UNIACTION: UNIFIED ACTION CONDITIONING FOR WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Actions that condition world models are highly heterogeneous, and thus existing action-conditioned world models are typically built for individual action types, fragmenting already limited action-annotated data across separate models. We introduce UniAction, a unified action-conditioning framework that represents any action as a sparse set of action elements, each with a discrete semantic identity and a continuous spatiotemporal grounding coordinate. By mapping these coordinates directly into the pretrained video model's native positional space, UniAction allows heterogeneous actions to condition a shared world model within the same spatiotemporal structure. We further introduce cardinality-calibrated attention to address the bias from differing action representation granularities. We evaluate UniAction on four representative action forms: force, camera control, hand pose, and robot pose. UniAction outperforms conditioning alternatives in action following, and a single jointly trained model surpasses the action-specific experts, showing that heterogeneous action data benefit from unified model and joint training. Anonymous project page: **`https://uniaction.pages.dev/`**

1 INTRODUCTION

A world model predicts how the world evolves in response to actions (Parker-Holder & Fruchter, 2025; Robbyant Team, 2026), and large-scale video generation models (MiniMax, 2026; Team Wan, 2025), which learn rich spatiotemporal dynamics from real-world videos, provide a natural foundation for building the world models (Brooks et al., 2024). The actions that condition such world models, however, are highly heterogeneous. They range from global viewpoint changes (Parker-Holder & Fruchter, 2025; Li et al., 2025; Tang et al., 2025; Sun et al., 2025; He et al., 2025) and physical forces applied to objects (Gillman et al., 2025; 2026; Wang et al., 2025a; 2026) to hand and robot motions (Sudhakar et al., 2024; Xie et al., 2026; Wang et al., 2025b; Zhu et al., 2025), and differ greatly in their semantics and spatiotemporal structures.

Given this heterogeneity, most existing action-conditioned world models are built for a single action type, with a conditioning mechanism designed for that type and training restricted to its data (Li et al., 2025; Sun et al., 2025; He et al., 2025; Gillman et al., 2025; Wang et al., 2026; Sudhakar et al., 2024; Zhu et al., 2025). This practice is costly because action-annotated data is scarce: compared with large-scale generic video, it is limited in scale and concentrated in a narrow range of scenes and interaction patterns (Gillman et al., 2025). Training separate models further fragments this data, and since each model learns how the world responds to only one type of action, knowledge of scene structure, object properties, and world dynamics that could be shared remains isolated across models (Fig. 1, left). As the number of action types grows, such separate training scales poorly.

To bring data from different actions into a single world model, we first need a conditioning representation that can accommodate their heterogeneous forms in a unified way. Existing generic mechanisms offer two natural choices. One renders actions into image or video conditions (Wang et al., 2025b; Wu & Gao, 2026), providing direct spatial correspondence with the generated content, but not every action has a visual form that can be directly rendered. The other encodes different actions into conditioning tokens (Li & Torralba, 2025; NVIDIA, 2025), which flexibly accommodates signals of different forms, but where and when an action takes effect in the generated video remains an implicit correspondence. These limitations lead us to reconsider what information heterogeneous actions actually need to convey to a world model.

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.

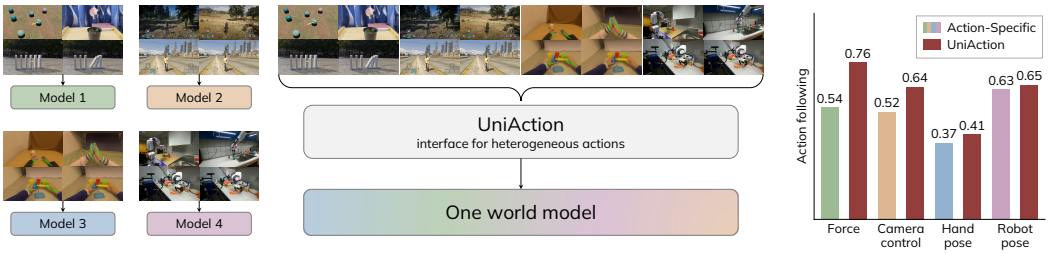


Figure 1: **One world model for heterogeneous actions.** Instead of training a separate model per action type (left), UniAction conditions a single world model (middle) on heterogeneous actions and outperforms the action-specific experts across all four action forms (right).

Our key observation is that, despite their heterogeneous raw forms, every action conveys two complementary pieces of information to the world model: what the action is, and where and when it takes effect. Based on this observation, we introduce UniAction, which represents each action as a sparse set of action elements. Each element describes its semantics with a discrete identity and its spatiotemporal grounding with a continuous coordinate. We place these coordinates directly in the pretrained video model's native positional geometry, allowing action and video to interact within a shared spatiotemporal structure while reusing the model's pretrained spatiotemporal modeling capabilities. Heterogeneous actions can thus condition a single model through the same mechanism without requiring a common raw representation or rendering into visual conditions. Since different actions may be represented by different numbers of elements, we further introduce cardinality-calibrated attention to prevent representation granularity from biasing conditioning strength.

We instantiate UniAction on four representative action forms: force, camera control, hand pose, and robot pose. Under matched training settings, UniAction outperforms alternative conditioning mechanisms in both action following and visual quality. More importantly, a single jointly trained UniAction model consistently outperforms the corresponding action-specific experts across all four action forms (Fig. 1, right), demonstrating the effectiveness of unified action conditioning and the benefit of jointly learning from heterogeneous action data. Our contributions are threefold:

- We introduce UniAction, a unified action-conditioning framework that represents heterogeneous actions through action semantics and spatiotemporal grounding, placing action coordinates in the pretrained video model's native positional geometry so that all actions condition the model through the same mechanism.
- We introduce cardinality-calibrated attention to mitigate the bias caused by varying numbers of action tokens, reducing the influence of representation granularity on conditioning strength.
- We evaluate UniAction across four representative action forms. UniAction achieves better action following and visual quality than visual- and token-based conditioning alternatives, while a single jointly trained model outperforms the corresponding action-specific experts, showing the benefit of jointly learning from heterogeneous action data.

2 Related Work

Video World Models. Large-scale video generation models, such as Wan (Team Wan, 2025) and MiniMax (MiniMax, 2026), learn visual structures and spatiotemporal dynamics through video pretraining, providing a generative foundation for predicting how the world evolves from current observations. Building on these powerful video models, a line of works (Parker-Holder & Fruchter, 2025; Li et al., 2025; He et al., 2025; Sun et al., 2025; Gao et al., 2026) further introduce user interactions, enabling the models to continuously generate future observations of an environment under the given controls. These video world models typically define their action spaces and control interfaces for a specific environment; for example, the keyboard inputs and camera movement signals (Li et al., 2025; He et al., 2025). In contrast, UniAction builds on the same class of pretrained video models and allows actions from different environments with different structures to condition a single video world model through a shared interface.

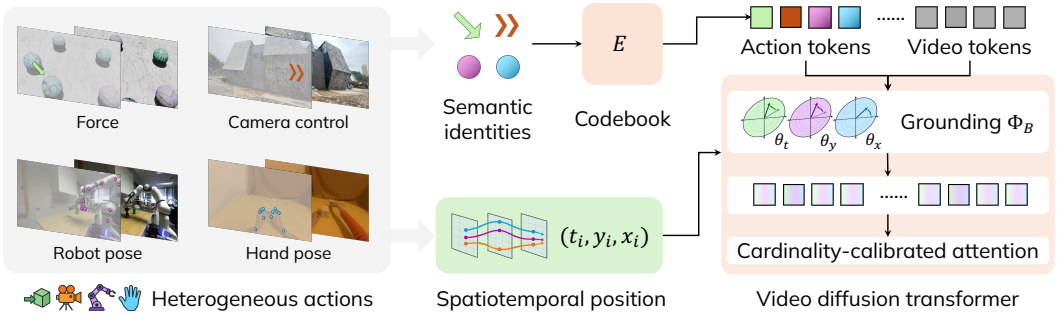


Figure 2: **Overview of UniAction.** Diverse action forms are represented as sparse action elements: semantic identities index tokens in a shared codebook (Sec. 3.1), spatiotemporal positions use the backbone's native positional encoding (Sec. 3.2), and action and video tokens interact via cardinality-calibrated attention (Sec. 3.3).

Action Conditioning in Video Models. Existing action-conditioned video models largely focus on specific action spaces. One line of work studies game commands, camera motion, and navigation controls (Li et al., 2025; He et al., 2025; Sun et al., 2025; Robbyant Team, 2026; Gao et al., 2026), while another models force-based physical interventions (Gillman et al., 2025; 2026; Wang et al., 2025a; 2026) with the environment. Other approaches condition video models on hand motion (Sudhakar et al., 2024) or robot manipulation trajectories (Zhu et al., 2025). These methods typically use representations and conditioning mechanisms tailored to their respective action spaces.

