10^{-4} and then linearly decayed to 3×10^{-5} ($0.3 \times$ the peak) by step 30k. It is subsequently reduced to 6×10^{-6} ($0.06 \times$ the peak) and held constant for the remainder of training. Gradients are clipped to a global norm of 1.0.

Table 3: Optimization and learning-rate schedule.

Hyperparameter	Value
Initialization	pretrained 2B video2world DiT
Optimizer	FusedAdamW
Peak learning rate	1×10^{-4}
Weight decay	1×10^{-3}
(β_1, β_2)	(0.9, 0.99)
ϵ	1×10^{-8}
LR schedule	linear warmup + linear decay
Warmup steps	1,000 (from $10^{-6} \cdot \text{lr}$ to peak)
Decay	to $0.3 \cdot \text{lr}$ over steps 1k–30k, then $0.06 \cdot \text{lr}$
Gradient clipping	global norm 1.0
Training iterations	40,000 for I2RT; $\sim 55,000$ for simulation
Mixed precision	bfloat16
Batch size (per-GPU / effective)	8 / 64
Parallelism	FSDP full shard (8 GPUs)

The three datasets differ in clip length, image resolution, and action/state representation. All actions, together with the proprioceptive states in the real-world YAM dataset, are normalized to $[-1, 1]$. For real-world YAM data, the normalization bounds are computed independently for each dimension using the 1st and 99th percentiles. RoboCasa and LIBERO instead use min–max normalization: the first three translation dimensions use symmetric bounds defined by $\pm \max |\cdot|$, while the remaining dimensions use their empirical minima and maxima. The real-world policy is conditioned on 14-dimensional proprioceptive states and predicts 14-dimensional bimanual actions. The six motion dimensions of each arm are represented as deltas relative to the initial state of the clip, while the gripper dimensions remain absolute, corresponding to the mask $(6, -1, 6, -1)$. RoboCasa and LIBERO predict 7-dimensional absolute actions without proprioceptive-state conditioning. For visual input, the real-world YAM dataset horizontally tiles three synchronized views (left, head, and right). RoboCasa arranges its three views in a 2×2 grid with the unused fourth tile left blank, whereas LIBERO horizontally tiles its two views.

Table 4: Dataset-specific data and dataloader settings.

Setting	I2RT Deployment	RoboCasa	LIBERO
Clip length (frames)	25	33	25
Resolution ($H \times W$)	128×384	256×256	128×256
Camera views	3 (tiled)	3 (tiled) + 1 (blank)	2 (tiled)
Frame rate	16 fps	16 fps	16 fps
Action dimension	14	7	7
State dimension	14	–	–
Delta-action targets	yes, mask $(6, -1, 6, -1)$	no	no
Action/state norm.	quantile (1/99 pct.)	min-max	min-max
Batch size (train / val)	8 / 4	8 / 4	8 / 4

The two auxiliary decoders receive features from disjoint sets of intermediate video-DiT blocks. Each decoder constructs a four-level feature pyramid with a finest resolution of $H/4 \times W/4$. Supervision is applied to latent frames at a temporal stride of 4, yielding $1 + (h - 1)/4$ supervised frames from each h -frame clip. The eight video blocks used across both decoders map one-to-one to the eight cross-attention layers in the action expert. The affordance decoder is trained using positive-class-weighted binary cross-entropy to address foreground–background imbalance, while the relative-depth

decoder combines masked L_1 reconstruction with an image-gradient loss for edge-aware geometric supervision. Both auxiliary losses are added to the video diffusion objective with unit weights.

Table 5: Auxiliary affordance and relative-depth decoder settings.

Hyperparameter	Affordance	Depth
Loss weight	1.0	1.0
Supervised video-DiT blocks	{6, 13, 20, 27}	{3, 10, 17, 24}
Pyramid resolutions	$H/4 \times W/4$, $H/8 \times W/8$, $H/16 \times W/16$, $H/32 \times W/32$	
# Supervised frames	$1 + \lfloor (h - 1)/4 \rfloor$ per h -frame clip	
Decoder channels	256	
Temporal-attn. blocks / heads	2 / 8	
Loss type	BCE with logits	Masked L_1 + masked grad.
Output-bias initialization	logit(0.05)	logit(0.15)
Target channels	1 (heatmap)	2 (depth + validity)
Target resolution	$H/4 \times W/4$	

Affordance targets are independently normalized by the maximum value in each frame, smoothed using a 3×3 Gaussian kernel with $\sigma=1.0$, and then downsampled to the decoder output resolution.

