COMPWORLD: COMPOSITIONAL ACTION CONTROL FOR VIDEO WORLD MODELS

Anonymous authors
Paper under double-blind review

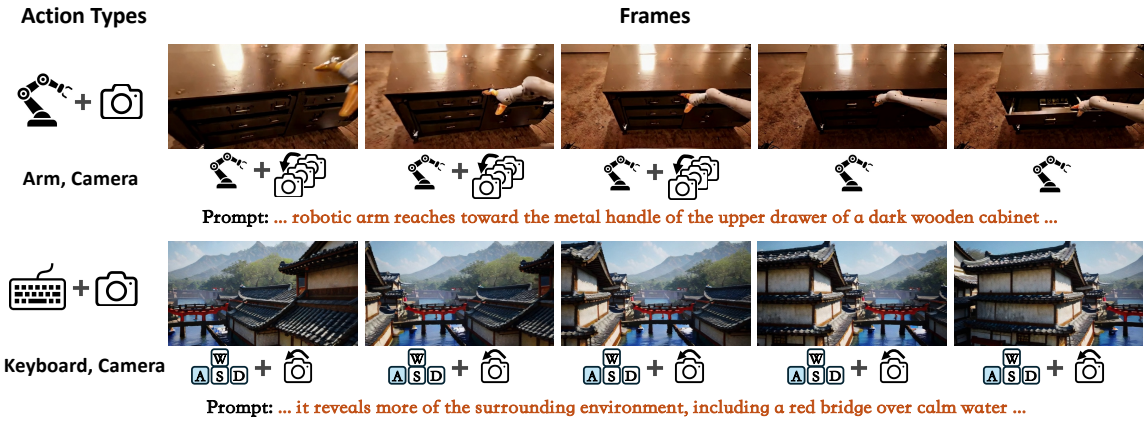

Figure 1: **Compositional action control across heterogeneous action types.** Given a text prompt and multiple action streams, CompWorld conditions on heterogeneous controls such as arm motion with camera pose and keyboard movement with camera rotation, producing coherent videos whose dynamics reflect the combined effects of the specified actions.

ABSTRACT

Interactive video world models increasingly serve as controllable simulators for domains such as gaming, camera-controlled navigation, and robot manipulation, where video generation models must support heterogeneous action types such as keyboard commands, mouse motion, camera poses, and robot-arm trajectories. However, most existing video world models build action-control modules for particular action types, leaving different action types in isolated representations that are difficult to combine reliably without paired multi-action supervision. We propose CompWorld, a framework for compositional action control in video world models. CompWorld keeps the original action streams as precise inputs, while projecting them into a text-aligned unified action space that uses the pretrained text space of a text-conditioned video generation model as a shared interface prior. With action type-specific encoders, learned "not provided" placeholders, and Action-to-Text Fusion, CompWorld enables different action types to condition the video generator through Fused-Text Cross-Attention and compose at inference time. We evaluate CompWorld in compositional and separate action-control settings, including two compositional benchmarks—keyboard plus camera pose and robot-arm plus camera pose—that pair heterogeneous controls only at inference, without paired multi-action training annotations. Experiments across keyboard, mouse, camera-pose, and robot-arm controls show that CompWorld improves compositional action control, action following, and video quality over FrameInject, TokenConcat, and GameFactory, and yields a more semantically aligned conditioning interface.

1 INTRODUCTION

Recent video world models increasingly serve as controllable simulators for interactive domains such as gaming, camera-controlled navigation, and robot manipulation (Bruce et al., 2024; Yu et al., 2025b; Zhu et al., 2024; Gao et al., 2025; NVIDIA, 2025). These applications require video generation