Recent work has begun to explore unifying related action spaces. Shared skeletal or kinematic representations connect human and robot motion (Wang et al., 2025b; Wu & Gao, 2026), while the recent Cosmos 3 (NVIDIA, 2026a) represents controls across embodied systems through common physical variables. These approaches build on shared kinematic structures or common physical-control parameterizations. In contrast, UniAction accommodates heterogeneous actions with distinct structures and parameterizations through a shared formulation of action semantics and spatiotemporal grounding, enabling joint learning from diverse action-conditioned data within a single world model.

Controllable Video Generation. Controllable video generation typically achieves unified control by converting external signals into shared conditioning representations compatible with the video model. Several approaches represent controls as visual conditions aligned with the generated video (Wang et al., 2023; Zhang et al., 2024; Jiang et al., 2025). However, this requires a suitable visual encoding for each control signal, which is less natural for abstract interventions such as forces. Alternatively, separate learned encoders can map different sensory inputs and actions into features that are fused and injected through a cross-attention mechanism (Li & Torralba, 2025). These approaches accommodate heterogeneous inputs through either visual renderings or learned encoders for specific action formats. In contrast, UniAction directly represents actions following the same conditioning formulation at the representation level, without requiring a separate renderer or learned encoder for each action type.

3 Method

Given an initial frame I_0, a text prompt T, and an action A, we model the conditional generation of a future video V as

$$p_\theta(V \mid I_0, T, A). \tag{1}$$

UniAction represents any action as a sparse set of action elements:

$$A = \{(s_i, \mathbf{p}_i)\}_{i=1}^{N}, \qquad \mathbf{p}_i = (t_i, y_i, x_i), \tag{2}$$

where s_i denotes the semantic identity of the i-th element and $\mathbf{p}_i$ its continuous grounding coordinate in video space-time. Different actions may use different semantic identities and numbers of elements, but all follow this unified representation and condition a shared video model through the same mechanism.

3.1 ACTION SEMANTICS

Action semantics describe the meaning represented by each action element. We define a shared discrete semantic vocabulary $\mathcal{S}$ that distinguishes semantic identities both within and across action forms. A semantic codebook $E \in \mathbb{R}^{|\mathcal{S}|\times d}$ associates each identity with a d-dimensional action token, where d matches the feature dimension of the video tokens. For the i-th action element, its semantic identity $s_i \in \mathcal{S}$ directly indexes the corresponding action token: $\mathbf{a}_i = E[s_i] \in \mathbb{R}^d$.

Each semantic identity is associated with an independently initialized action token, and the model learns its correspondence to scene evolution through joint action-conditioned training. Elements sharing the same identity reuse the same action token across different locations and times, while these spatiotemporal differences are modeled separately by the spatiotemporal grounding described in the next section. New semantic identities can be added directly to the same vocabulary, allowing the representation to naturally extend to new action forms.

3.2 SPATIOTEMPORAL GROUNDING

Spatiotemporal grounding specifies where and when an action element takes effect. Although heterogeneous actions may be defined in different coordinate and control spaces, their effects ultimately manifest in the spatiotemporal evolution of the generated video. We therefore use the video as a common spatiotemporal reference and represent the grounding of the i-th action element with a continuous coordinate $\mathbf{p}_i = (t_i, y_i, x_i)$.

To align action grounding with the spatiotemporal structure established by the pretrained video model, UniAction directly reuses the backbone's native positional system. Let $\Phi_{\mathcal{B}}(\mathbf{x}, \mathbf{p})$ denote the backbone's positional mechanism for a token $\mathbf{x}$ at coordinate $\mathbf{p}$. For an action token $\mathbf{a}_i$ and a video token $\mathbf{v}_j$,

$$\widetilde{\mathbf{a}}_i = \Phi_{\mathcal{B}}(\mathbf{a}_i, \mathbf{p}_i), \qquad \widetilde{\mathbf{v}}_j = \Phi_{\mathcal{B}}(\mathbf{v}_j, \mathbf{p}_j^v), \tag{3}$$

where $\mathbf{p}_j^v$ denotes the spatiotemporal coordinate of the video token. Action and video tokens thus establish their spatiotemporal relations within the same positional geometry and interact through the backbone's native self-attention (Vaswani et al., 2017). Since this positional geometry is already established during video pretraining, action–video interactions can directly reuse the backbone's pretrained spatiotemporal modeling capabilities.

3.3 CARDINALITY-CALIBRATED ATTENTION

Different action forms can be represented at different granularities, leading to potentially large differences in the number of action tokens. Since action and video tokens are jointly processed by the backbone's self-attention, we examine how token cardinality affects attention allocation. For a video query, let ℓ_i^a denote its attention logit to the i-th action token. For an action represented by N tokens, its total unnormalized attention mass is

$$Z_a = \sum_{i=1}^{N} \exp(\ell_i^a) = N \cdot \mathrm{mean}_i[\exp(\ell_i^a)]. \tag{4}$$

The factor N is determined solely by representation granularity and is independent of the average action–video compatibility, thereby introducing an additional cardinality bias. To remove this bias, we apply the same calibration term to all action logits:

$$\hat{\ell}_i^a = \ell_i^a + \log \frac{N_{\mathrm{ref}}}{N}, \tag{5}$$

where N_{ref} is a fixed reference cardinality shared across all actions, while the video logits remain unchanged. After calibration, the total action mass becomes $N_{\mathrm{ref}} \cdot \mathrm{mean}_i[\exp(\ell_i^a)]$, removing its dependence on the actual number of action tokens and placing different actions on a common attention scale. Since the same offset is added to all action logits, the relative attention relationships among action tokens remain unchanged. We detail the mathematical proof in Supplementary B.

4 EXPERIMENTS

We first describe how UniAction is instantiated in a pretrained video backbone (Sec. 4.1) and on four representative diverse action forms (Sec. 4.2), and then present the experimental setup (Sec. 4.3), and the results on comparisons and ablations in the following sections.

4.1 BACKBONE INSTANTIATION

Our model backbone, Wan2.2 (Team Wan, 2025), is a latent video transformer with 3D RoPE, into which UniAction is instantiated directly through its existing token sequence and self-attention. The semantic identity s_i of each action element indexes an action token $\mathbf{a}_i$ as in Sec. 3.1, and all action tokens are concatenated after the latent video token sequence: $[\,\mathbf{v}_1, \ldots, \mathbf{v}_M,\ \mathbf{a}_1, \ldots, \mathbf{a}_N\,]$.

The action location $\mathbf{p}_i$ in video space is mapped into the backbone's latent coordinate system through its own VAE (Kingma & Welling, 2014) compression and tokenization,

$$\mathbf{p}_i^{\mathrm{lat}} = g_{\mathcal{B}}\left(\mathbf{p}_i\right), \tag{6}$$

and in self-attention the same rotary positional encoding (Su et al., 2024) is applied to action tokens at $\mathbf{p}_i^{\mathrm{lat}}$ and to video tokens at their latent grid coordinates $\mathbf{p}_j^{\mathrm{v}}$.

Consequently, for two action elements that share a semantic identity but occur at different locations,

$$s_i = s_j \implies \mathbf{a}_i = \mathbf{a}_j = E[s_i], \qquad \mathbf{p}_i^{\mathrm{lat}} \neq \mathbf{p}_j^{\mathrm{lat}}, \tag{7}$$

so a semantic identity can be reused across locations, while where it takes effect is distinguished entirely by the backbone's native positional geometry.

4.2 ACTION INSTANTIATION

The four action types, force, camera control, hand pose, and robot pose, differ substantially in their raw forms: some are instantaneous physical vectors, some are discrete commands with no intrinsic spatial location, and some are dense per-frame, per-joint annotations.

Hand pose. A hand pose is a dense per-frame, per-joint action, and the image location of each joint is itself where the intervention takes effect. The semantic identities are joint indices, one per joint; at each frame, every visible joint yields one action element located at that joint's position in the frame.

Robot pose. A robot pose is given by a mechanical structure of rigid links, whose number of joints and geometry are determined by the particular robot. The semantic identities are robot joint indices; at each frame, every visible joint yields one action element at that joint's position in the frame.

Camera control. We consider keyboard-based camera control, where viewpoint changes are specified by discrete key commands. The semantic identities correspond to eight camera commands: four translations (forward, backward, left, right; keys W, S, A, D) and four rotations (pitch up, pitch down, yaw left, yaw right; keys I, K, J, L). Since these commands act on the global view rather than at a specific location, they have no intrinsic spatial grounding. We therefore assign each direction a fixed canonical location on the corresponding side of the frame; translation and rotation commands with the same direction share this location and are distinguished by their semantic identities. At each frame, every active command yields one action element at its canonical location and the current time.

Force. A force is an instantaneous vector applied at a point, and its direction and magnitude cannot be expressed by a single location. We represent one applied force with two elements at the same instant, one at the tail and one at the head of the force vector, with the semantic identities marking these two roles. The spatial offset between them is the force vector.