B Data Processing

We train DeVA on each benchmark with all the training data in a multi-task manner.

B.1 Training Data

LIBERO. We train on all four LIBERO suites: LIBERO-Spatial, LIBERO-Object, LIBERO-Goal, and LIBERO-Long. Each suite contains 10 tasks with 50 demonstrations per task, yielding 2,000 demonstrations in total before filtering. Because some demonstrations do not complete successfully when replayed, we use the filtered dataset introduced by Cosmos-Policy [6] to improve data quality and ensure a fair comparison with prior baselines. This dataset replays each demonstration in the simulator and retains only successful trajectories.

RoboCasa The RoboCasa training set contains 24 tasks, with more than 50 demonstrations available for each task. Following the LIBERO setup, we use the replay-filtered dataset introduced by Cosmos-Policy [6], which retains successful trajectories and contains 1,199 episodes in total.

I2RT YAM Deployment We evaluate real-world deployment on the I2RT YAM bimanual robot platform using demonstrations collected through teleoperation. The evaluation includes three tasks: (1) picking up a marker with one arm and handing it over to the other; (2) lifting a pot with both arms; and (3) picking up two bottles, one with each arm. We collect 30 demonstrations per task based on VR teleoperation [69], covering diverse initial object configurations. Each timestep contains three synchronized views: an external head view and two wrist-mounted end-effector views.

B.2 Depth Map Supervision

To provide the depth maps for auxiliary decoder supervision, we use an offline video depth prediction model [57] to generate relative-depth targets. The predictions are reversed and normalized to $[0, 1]$, such that nearby regions have values close to 0 and distant regions have values close to 1. To avoid boundary artifacts caused by the tiled multi-view input, we estimate depth independently for each camera view and then tile the resulting maps using the corresponding view layout. Because these targets serve as auxiliary supervision for distilling relative scene geometry into the intermediate video features, they do not require metric accuracy or exact scale consistency across views. This formulation is therefore robust to minor cross-view inconsistencies and temporal drift over long prediction horizons.

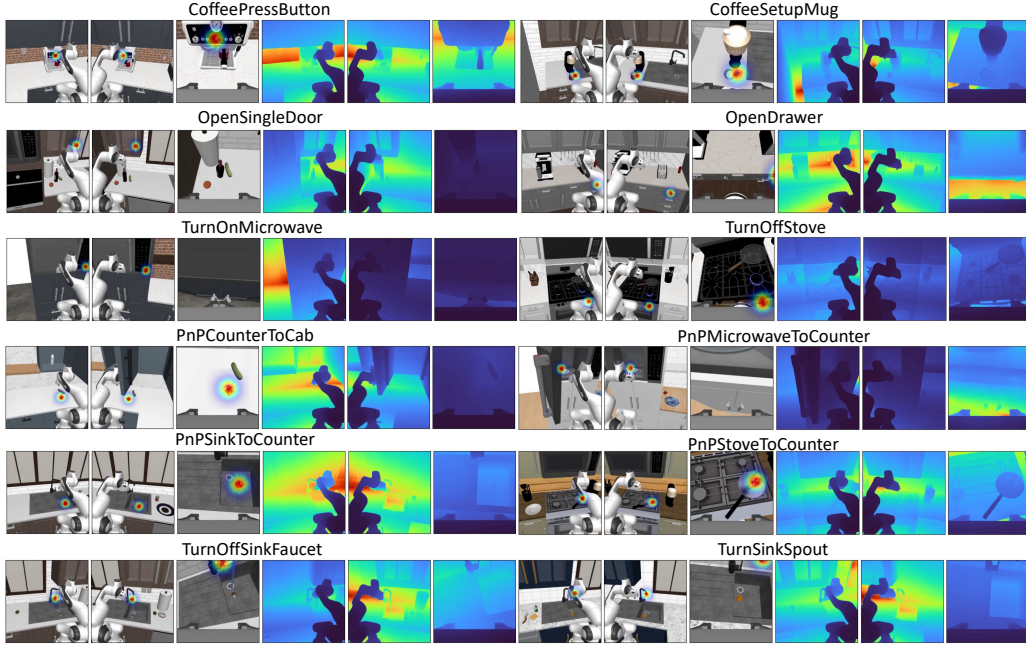


Figure 9: Visualization of physical guidance across representative manipulation examples. Affordance maps highlight task-relevant interaction regions, while relative-depth maps capture the near-to-far geometric structure of each scene.