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

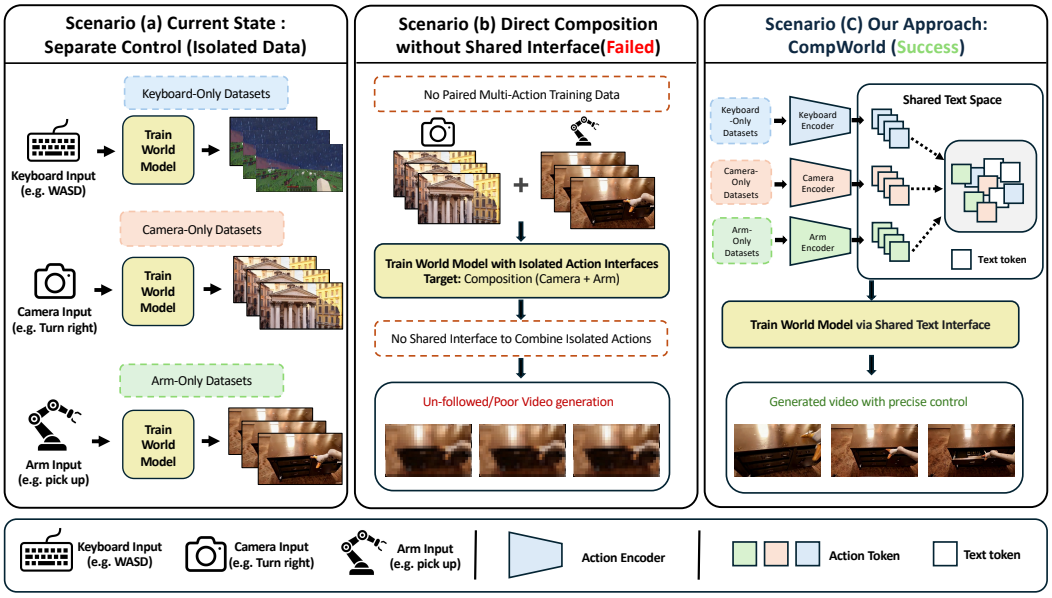

Figure 2: **Overview of our approach.** (a) Existing datasets provide separate action annotations (e.g., keyboard, camera, arm), leading to separate action-control interfaces. (b) Without a shared interface for isolated action representations, direct composition can yield unreliable joint control. (c) Our method maps heterogeneous actions into a shared text-conditioned interface, enabling unified training across disjoint datasets and cross-action-type composition at inference.

models to support heterogeneous action control types. For example, in a first-person game, a player simultaneously presses WASD keys to move through the environment while dragging the mouse to look around, so locomotion and viewpoint control must take effect together in every frame; in a tabletop manipulation scene, a robot arm reaches for an object while the camera orbits to provide a better view of the grasp, requiring arm trajectory and camera pose control to remain coordinated in one coherent video. These scenarios require **compositional action control** across heterogeneous action types for interactive video generation models, as shown in Fig. 1.

However, most existing video world models build and train action control modules only for a particular action type, such as camera or motion control (He et al., 2024; Wang et al., 2024c), as shown in Fig. 2(a). Although these modules work well for their target action type, they represent different action types in isolated forms: keyboard inputs as discrete embeddings, camera poses as continuous pose vectors, and arm commands as trajectory vectors, *etc*. This isolation makes inference-time composition difficult: if action controls are combined only after being learned in separate forms, one action stream may dominate, another may be ignored, or the video may fail to follow the intended joint command coherently (Fig. 2(b)). Since paired multi-action supervision is rarely available, the model needs a shared interface for combining these isolated action representations.

We address this problem with CompWorld, a video generation framework for *Compositional Action Control*. The key idea is to use the pretrained text space of a text-conditioned video generation model as a shared interface prior for heterogeneous action representations, as shown in Fig. 2(c). This choice is motivated by two complementary observations. On the one hand, pretrained text embedding spaces offer such an interface: they encode semantic relations among concepts such as moving, rotating, and grasping, and text-to-video pretraining aligns this space with visual dynamics. On the other hand, this does not mean replacing precise action inputs with natural language prompts: language can describe actions at a high level, but it cannot faithfully specify cursor displacements, camera rotations, or end-effector trajectories. CompWorld therefore keeps the original action streams as precise inputs, while projecting them into a text-aligned unified action space that provides a pretrained semantic prior for conditioning the video generation model.