4.3 EXPERIMENTAL SETUP

We summarize the data, implementation, and evaluation protocol below; full configurations and details are given in Supplementary C.

Table 1: Comparison with visual-based (VACE (Jiang et al., 2025)) and token-based (Cross-Attn (NVIDIA, 2026b)) conditioning mechanisms across four action forms. Best and second-best results are **bold** and underlined.

Action Form	Method	Automatic Evaluation		Human Evaluation	
		Action Following ↑	LPIPS ↓	Action Following ↑	Visual Quality ↑
Force	VACE	0.6046	0.1178	2.47	3.40
	Cross Attention	0.1618	0.1341	1.96	3.13
	UniAction (Ours)	**0.7630**	**0.1163**	**4.37**	**4.00**
Camera Control	VACE	0.4997	0.5100	3.30	3.55
	Cross Attention	0.5306	0.5124	3.13	3.35
	UniAction (Ours)	**0.6371**	**0.4863**	**4.02**	**3.68**
Hand Pose	VACE	0.2938	0.1122	3.30	**3.43**
	Cross Attention	0.0678	0.1695	2.34	3.12
	UniAction (Ours)	**0.4119**	**0.0910**	**3.36**	3.42
Robot Pose	VACE	0.5245	0.1389	3.60	3.73
	Cross Attention	0.0859	0.1859	2.07	3.44
	UniAction (Ours)	**0.6521**	**0.1164**	**4.32**	**3.99**

Datasets. The training data for force, camera control, hand pose, and robot pose are drawn from GoalForce (Gillman et al., 2026), SolarWM (Huang et al., 2026), EgoDex (Hoque et al., 2026), and DROID (DROID Dataset Team, 2024), respectively. We sample 10K clips from each dataset, 40K in total, each with 81 frames at 480×832 resolution.

Implementation details. Our main experiments use the high-noise expert of Wan2.2-I2V-A14B (Team Wan, 2025) as the generative backbone and adapt it with LoRA (Hu et al., 2022).

Evaluation. We primarily evaluate *Action Following*, i.e., whether the generated video evolves as specified by the given action. For each test sample, we fix the initial frame and text prompt while varying only the action condition, such as rotated or reversed forces, different camera directions, and flipped or stationary pose trajectories. This yields 410 counterfactual generations that isolate the model's response to the specified intervention. We then measure each response with an off-the-shelf estimator: DIS optical flow (Kroeger et al., 2016) to estimate object motion in force scenes, CoTracker3 (Karaev et al., 2025) to track hand keypoints for hand pose and control points for robot pose, and ViPE (Huang et al., 2025) to estimate camera pose trajectories for camera control. From these motion measurements, we compute an Action Following score for each action form and normalize all scores to $[0, 1]$, where higher values indicate a more accurate response to the given action. Full metric definitions and normalization are provided in the supplementary material. We additionally report LPIPS (Zhang et al., 2018) on original condition's generation as a complementary measure of perceptual similarity to the ground truth frames.

In addition, we conduct a human evaluation on 60 cases sampled uniformly across the four action forms, with 100 participants. In each case, the videos generated by the three methods are presented together, with method names hidden and their order randomized. Participants rate each video on a 1–5 scale along two dimensions: *Action Following*, i.e., whether the video performs the given action, and *Visual Quality*, covering clarity, stability, and the plausibility of physical interactions. We report the average scores; the full evaluation protocol is provided in Supplementary C.

4.4 Comparison with Generic Conditioning Mechanisms

We compare UniAction with two generic conditioning mechanisms. All three methods start from the same pretrained model and are trained jointly on the same 40K samples of four action types with matched optimization settings. They differ only in how actions are represented and injected into the video model, so the comparison isolates the conditioning mechanism itself.

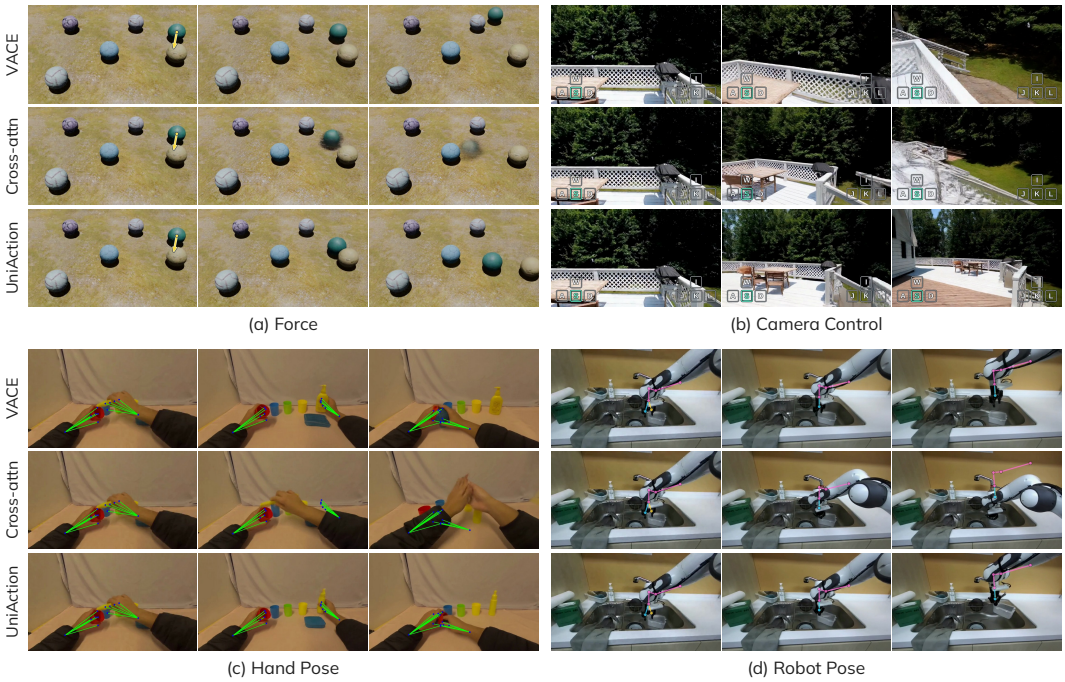


Figure 3: **Qualitative comparison across four action forms.** UniAction follows the input actions more faithfully than other conditioning mechanisms. Overlays indicate the input action.

Visual Conditioning (VACE). We adopt a VACE (Jiang et al., 2025) style visual conditioning baseline that renders different actions as control videos, encodes them with the VAE, and feeds them into the VACE conditioning branch. Force is represented by an application point and a directional arrow, camera control by temporally highlighted key indicators, and hand and robot poses by their corresponding skeletons. The VACE conditioning branch consists of eight replicated video DiT blocks, with approximately 3.1B trainable parameters.

Token Conditioning (Cross-Attn). We adopt a cross-attention conditioning baseline (NVIDIA, 2026b). We use separate encoders for different action forms, to directly map action signals and their spatiotemporal information into conditioning tokens of a shared dimensionality, which are then injected into the video model through cross-attention layer. The action encoders and cross-attention modules are both trained, with LoRA applied to the cross-attention modules, resulting in approximately 640M trainable parameters.

UniAction. All action forms share the same conditioning pathway, where action tokens interact with video tokens through the backbone's native self-attention. UniAction trains a shared action-branch LoRA that is applied only to action tokens, with approximately 404M trainable parameters.

As shown in Table 1, UniAction achieves the best action following performance across all four action forms, together with the lowest LPIPS. Human evaluation leads to a consistent conclusion: UniAction receives the highest action following ratings on all action forms, along with comparable or better overall visual quality. With the spatiotemporal information in a learned encoder space rather than the backbone's positional geometry, Cross-Attn performs markedly worse on actions that require spatial localization like robot and hand pose. VACE is competitive where an action has a natural visual rendering, but with the largest gap to UniAction on force and camera control, which lack a native visual form. In contrast, UniAction responds effectively to both types of actions through semantic identities and explicit spatiotemporal coordinates. Figure 3 presents qualitative comparisons across the four action forms, more results in Supplementary E.