B.3 Affordance Map Supervision

In addition to relative depth, we apply affordance-map supervision to intermediate video features. Because oracle interaction signals are available only in simulation, we use different annotation strategies for simulated and real-world data. For simulation benchmarks, we extract contact information from the RoboSuite MuJoCo simulator and project the contact points into the image plane. For frames containing valid contacts, the mean projected contact position defines the affordance center. For frames without direct contact, we use the first valid contact location in the corresponding trajectory, representing the region that the gripper is expected to approach. The resulting locations are converted into smooth Gaussian heatmaps for supervision.

For real-world demonstrations, where oracle contact information is unavailable, we use the task-conditioned affordance model [50] in an out-of-the-box manner to generate affordance pseudo-labels. The pretrained model distill affordance knowledge from vision and vision-language foundation models into a lightweight prediction model, allowing efficient annotation without task-specific manual labels. In practice, it provides usable affordance targets for most of our real-world manipulation scenes.

As with relative-depth supervision, affordance maps provide an auxiliary signal that encourages the video features to emphasize task-relevant interaction regions rather than serving as exact spatial targets. The supervision is therefore tolerant to moderate noise and localization errors in the generated pseudo-labels.

B.4 Guidance Visualization

Figure 9 presents representative visualizations of the physical guidance used to supervise the intermediate video-DiT features and condition the action expert. The task-conditioned affordance maps highlight interaction regions relevant to the language instruction, such as target objects and expected end-effector contact locations. The relative-depth maps provide complementary geometric information by capturing object boundaries, free space, and near-to-far scene structure. Although the supervision may contain moderate prediction noise, both signals remain spatially coherent across

diverse manipulation scenes. These examples illustrate how affordance and relative depth provide complementary semantic and geometric guidance for video-to-action transfer.

C Real-World Deployment

C.1 Task Details

We evaluate three real-world bimanual manipulation tasks:

- **Handover Marker:** A blue, black, or red marker is initially placed on either side of the workspace. The nearest arm must grasp the marker, transfer it to the opposite gripper, and place it safely on the table. The primary challenge is the handover stage, which requires precise alignment and coordinated gripper timing to prevent the marker from being dropped.
- **Lift Pot:** Both grippers approach the corresponding handle regions and align approximately parallel to the handles. They must close synchronously, lift the pot to the target height while maintaining a stable orientation, hold it securely, and return it steadily to the table. This task requires robustness to variations in pot orientation, coordination between both arms, synchronized grasping, and sufficient grip force throughout the motion.
- **Pick Up Diverse Bottles:** The task includes four bottle instances, from which two are randomly selected and placed on the table for each episode. Each arm must approach and align its gripper with one bottle, grasp both bottles simultaneously, lift them, and place them stably back on the table. The policy must coordinate both arms across diverse bottle geometries and distinguish visually similar stages of the trajectory, particularly between lifting and placing.

Figures 10, 11, and 12 illustrate the keyframes required for successful completion of all three real-world tasks.