Concretely, CompWorld uses action type-specific encoders to convert raw action inputs such as keyboard, mouse, camera-pose, and robot-arm sequences into action tokens with the same dimen-

sionality as text tokens. Since a training video usually contains only a subset of the supported action types, CompWorld keeps a fixed action slot for each supported action type and fills unavailable action types with a learned "not provided" placeholder. This placeholder explicitly represents the absence of an action type, so videos paired with different action types can still be trained under the same text-aligned unified action space. Action-to-Text Fusion then enables text tokens to attend to action tokens. The resulting fused text tokens replace the original text tokens and condition the video generation model through Fused-Text Cross-Attention. In this way, each action type learns how to modify the same pretrained text space, while multiple available action types can be fused jointly at inference time to produce compositional action control.

Our work makes three main contributions:

1. While most previous works focus on conditioning video world models with single action control types, we are the first to formulate compositional action control in video world models trained on separate single-action-type datasets, where heterogeneous action controls must compose at inference without paired multi-action supervision.
2. We propose CompWorld, a framework that projects heterogeneous action streams into a text-aligned action space, using action type-specific encoders, learned "not provided" placeholders, and Action-to-Text Fusion to condition video generation through the pretrained text space.
3. We conduct experiments across keyboard, mouse, camera-pose, and robot-arm controls, including two compositional benchmarks—keyboard plus camera pose and robot-arm plus camera pose—that pair heterogeneous controls at inference without paired multi-action training data, and show that the design transfers to the open Wan2.1-1.3B model. CompWorld follows heterogeneous controls better than FrameInject, TokenConcat, and GameFactory, with a more semantically aligned interface than FrameInject and TokenConcat.

2 Related Work

Action Control in Video World Models. World models learn predictive dynamics for sequential decision making (Ha & Schmidhuber, 2018; Hafner et al., 2020), and video world models increasingly serve as controllable simulators for driving, gaming, and robotics (Kim et al., 2021; Hu et al., 2023; Gao et al., 2024; Bruce et al., 2024; Yu et al., 2025b; Zhu et al., 2024; Zhou et al., 2024b; Gao et al., 2025; NVIDIA, 2025). Game world models respond to keyboard and mouse actions (Decart et al., 2024; Alonso et al., 2024; Parker-Holder et al., 2024; Zhang et al., 2025), and Hunyuan-GameCraft (Li et al., 2025) unifies both into a shared camera representation; these models learn joint control from gameplay in which the inputs co-occur. Other methods condition generation on camera trajectories (He et al., 2024; Yang et al., 2024b; Bai et al., 2024), motion cues (Wang et al., 2024c; Guo et al., 2025; Fu et al., 2025), or physics-based interactions (Montanaro et al., 2024; Zhang et al., 2024; Liu et al., 2024), each through a module for one action type. Since action-annotated datasets usually cover a single action type or interaction setting (Guss et al., 2019; Ebert et al., 2021; Caesar et al., 2020; Yu et al., 2025b), paired multi-action supervision is rare, and these designs do not directly address how heterogeneous controls should be combined without it. UniSim (Yang et al., 2024a) also learns from datasets that each cover a different aspect of interaction; we focus on composing action types that were only ever annotated separately.

Composing Heterogeneous Conditions. Controllable diffusion models combine heterogeneous conditions in two ways. Some train one model on conditions derived from the same sample (Huang et al., 2023; Zhao et al., 2023; Wang et al., 2023), which presumes jointly annotated data; others combine separately trained modules or conditional scores at inference (Zhang et al., 2023; Mou et al., 2024; Liu et al., 2022), which leaves the interaction between conditions unlearned (Fig. 2(b)). CompWorld lies between the two: each action type is learned from its own dataset, yet all of them modify the same pretrained text space, which mediates their composition.