4.5 Benefits of Unified Training

We next examine whether heterogeneous action data benefit one another when trained within a single model. To test this, we train an action-specific expert baseline for each action form. Each expert has

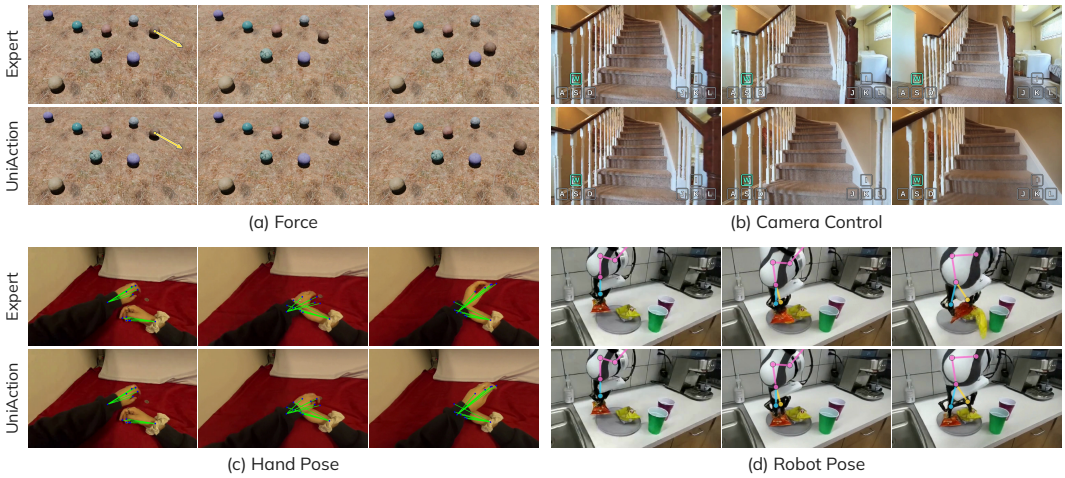


Figure 4: **Qualitative comparison with action-specific experts.** Compared with separate training, the jointly trained UniAction follows the input actions more faithfully and produces more plausible dynamics and visual appearance. Overlays indicate the input action.

the same architecture as UniAction and is trained on the data of its corresponding action form. For each action form, the expert and the jointly trained UniAction use exactly the same training samples, and each sample is seen the same number of times. The four experts are thus equivalent to a single model with $4\times$ the parameters that hard-routes samples by action form, trained with the same total compute as the jointly trained UniAction.

As shown in Table 2, the jointly trained UniAction outperforms the corresponding experts in action following across all four action forms. The jointly trained model also achieves lower or comparable LPIPS. Figure 4 further shows the qualitative comparisons.

Table 2: Jointly trained UniAction vs. action-specific experts. Best in **bold**.

Action	Action Following ↑		LPIPS ↓	
	Expert	Ours	Expert	Ours
Force	0.5436	**0.7630**	0.1295	**0.1163**
Camera	0.5182	**0.6371**	**0.4780**	0.4863
Hand	0.3701	**0.4119**	0.0934	**0.0910**
Robot	0.6310	**0.6521**	0.1286	**0.1164**

We attribute these gains to shared learning of world dynamics. An expert learns how the world responds to actions only from the limited annotated data of one action form. In contrast, the jointly trained model learns scene properties and dynamics from all four forms, so each form's data is complemented by data from the others.

4.6 Ablation Studies

We ablate the two key designs of UniAction: spatiotemporal grounding and cardinality calibration. All ablation studies use the Wan2.2-TI2V-5B (Team Wan, 2025) model under the same training and evaluation protocol. Quantitative results are reported in Table 3.

4.6.1 Spatiotemporal Grounding

UniAction maps action coordinates into the backbone's native positional system, so action and video tokens share the same positional geometry learned during video pretraining. We test this design with three variants, each changing one aspect:

Self-attn → Cross-attn keeps the RoPE (Su et al., 2024) encoding of action coordinates but injects action tokens through the text embedding's cross-attention pathway instead of self-attention. In this case, action and video tokens' coordinates no longer share the positional system. This variant tests whether encoding action coordinates is sufficient on its own, without the shared positional geometry.

RoPE → Sinusoidal APE and **RoPE → Learned APE** keep the shared self-attention but encode action coordinates with, respectively, fixed sinusoidal absolute positional encodings

(Vaswani et al., 2017) and learned absolute positional embeddings (Dosovitskiy et al., 2021), instead of the backbone's pretrained 3D RoPE. These variants test whether reusing the pretrained positional geometry matters, as opposed to merely sharing an attention pathway.

As shown in Table 3, under the same training budget, all three variants achieve low action following across all action forms and nearly fail on hand pose and robot pose, which require precise spatial grounding. These variants all need to learn the spatiotemporal correspondence between actions and videos anew, whereas UniAction places action coordinates in the backbone's pretrained positional system and directly reuses its existing spatiotemporal modeling capabilities, thereby acquiring precise control from limited action data.

Table 3: Ablations of spatiotemporal grounding and cardinality calibration. Each variant changes *one* component ($A \rightarrow B$: A replaced by B). Best in **bold**, second-best underlined.

Variant	Action Following ↑			
	Force	Camera	Hand	Robot
Self-attn → Cross-attn	0.2975	0.5050	0.0533	0.0733
RoPE → Sinusoidal APE	0.2084	<u>0.5278</u>	0.0603	0.0755
RoPE → Learned APE	0.1485	0.5206	0.0656	0.0693
w/o Cardinality Calib.	<u>0.5186</u>	0.4966	**0.4303**	<u>0.6995</u>
UniAction (Ours)	**0.7190**	**0.5842**	<u>0.4266</u>	**0.7004**

Temporal Generalization. UniAction reuses the backbone's pretrained positional geometry to encode action coordinates, enabling generalization to unseen action times. In the force training data, forces are always applied at the first frame ($t = 0$). At test time, we apply forces at $t = 8$ and $t = 24$ (Figure 5). UniAction starts the induced motion at the specified time, whereas Learned APE changes little as the force time shifts because unseen action-time embeddings receive no supervision. This suggests that UniAction leverages temporal structure learned during video pretraining rather than relearning action-time positions from limited action data.

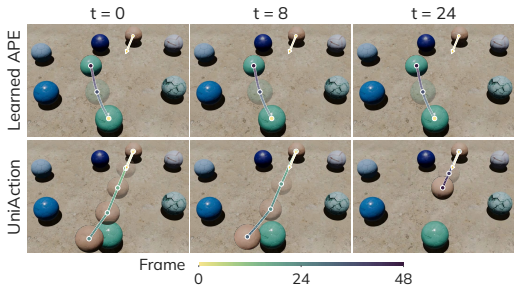


Figure 5: **Generalization.** Both models are trained only on force data, with all forces applied at $t = 0$; columns show test-time forces at different t. Trajectories are colored by frame.

4.6.2 Cardinality Calibration

To evaluate the cardinality calibration introduced in Sec. 3.3, we remove the calibration term during both training and inference and keep everything else unchanged. As shown in Table 3, removing calibration leads to clear drops in action following for force and camera control, whereas pose actions remain largely unaffected.

This pattern follows the analysis in Sec. 3.3. Without calibration, the total attention mass grows with the number of tokens: pose actions, represented by many tokens, naturally dominate the attention, while force and camera control, represented by only a few tokens, are suppressed. Cardinality calibration removes this bias and places different actions on a common attention scale, allowing actions with sparse tokens to control generation just as effectively.

5 Conclusion

We presented UniAction, a unified action-conditioning framework that represents heterogeneous actions as sparse sets of elements, each defined by the semantic identity and the spatiotemporal position. By placing them in the pretrained video model's native positional geometry, UniAction lets action and video tokens interact within a shared space, while cardinality-calibrated attention prevents representation granularity from biasing conditioning strength. Across four representative action types including force, camera control, hand pose, and robot pose, UniAction outperforms visual- and token-based conditioning alternatives, and a single jointly trained model surpasses independently trained action-specific experts, showing that heterogeneous action data can benefit one another within one world model. More broadly, UniAction offers an extensible interface through which new forms of intervention data can expand a world model's control capabilities.

AI USE STATEMENT

We used large language models (LLMs) as AI tools to aid and polish the writing of this paper, including improving grammar, clarity, and conciseness of text drafted by the authors. AI tools were also used to create some decorative icons in the figures. Initial frames used as inputs for some qualitative examples in Supplementary G and H were generated by GPT Image 2. In addition, an AI coding assistant was used to help produce the introductory video in the supplementary file, by writing the rendering code for its animated diagrams, layout, and on-screen captions. All experimental results shown in the video, including generated videos and quantitative numbers, are the actual outputs and reported results of our models and baselines; AI tools were not used to generate, modify, or select these results. The authors reviewed all AI-assisted content and take full responsibility for this paper and its supplementary materials.

ETHICS STATEMENT

Our human evaluation recruited 100 participants through Prolific to rate generated videos on action following and visual quality. Participants were compensated in accordance with Prolific's payment policy and provided informed consent before participating, and no personally identifiable information was collected. All training data used in this work are drawn from publicly available datasets and are used in accordance with their licenses.

REFERENCES

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021.

DROID Dataset Team. DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset. In *Robotics: Science and Systems*, 2024.

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, Ka Leong Cheng, Haojie Zhang, Jian Gao, Tianrui Feng, Yuzheng Liu, Yao Yao, Yinghao Xu, Xing Zhu, Yujun Shen, and Hao Ouyang. Infinite Worlds with Versatile Interactions. *ArXiv preprint*, 2026.

Nate Gillman, Charles Herrmann, Michael Freeman, Daksh Aggarwal, Evan Luo, Deqing Sun, and Chen Sun. Force Prompting: Video Generation Models Can Learn And Generalize Physics-based Control Signals. In *Conference and Workshop on Neural Information Processing Systems*, 2025.