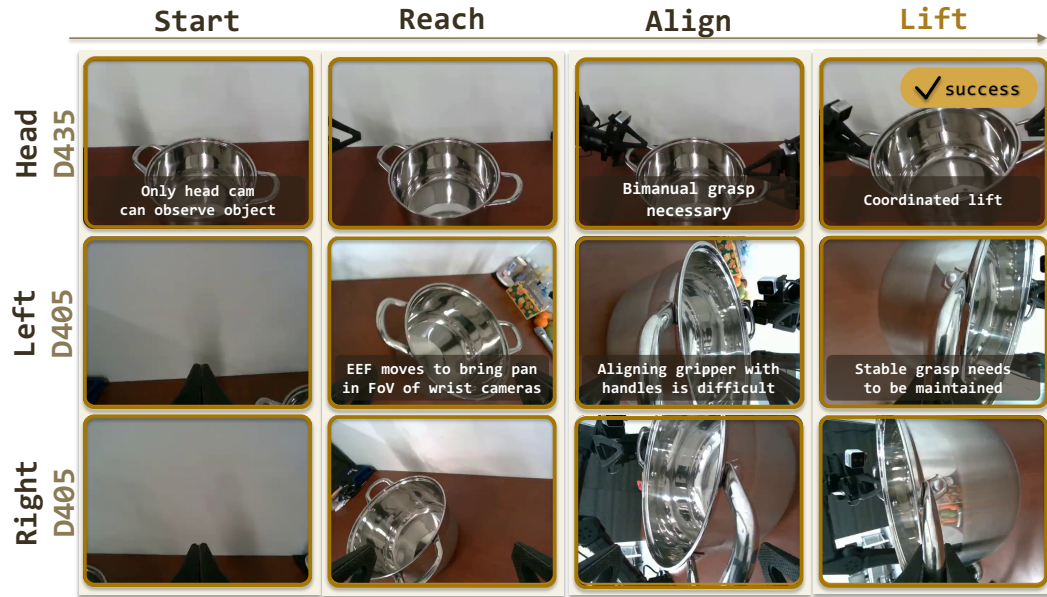


Figure 10: Keyframes of the Lift Pot task, including coordinated approach, synchronized grasping, stable lifting, and placement.

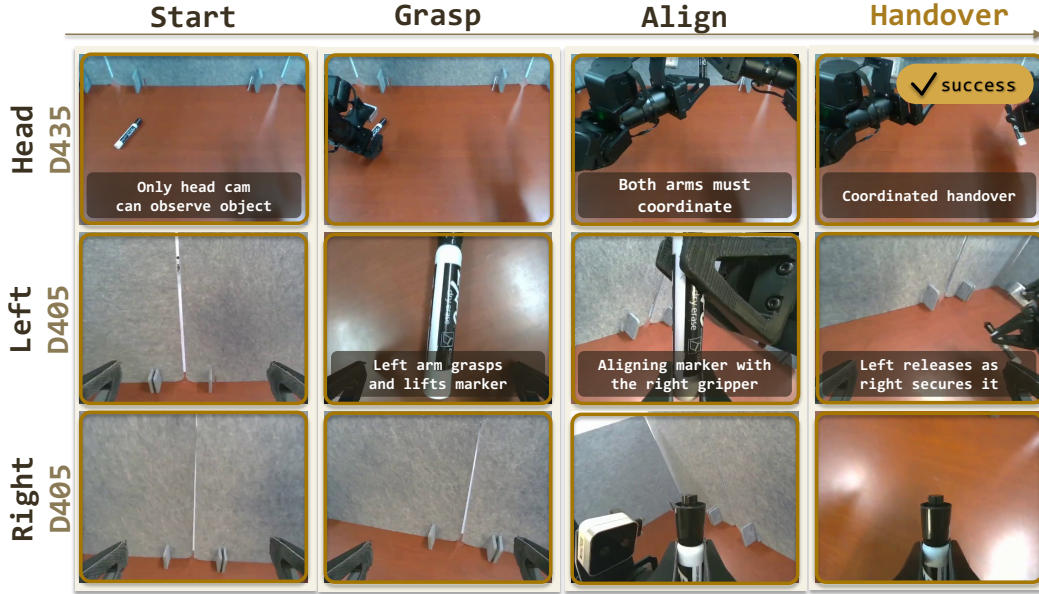


Figure 11: Keyframes of the Handover Marker task, highlighting grasping, precise bimanual transfer, and stable placement.

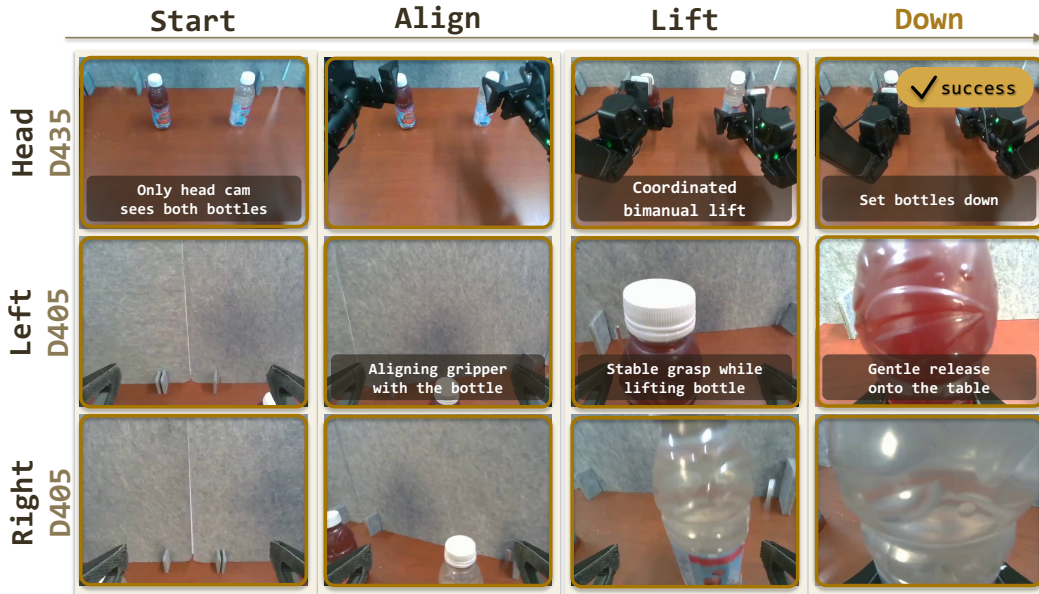


Figure 12: Keyframes of the Pick Up Diverse Bottles task, including coordinated dual-arm grasping, lifting, and stable placement.

C.2 Evaluation Score Criteria

For more detailed analysis of rollout progress, we assign each episode a stage-completion partial credit in $[0, 1]$. The score corresponds to the most advanced criterion completed during the rollout; episodes that do not satisfy the first criterion receive a score of 0.

Handover Marker.

- **0.5:** One gripper lifts the marker, but the transfer to the other gripper fails.

Success Rate	Handover Marker	Lift Pot	Pick Up Bottles	Average
GR00T-n1.6 [22]	<u>0.88</u>	0.45	0.10	0.48
$\pi_{0.5}$ [3]	0.70	<u>0.55</u>	0.75	<u>0.67</u>
Cosmos-Policy [6]	0.45	<u>0.23</u>	0.35	0.34
Ours	0.90	0.68	<u>0.65</u>	0.74

Table 6: Real-world deployment comparison.

- **1.0:** The receiving gripper successfully secures the marker and completes the handover.

Lift Pot.

- **0.25:** One gripper lifts its handle, while the other gripper fails to establish a successful grasp.
- **0.5:** Both grippers successfully lift the pot, but the pot is not returned to the table.
- **0.75:** Both grippers lift the pot, but only one gripper successfully completes the placement stage.
- **1.0:** Both grippers coordinate to lift the pot and return it stably to the table.

Pick Up Diverse Bottles.

- **0.25:** One gripper lifts one bottle, but does not return it to the table.
- **0.5:** Both bottles are lifted but not returned to the table, or one arm completes the full pick-and-place sequence while the other arm remains inactive.
- **0.75:** Both bottles are successfully lifted, but only one is returned stably to the table.
- **1.0:** Both arms successfully lift their assigned bottles and return them stably to the table.

C.3 Additional Baseline Comparisons

Baselines DeVA is compared with Cosmos-Policy, the most closely related Video-Action Model, and two other pretrained VLA baselines after task-specific finetuning:

- **Cosmos-Policy:** As the most closely related Video-Action Model, Cosmos-Policy adapts a pretrained Cosmos video generative backbone for robot control. We initialize it from the pretrained video checkpoint and train the action modality from scratch on the same task demonstrations used for all real-world comparisons. The model is trained for 40k steps using its official training pipeline and configuration, without modifying its architecture or objective.
- $\pi_{0.5}$: We initialize $\pi_{0.5}$ from its released pretrained checkpoint and fine-tune it on the same task demonstrations used for Cosmos-Policy and DeVA. The model is trained for 40k steps using the official training pipeline and model-specific configuration, without modifying its architecture or training objective. The pretrained model has been pretrained on a large and diverse corpus of robot demonstrations.
- **GR00T-N1.6:** We similarly initialize GR00T-N1.6 from its released pretrained checkpoint and fine-tune it on the same task demonstrations for 40k steps. We retain the official training pipeline and configuration provided with the model.