Nate Gillman, Yinghua Zhou, Zitian Tang, Evan Luo, Arjan Chakravarthy, Daksh Aggarwal, Michael Freeman, and Chen Sun. Goal Force: Teaching Video Models To Accomplish Physics-Conditioned Goals. In *Conference on Computer Vision and Pattern Recognition*, 2026.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, Baixin Xu, Hao-Xiang Guo, Kaixiong Gong, Cyrus Wu, Wei Li, Xuchen Song, Yang Liu, Eric Li, and Yahui Zhou. Matrix-Game 2.0: An Open-Source, Real-Time, and Streaming Interactive World Model. *ArXiv preprint*, 2025.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. EgoDex: Learning Dexterous Manipulation from Large-Scale Egocentric Video. In *International Conference on Learning Representations*, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixé, and Sanja Fidler. ViPE: Video Pose Engine for 3D Geometric Perception. *ArXiv preprint*, 2025.

Junchao Huang, Guian Fang, Shengju Qian, Xianghao Kong, Zhuoran Zhao, Wei Huang, Yihua Du, Zixin Zhang, Justin Cui, Yuchao Gu, Yukang Chen, Xinting Hu, Tianyu He, Shaoshuai Shi, Zhuotao Tian, Xin Wang, Mike Zheng Shou, and Li Jiang. SolarWM: Open Data and Scalable Training for Long-Horizon Video World Models. *ArXiv preprint*, 2026.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-One Video Creation and Editing. In *International Conference on Computer Vision*, 2025.

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and Better Point Tracking by Pseudo-Labelling Real Videos. In *International Conference on Computer Vision*, 2025.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*, 2014.

Till Kroeger, Radu Timofte, Dengxin Dai, and Luc Van Gool. Fast Optical Flow Using Dense Inverse Search. In *European Conference on Computer Vision*, 2016.

Jiaqi Li, Junshu Tang, Zhiyong Xu, Longhuang Wu, Yuan Zhou, Shuai Shao, Tianbao Yu, Zhiguo Cao, and Qinglin Lu. Hunyuan-GameCraft: High-dynamic Interactive Game Video Generation with Hybrid History Condition. *ArXiv preprint*, 2025.

Yichen Li and Antonio Torralba. MultiModal Action Conditioned Video Simulation. In *International Conference on Computer Vision*, 2025.

MiniMax. MiniMax H3: An Open Model Breaking the Boundaries Between Tasks and Modalities, 2026. URL https://www.minimax.io/blog/minimax-h3.

NVIDIA. Cosmos World Foundation Model Platform for Physical AI. *ArXiv preprint*, 2025.

NVIDIA. Cosmos 3: Omnimodal World Models for Physical AI. *ArXiv preprint*, 2026a.

NVIDIA. World simulation with video foundation models for Physical AI. *ArXiv preprint*, 2026b.

Jack Parker-Holder and Shlomi Fruchter. Genie 3: A new frontier for world models, 2025. URL https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/.

Robbyant Team. Advancing Open-source World Models. *ArXiv preprint*, 2026.

Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding. *Neurocomputing*, 2024.

Sruthi Sudhakar, Ruoshi Liu, Basile Van Hoorick, Carl Vondrick, and Richard S. Zemel. Controlling the World by Sleight of Hand. In *European Conference on Computer Vision*, 2024.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. WorldPlay: Towards Long-Term Geometric Consistency for Real-Time Interactive World Modeling. *ArXiv preprint*, 2025.

Junshu Tang, Jiacheng Liu, Jiaqi Li, Longhuang Wu, Haoyu Yang, Penghao Zhao, Siruis Gong, Xiang Yuan, Shuai Shao, and Qinglin Lu. Hunyuan-GameCraft-2: Instruction-following Interactive Game World Model. *ArXiv preprint*, 2025.

Team Wan. Wan: Open and Advanced Large-Scale Video Generative Models. *ArXiv preprint*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Conference and Workshop on Neural Information Processing Systems*, 2017.

Chen Wang, Chuhao Chen, Yiming Huang, Zhiyang Dou, Yuan Liu, Jiatao Gu, and Lingjie Liu. PhysCtrl: Generative Physics for Controllable and Physics-Grounded Video Generation. In *Conference and Workshop on Neural Information Processing Systems*, 2025a.

Hanhui Wang, Yiming Xie, Haiwen Feng, Zhaoyang Lv, Shenlong Wang, and Huaizu Jiang. Streaming Video Generation with Streaming Force Control. *ArXiv preprint*, 2026.

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. VideoComposer: Compositional Video Synthesis with Motion Controllability. In *Conference and Workshop on Neural Information Processing Systems*, 2023.

Yuang Wang, Chao Wen, Haoyu Guo, Sida Peng, Minghan Qin, Hujun Bao, Xiaowei Zhou, and Ruizhen Hu. Precise Action-to-Video Generation Through Visual Action Prompts. In *International Conference on Computer Vision*, 2025b.

Zhuoyuan Wu and Jun Gao. OSCAR: Omni-Embodiment Action-Conditioned World Model for Robotics. *ArXiv preprint*, 2026.

Linxi Xie, Lisong C. Sun, Ashley Neall, Tong Wu, Shengqu Cai, and Gordon Wetzstein. Generated Reality: Human-centric World Simulation using Interactive Video Generation with Hand and Camera Control. *ArXiv preprint*, 2026.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Conference on Computer Vision and Pattern Recognition*, 2018.

Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. ControlVideo: Training-free Controllable Text-to-video Generation. In *International Conference on Learning Representations*, 2024.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. IRASim: A Fine-Grained World Model for Robot Manipulation. In *International Conference on Computer Vision*, 2025.

SUPPLEMENTARY MATERIAL

We provide an introduction video of the method in the supplementary package. We provide additional video results and comparisons in the anonymous project page:
https://uniaction.pages.dev/

A SEMANTIC REPRESENTATION AND EXTENSIBILITY

Each semantic identity $s \in \mathcal{S}$ indexes an independently initialized action token $E[s] \in \mathbb{R}^d$. Under isotropic initialization, let $\bar{\mathbf{e}}_s = E[s]/\|E[s]\|_2$ denote its normalized representation. For two distinct identities $s \neq s'$,

$$\mathbb{E}\left[\bar{\mathbf{e}}_s^\top \bar{\mathbf{e}}_{s'}\right] = 0, \qquad \mathrm{Var}\left[\bar{\mathbf{e}}_s^\top \bar{\mathbf{e}}_{s'}\right] = \mathcal{O}\left(\frac{1}{d}\right). \tag{A.1}$$

Thus, in a high-dimensional feature space, different semantic identities are initialized with low correlation and approximate orthogonality, while their correspondence to scene evolution is learned through joint action-conditioned training.

Extensibility. The semantic vocabulary can be extended by appending independently initialized tokens to the same codebook. New identities inherit the same low-correlation initialization with existing identities, allowing additional action forms to be incorporated under the same semantic representation and learned through the shared world model.

B CARDINALITY-CALIBRATED ATTENTION

B.1 DERIVATION

UniAction processes video and action tokens through a shared self-attention softmax. For a video query $\mathbf{q}_v$, let the attention logits to the j-th video key and the k-th action key be

$$L_{v,j} = \frac{\mathbf{q}_v^\top \mathbf{k}_{v,j}}{\sqrt{d_h}}, \qquad L_{a,k} = \frac{\mathbf{q}_v^\top \mathbf{k}_{a,k}}{\sqrt{d_h}}, \tag{B.1}$$

where d_h is the attention head dimension. The video and action logits are normalized by the same softmax. The total attention share assigned to the action tokens is therefore

$$P_{\text{action}} = \frac{\sum_{k=1}^{N_a} \exp(L_{a,k})}{\sum_{j=1}^{N_v} \exp(L_{v,j}) + \sum_{k=1}^{N_a} \exp(L_{a,k})}. \tag{B.2}$$

Define the unnormalized video and action contributions as

$$Z_v = \sum_{j=1}^{N_v} \exp(L_{v,j}), \qquad Z_a = \sum_{k=1}^{N_a} \exp(L_{a,k}),$$

such that $P_{\text{action}} = Z_a/(Z_v + Z_a)$.

Cardinality bias. The action contribution can be rewritten as

$$Z_a = N_a \cdot \mathrm{mean}_k[\exp(L_{a,k})]\,. \tag{B.3}$$

Here, $\mathrm{mean}_k[\exp(L_{a,k})]$ captures the average compatibility between the video query and the action tokens, while N_a is determined by the representation granularity of the action. Consequently, actions with comparable average compatibility can receive different total attention shares solely because they are represented by different numbers of tokens.