This setup matches the training data and number of optimization steps across methods while preserving the model-specific training procedures recommended by their respective authors.

Analysis We evaluate DeVA and the three baselines on all three tasks, using 10 rollouts for each method-task pair. The quantitative results are reported in Table 6. Qualitative inspection further reveals several recurring failure modes.

VLA models: The two VLA baselines exhibit different behaviors across tasks. Notably, $\pi_{0.5}$ is initialized from a checkpoint with extensive robot-centric pretraining on large-scale and diverse robot trajectories, providing strong manipulation priors before task-specific fine-tuning. It generalizes well to variations in initial object placement, but still encounters coordination failures. During Handover Marker, the transferring gripper sometimes releases the marker before the receiving gripper establishes a stable grasp, causing the marker to fall. During Lift Pot, performance is sensitive to changes in the pot and handle orientation relative to the demonstrations. In some rollouts, the policy successfully lifts the pot but fails to initiate the placement stage, leaving it suspended. GR00T-N1.6 exhibits similar pose sensitivity on Lift Pot. On Pick Up Diverse Bottles, it frequently approaches the bottles but does not close the grippers or initiate lifting. One possible explanation is ambiguity between visually similar stages before grasping and after placement, which may cause premature task completion.

VAM models: Under the same 40K-step fine-tuning budget, Cosmos-Policy adapts more slowly than all others and remains less robust to variations in object placement. During Handover Marker, it often reaches toward a fixed workspace region rather than adapting precisely to the marker location. Similar sensitivity to object pose is observed in the pot and bottle tasks. In contrast, DeVA more consistently adjusts its trajectories to these variations.

Among all methods, $\pi_{0.5}$ performs particularly well on Pick Up Diverse Bottles. We hypothesize that this advantage is partly attributable to its extensive robot-centric pretraining and its longer action horizon of 50 steps, compared with 25 steps for the other methods, which may help maintain coherent actions across the grasp, lift, and placement stages.

C.4 Robustness Analysis for DeVA

We conduct qualitative stress tests along three axes: (1) unseen object instances, (2) distractors and occlusions, and (3) out-of-distribution layouts. We use Handover Marker as a representative task. For unseen instances, we progressively vary the target object’s color, thickness, and category, including shifts from markers to ballpoint pens and bottles. For distractor evaluation, we place a similarly elongated object, such as a banana, near the target marker or introduce a can that partially occludes the target from different camera views. Finally, we evaluate challenging layouts with object placements substantially different from those observed during training. Representative examples are shown in Figures 13a, 13b, and 13c.

Qualitatively, DeVA remains effective under changes in object appearance and retains successful behavior for several category-level variations. It generally identifies the instructed target when distractors are spatially separated and remains robust in more challenging cases involving adjacent, shape-similar, or partially occluding objects. The model also completes several tasks under substantially shifted object layouts. Notably, these behaviors emerge from only 30 task-specific demonstrations without large-scale robot-action pretraining, suggesting that DeVA can adapt its pretrained video representations to previously unseen real-world configurations.

D More Ablation Study

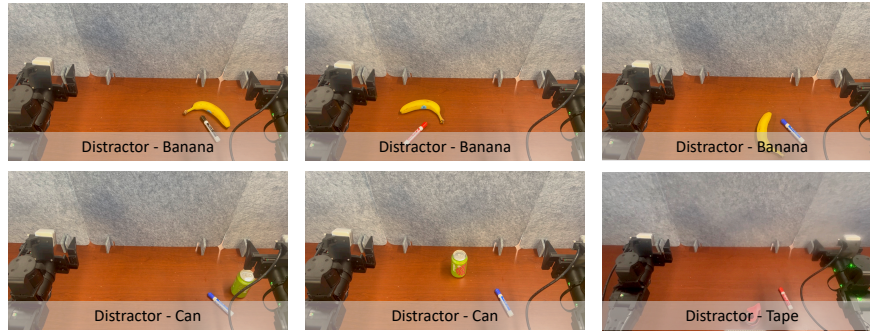
We report additional ablations conducted during model development, focusing on the feature-interaction mechanism, observation-conditioning sources, and the use of representations from multiple video denoising steps. These studies complement the component ablations presented in the main paper.

D.1 Feature-Interaction Mechanism

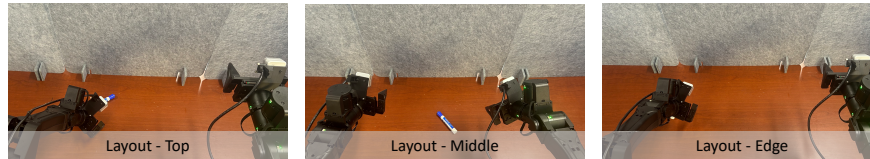
We compare several mechanisms for transferring video representations to the action expert, including our cross-attention with bridge tokens, self-attention-based fusion, and cross-attention augmented with additional transformation layers. As shown in Table 7, most direct interaction mechanisms achieve comparable performance, indicating that predictive video representations provide useful



(a) Instance axis.



(b) Distraction axis.



(c) Layout axis.

Figure 13: Robust analysis examples along three axes: instance, distraction, and layout.

conditioning through multiple transfer interfaces. However, introducing additional transformation layers degrades performance, suggesting that excessive processing may weaken the information already encoded by the pretrained video backbone. We therefore adopt cross-attention with bridge tokens as a compact interface that supports both layer-wise feature transfer and global information aggregation.

D.2 Conditioning Sources

We further examine whether the action expert benefits from a separately encoded initial observation in addition to the multi-level video features and language embeddings. This ablation is conducted on

Model	Success Rate (%)
Decoupled Branches with direct self-attention in action head	66.79
Decoupled Branches with direct cross-attention	66.50
Decoupled Branches with specific layers for each cross attention	63.75
Decoupled Branches with aggregated attention before cross attention	62.75

Table 7: Ablation of video-to-action interaction mechanisms on RoboCasa. Direct self-attention and cross-attention perform similarly, while additional transformation or aggregation layers reduce performance.

RoboCasa using the base model. As reported in Table 8, explicitly adding the initial-frame features does not provide a consistent improvement. Because the video backbone is already conditioned on the initial observation, its intermediate representations retain the corresponding visual context. A separate initial-frame pathway therefore introduces largely redundant information.

Model	Success Rate (%)
Decoupled Branches (T5 + Initial Observation + Video for Action)	66.75
Decoupled Branches (Only T5 + Video for Action)	66.50

Table 8: Ablation of conditioning sources for the action expert on RoboCasa. Adding a separately encoded initial observation yields only a marginal 0.25-point improvement, suggesting that its visual information is already captured by the video features.

D.3 Denoising-Step Conditioning

We additionally compare conditioning the action expert on features from only the final video denoising step with aggregating features across multiple denoising steps. Although the final-step representation corresponds to the least corrupted video latent, it may be more specialized toward video reconstruction and does not necessarily provide the most informative representation for action prediction. As shown in Table 9, final-step conditioning achieves a success rate of 49.6%, while multi-step conditioning improves it to 53.2%. This 3.6-point gain suggests that intermediate denoising representations provide complementary information to the action expert. Moreover, freezing the video backbone under final-step conditioning reduces performance to 22.4%, demonstrating that joint adaptation of the video representations is more important than simply extracting features from the pretrained backbone.

Video Conditioning Strategy	Success Rate (%)
Final denoising step	49.6
Final denoising step + frozen video backbone	22.4
Multi-step denoising features	53.2

Table 9: Ablation of video denoising-step conditioning on RoboCasa. Multi-step conditioning improves over final-step conditioning by 3.6 points, while freezing the video backbone substantially degrades performance. All results are obtained from 30K-step checkpoints.

E Reproducibility Statement

To facilitate reproducibility, we release the training and evaluation code, model configurations, and benchmark scripts. We provide detailed implementation specifications covering data preprocessing, action representations, model architecture, training schedules, hyperparameters, and pretrained checkpoint initialization. For simulation experiments, we follow the standard protocols of LIBERO, LIBERO-Plus, and RoboCasa and report results using consistent numbers of evaluation trials and random seeds when applicable. For real-world experiments, we document the robot platform, task definitions, observation modalities, control frequency, and evaluation criteria.