Calibration. Cardinality calibration adds the same bias to every action logit,

$$b = \log \frac{N_{\text{ref}}}{N_a}, \qquad L'_{a,k} = L_{a,k} + b, \tag{B.4}$$

where N_{ref} is a fixed reference cardinality shared across action forms. The calibrated action contribution becomes

$$\begin{aligned} Z'_a &= \sum_{k=1}^{N_a} \exp(L_{a,k} + b) = \frac{N_{\text{ref}}}{N_a} \sum_{k=1}^{N_a} \exp(L_{a,k}) \\ &= N_{\text{ref}} \cdot \mathrm{mean}_k[\exp(L_{a,k})]\,. \end{aligned} \tag{B.5}$$

The corresponding action attention share is

$$P'_{\text{action}} = \frac{N_{\text{ref}} \operatorname{mean}_k[\exp(L_{a,k})]}{Z_v + N_{\text{ref}} \operatorname{mean}_k[\exp(L_{a,k})]}. \tag{B.6}$$

The explicit dependence on the actual action cardinality N_a is thus removed from the total action contribution, while the average action–video compatibility is preserved.

Since the same bias is applied to all action logits, $L'_{a,i} - L'_{a,j} = L_{a,i} - L_{a,j}$, so cardinality calibration preserves the relative attention relationships among action tokens while placing different action forms on a common attention scale.

B.2 Forward, Backward, and Inference

Forward computation. Let L_v and L_a denote the video and action logit matrices for the video queries. The calibrated attention weights are

$$[A_v \mid A_a] = \operatorname{softmax}([L_v \mid L_a + b]), \qquad b = \log \frac{N_{\text{ref}}}{N_a}, \tag{B.7}$$

and the video output is

$$O_v = A_v V_v + A_a V_a. \tag{B.8}$$

Cardinality calibration therefore enters the forward computation as a scalar offset applied to the action-logit block before the shared softmax.

Backward computation. The backward pass follows the standard softmax-attention derivatives. Let $G = \partial \mathcal{L} / \partial O_v$. For the k-th action token,

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial V_{a,k}} &= A_{a,k} G, \\ \frac{\partial \mathcal{L}}{\partial L'_{a,k}} &= A_{a,k} G^\top (V_{a,k} - O_v). \end{aligned} \tag{B.9}$$

The calibrated attention weight $A_{a,k}$ therefore controls both the forward contribution of each action token and the training signal propagated through the action branch.

Training and inference. The reference cardinality N_{ref} is fixed, and N_a is determined by the current action representation. The calibration introduces no additional learnable parameters. The same calibrated attention computation is used during training and inference.

B.3 Algorithm

For each attention layer, cardinality calibration modifies only the action logits before the joint softmax. Given video queries Q_v, video keys and values (K_v, V_v), and action keys and values (K_a, V_a), we compute:

Algorithm 1 Cardinality-calibrated attention

Input: Queries Q_v; keys/values (K_v, V_v) and (K_a, V_a); d_h, N_a, N_{ref}.
Output: Video-token output O_v.

1: $L_v \leftarrow Q_v K_v^\top / \sqrt{d_h}$
2: $L_a \leftarrow Q_v K_a^\top / \sqrt{d_h}$
3: $b \leftarrow \log(N_{\text{ref}} / N_a)$
4: $L'_a \leftarrow L_a + b$
5: $[A_v \mid A_a] \leftarrow \operatorname{softmax}([L_v \mid L'_a])$
6: $O_v \leftarrow A_v V_v + A_a V_a$

Thus, the calibrated attention retains the original attention computation and introduces a single cardinality-dependent offset on the action-key logits.

C EXPERIMENTAL DETAILS

Data and implementation. We select 10K action-annotated training clips from each of GoalForce (Gillman et al., 2026), SolarWM (Huang et al., 2026), EgoDex (Hoque et al., 2026), and DROID (DROID Dataset Team, 2024), for a total of 40K samples. Each clip is sampled to 81 frames and resized and cropped to 480×832 (height $\times$ width) to match the input specification of the base model. The 10K GoalForce samples consist of 5K ball, 2.5K domino, and 2.5K flower scenes.

We use the high-noise expert of Wan2.2-I2V-A14B (Team Wan, 2025) as the base model, train it over the full noise range, and use this expert at all denoising timesteps during inference. LoRA (Hu et al., 2022) is applied to the Q/K/V/O projections of the attention modules and the two linear layers of the FFN, with both rank and alpha set to 128. During action adaptation, only the action branch is updated, i.e., LoRA is applied only to action tokens, while the video branch remains frozen. The total training budget is two full passes over the training data. All compared methods use the same data preprocessing, per-task training data budget, and inference configuration, and are evaluated under the same task and evaluation protocol. We fix $N_{\text{ref}} = 784$.

Ablation studies are conducted on Wan2.2-TI2V-5B, with video clips of 480×832 resolution and 49 frames for both training and evaluation; all other settings follow the main experiments.

Test set construction. For each source sample, we fix the initial frame and text prompt and vary only the action condition. We use 107 source samples, none of which appear in the training set, and construct 410 test conditions per model, including both original action conditions and action-intervention variants (Table 4).

Table 4: Construction of the test set.

Action Form	Construction	#Conditions
Force	8 ball scenes $\times$ 5; 16 domino scenes $\times$ 2; 16 flower scenes $\times$ 2	104
Camera Control	13 scenes $\times$ 8 directions (`W/A/S/D/I/J/K/L`)	104
Hand Pose	20 samples $\times$ {original, stationary, horizontal flip, vertical flip, flip along both axes}	100
Robot Pose	34 samples $\times$ {original, horizontally flipped center trajectory, reversed gripper trajectory}	102
Total		**410**

The ball variants consist of the original direction and its rotations by $\pm 45^\circ$ and $\pm 90^\circ$; the domino variants consist of the original direction and its 180° reversal; the flower variants consist of the original action and a paired action that applies force in the opposite direction.

Action Following scores. Let $C(x) = \min(1, \max(0, x))$. Each task uses the following fixed scoring rule.

- **Force.** For ball and domino scenes, we estimate the initial motion response of the generated video with DIS optical flow (Kroeger et al., 2016). For flower scenes, since the bending and swaying direction of the flower does not necessarily coincide with the direction of the applied force, we extract DIS optical flow features at multiple locations and estimate the force direction with a ridge regression mapping. The mapping is fitted only on GoalForce training data and kept fixed throughout the evaluation of all models. Let $\hat{\mathbf{f}}$ and $\mathbf{f}$ denote the estimated and target force directions, respectively. The score is

$$s_{\text{F}} = C\left(\frac{\hat{\mathbf{f}}^\top \mathbf{f}}{\|\hat{\mathbf{f}}\| \, \|\mathbf{f}\|}\right).$$

  Samples that do not produce a valid motion response receive a score of 0.
- **Hand Pose and Robot Pose.** We use CoTracker3 (Karaev et al., 2025) to track the locations corresponding to the action annotations, and compare the actual displacement $\Delta\mathbf{p}$ relative to the initial position with the target displacement $\Delta\mathbf{c}$. The score is obtained by normalizing the error between the actual and target displacements, so that results more consistent with the target motion

receive higher scores. For the stationary condition, the score is based on the residual motion relative to the initial position.

- **Camera Control.** We estimate the camera trajectory with ViPE (Huang et al., 2025), taking the first frame as the origin. To account for camera drift inherent to the model and scene, we use the no-action generation of the same scene and model as a reference and subtract this baseline from the predicted response. We measure the cosine similarity between the residual motion vector and the commanded direction, separately for translation and rotation, normalize it to $[0, 1]$, and report the average score.

All Action Following scores lie in $[0, 1]$, with higher scores indicating better action following. We first average the different action variants of each source sample with equal weights and then aggregate over all source samples. For Force, the ball, domino, and flower subcategories are further averaged with equal weights.

LPIPS. We compute AlexNet-based LPIPS (Zhang et al., 2018) only under the original action condition, comparing generated frames with the corresponding ground-truth frames at 480×832 resolution. Since the first frame is given as model input, we exclude it, compute LPIPS on each of the remaining 80 frames, and average the results.

Human evaluation. A total of 100 participants completed the questionnaire. Each participant was randomly assigned 8 groups of videos (2 per action form), where each group contains one video generated by each of UniAction, VACE, and Cross-Attention, resulting in 24 videos per participant. The three videos in each group are displayed side by side in a randomized order. Referring to the action markers overlaid on the videos, participants rate each video on a 1–5 scale along the following two dimensions:

- **Q1 (Action Following):**
  "Does the video do what the action says? (Compare with the other videos.)"
- **Q2 (Visual Quality):**
  "How good is the video quality, including clarity, stability, and plausible physical interactions? (Compare with the other videos.)"

D Efficiency Analysis

We measure action-conditioning overhead for one denoising forward pass on a single GH200, relative to the backbone without action input (1678.4 TFLOPs and 3.06 s). Table 5 reports trainable parameters, additional TFLOPs, and relative latency increases.

VACE runs a separate conditioning branch at full video resolution, adding approximately 21% to both computation and latency regardless of the action form. Cross Attention also has nearly fixed overhead. UniAction only appends action tokens, so its overhead varies with the number of action tokens. It incurs the lowest additional latency across all four action forms.

Table 5: Additional cost of action conditioning per denoising forward pass, relative to the backbone without action input.

Method	Params	ΔTFLOPs				ΔLatency (%)			
		Force	Camera	Hand	Robot	Force	Camera	Hand	Robot
VACE	3.1B	351.2	351.2	351.2	351.2	21.3	21.0	21.2	21.1
Cross Attention	640M	33.5	35.9	36.0	36.0	8.8	8.6	9.0	9.0
UniAction	404M	0.1	5.9	72.9	37.2	3.6	3.0	6.6	4.9

E Additional Qualitative Comparisons

Figures 6–8 provide 13 additional comparisons across the four action forms. Each example shows six frames per method, ordered from left to right. Arrows, key indicators, and pose overlays visualize the input action conditions rather than generated content.

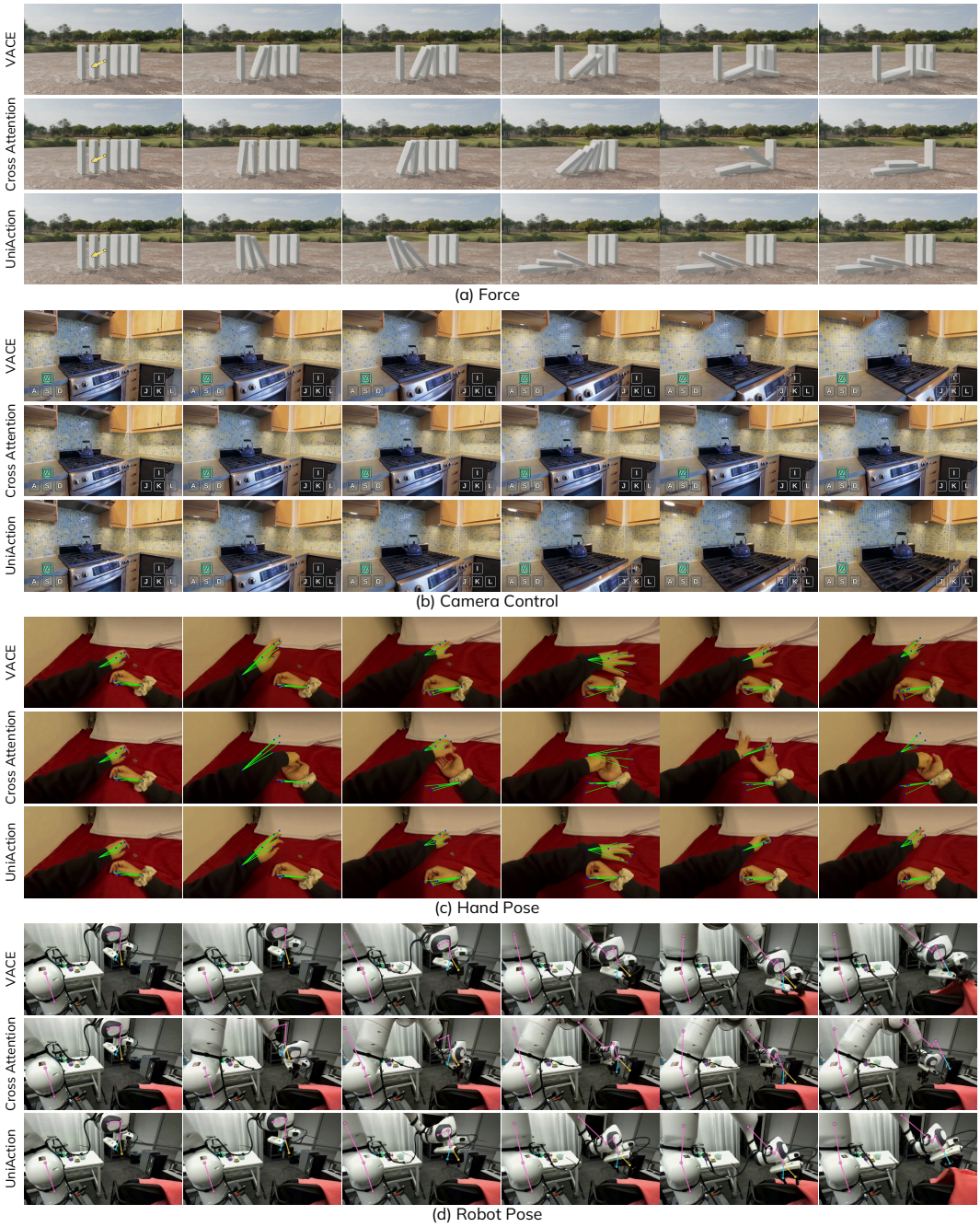


Figure 6: **Additional qualitative comparisons.** The camera input in (b) is `W` (forward camera motion). All overlays indicate input conditions.

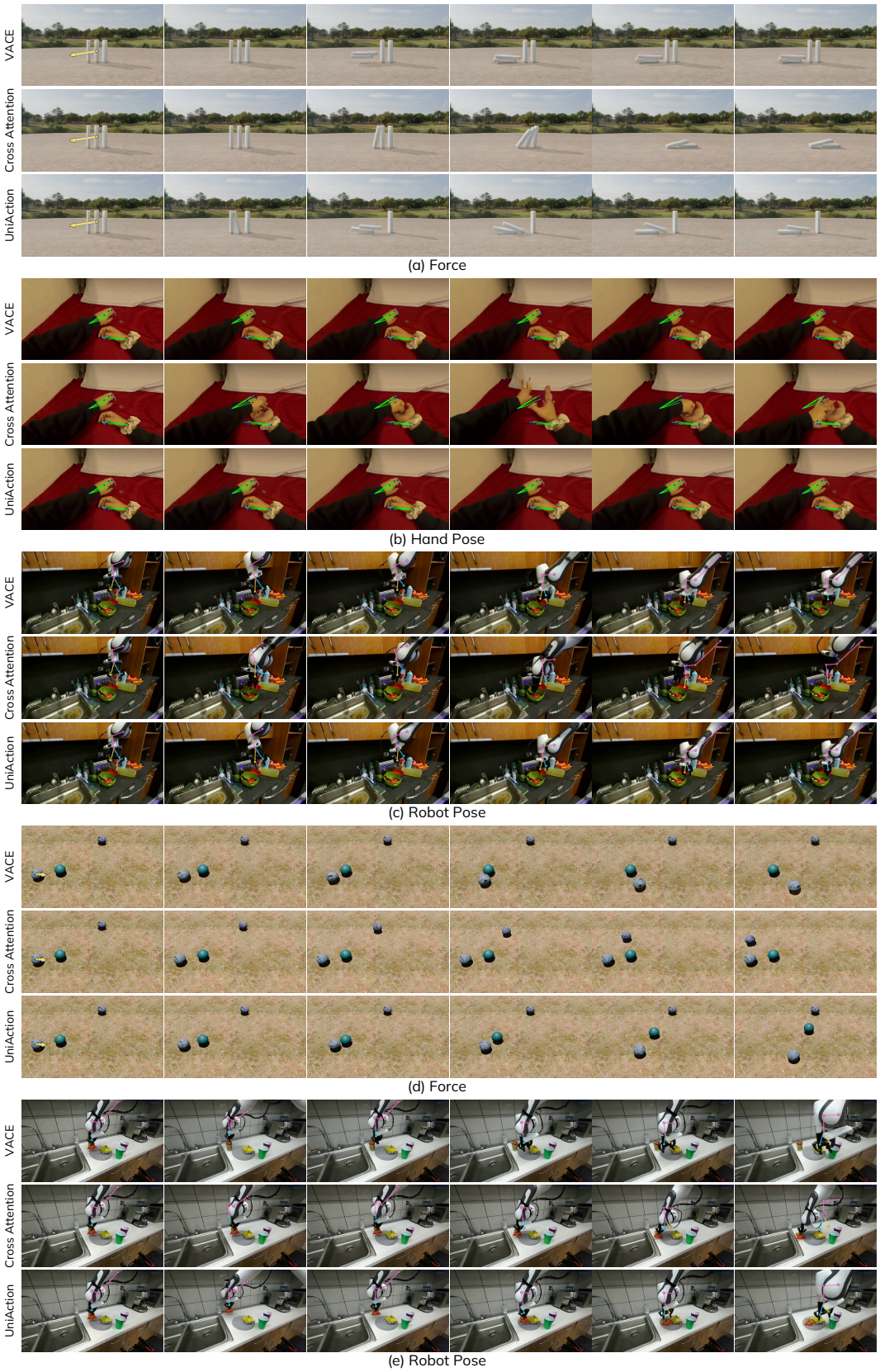


Figure 7: **Additional qualitative comparisons.** Overlays indicate input conditions.

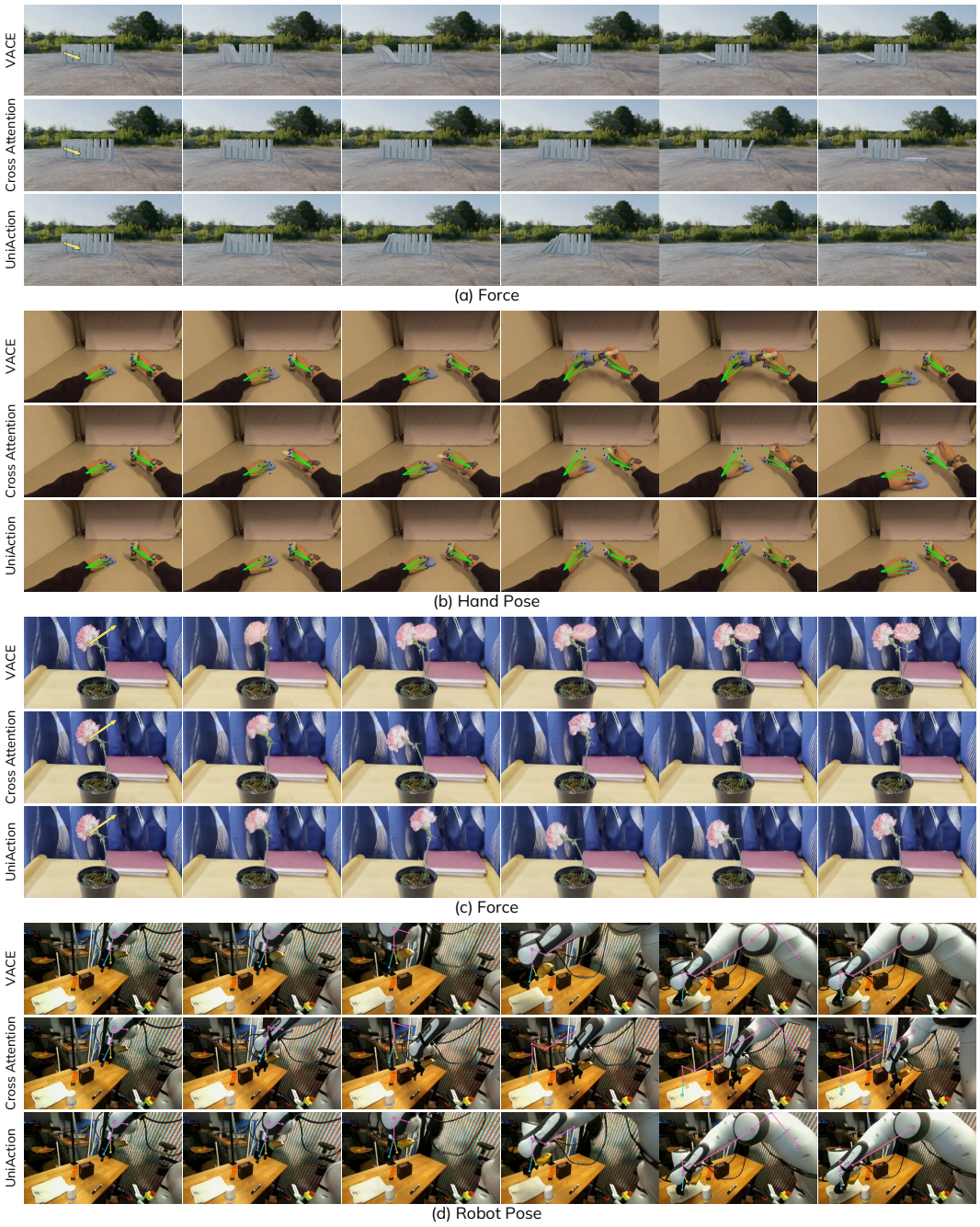


Figure 8: **Additional qualitative comparisons.** Overlays indicate input conditions.

F Additional Temporal Generalization Results

Following the setup in Sec. 4.6.1 and Figure 5, we train UniAction only on force data, with all forces applied at $t = 0$. Figures 9 and 10 show that delaying force application to $t = 8$ or $t = 24$ delays object motion accordingly, demonstrating generalization to unseen intervention times. Yellow arrows indicate the input force.



(a) Force applied at t = 0



(b) Force applied at t = 8



(c) Force applied at t = 24

Figure 9: **Temporal generalization on dominoes.**



(a) Force applied at t = 0



(b) Force applied at t = 8

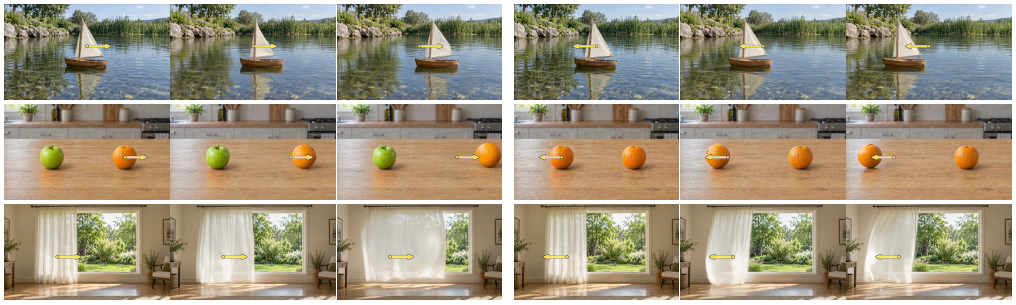


(c) Force applied at t = 24

Figure 10: **Temporal generalization on balls.**

G SCENE GENERALIZATION

Figure 11 shows 24 generations from the jointly trained UniAction. Except for the second row in the left column of (c) and the upper-left example in (d), whose initial frames come from the test set, all initial frames were generated by GPT Image 2 and were unseen during training. Each strip shows the initial frame and two later frames; paired left and right strips use matching frame indices.



(a) Force



(b) Camera Control



(c) Hand Pose

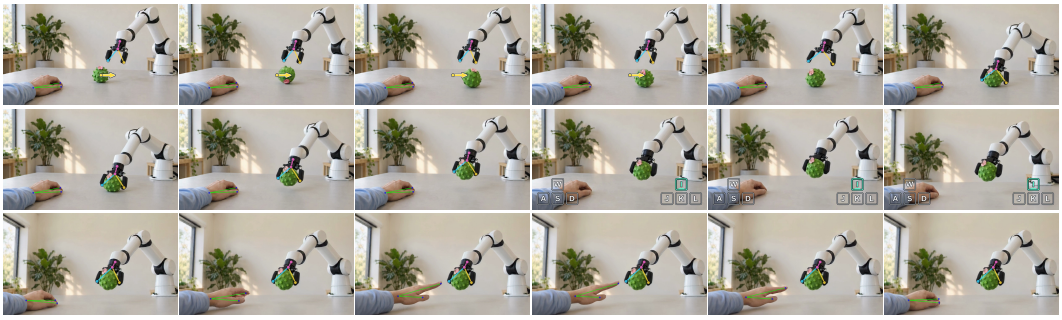


(d) Robot Pose

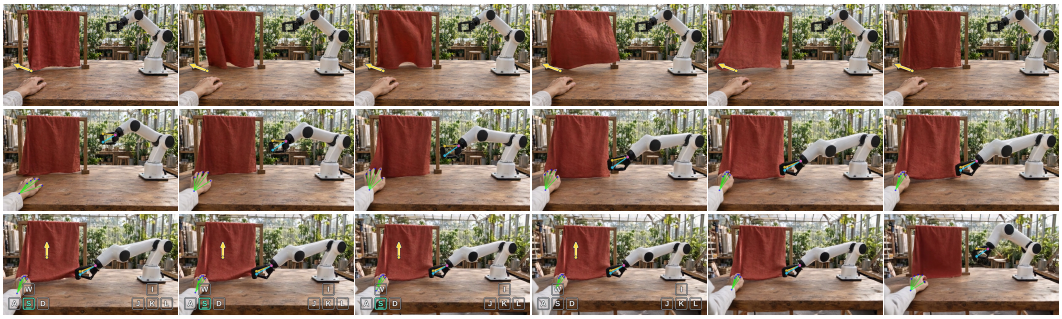
Figure 11: **Scene and appearance generalization across four action forms.** All six variants in (c) share the same hand-pose input. Arrows and highlighted keys indicate the input actions.

H Compositional Action Generalization

Figure 12 shows three long videos generated by the jointly trained UniAction. Each training sample contains only one action form. Since all actions share the same representation and conditioning pathway, tokens from different action forms can be directly concatenated into a single sequence at inference time without additional training. These results combine force, camera control, hand pose, and robot pose, with both simultaneous actions and transitions between actions over time. The initial frames were generated by GPT Image 2.



(a)



(b)



(c)

Figure 12: **Generalization to unseen action compositions.** Each block (a)–(c) shows one video in three rows of six selected frames, read from left to right and then from top to bottom. Arrows, pose overlays, and highlighted keys indicate the input actions.