Pretrained Text Space for Action Representations. Text-conditioned video generation has advanced from early transformer and diffusion models (Hong et al., 2022; Ho et al., 2022; Blattmann et al., 2023; Chen et al., 2024; Kondratyuk et al., 2023) to recent large-scale foundation models (Bao et al., 2024; Kong et al., 2024; The Movie Gen Team, 2024; Wang et al., 2025a; Seawead et al., 2025; DeepMind, 2024; Mao et al., 2025). Related multimodal generation and embodied AI studies

further show that language and text-aligned representations can support semantic abstraction and compositional generalization (Yu et al., 2023; Wang et al., 2024b; Deng et al., 2024; Zhou et al., 2024a; Jiang et al., 2022; Brohan et al., 2023; Du et al., 2023). Robot learning, in turn, handles heterogeneous action spaces with shared transformer policies across embodiments (Octo Model Team et al., 2024; Wang et al., 2024a; Doshi et al., 2024) or with latent actions learned from unlabeled video (Bruce et al., 2024; Gao et al., 2025; Ye et al., 2025), but these shared spaces are learned from robot or video data, whereas CompWorld inherits a pretrained text space. It still keeps actions as precise inputs rather than prompts, since language alone cannot faithfully specify cursor displacements, camera rotations, or end-effector trajectories.

3 METHOD

CompWorld is designed to learn compositional action control when paired multi-action supervision is rarely available. Instead of combining isolated action-control modules only at inference time, CompWorld maps heterogeneous action streams into a text-aligned action space and lets them modify the pretrained text space before the video generation model consumes the condition. This keeps raw actions as precise inputs, while using the pretrained text space as a shared interface prior for different action types. In other words, CompWorld does not learn an unstructured action space from scratch; it learns how each action type should enter a text space already aligned with video dynamics.

We first summarize the text-to-video backbone in Sec. 3.1. We then describe how CompWorld represents available and unavailable action types with fixed action slots in Sec. 3.2, and how Action-to-Text Fusion turns these action tokens into fused text tokens in Sec. 3.3.

3.1 PRELIMINARIES

Text-to-Video Model. Our backbone is a full-sequence latent text-to-video diffusion model (Fig. 3): a causal 3D VAE followed by a Diffusion Transformer (DiT) whose blocks contain 2D attention, 3D attention, cross-attention, and FFN modules (He et al., 2022; Peebles & Xie, 2023). Given a video $\mathbf{X} = \{x^0, x^1, \ldots, x^{nr}\}$, the VAE encoder produces latent video tokens $\mathbf{Z}_0 = \mathcal{E}_v(\mathbf{X}) = \{z^0, z^1, \ldots, z^n\}$, where the temporal compression factor r maps $1 + nr$ frames to $1+n$ latent steps. A text encoder maps the prompt c to L text tokens $\mathbf{Y} = \mathcal{E}_{\text{text}}(c) \in \mathbb{R}^{L\times d}$, where d is the hidden dimension. During diffusion training, noise ϵ is added to obtain $\mathbf{Z}_t$, and the denoiser predicts the noise conditioned on text tokens:

$$\mathcal{L}_{\text{t2v}} = \mathbb{E}\left[\|\epsilon_\theta(\mathbf{Z}_t, t, \mathbf{Y}) - \epsilon\|_2^2\right].$$

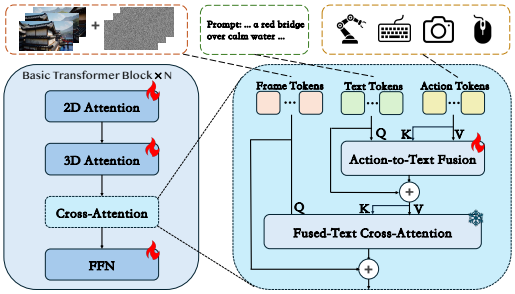

Figure 3: Overview of the model architecture. Left: the pretrained video backbone with 2D attention, 3D attention, cross-attention, and FFN blocks. Right: action tokens are first fused into text tokens by **Action-to-Text Fusion**; the resulting fused text features then condition frame tokens through **Fused-Text Cross-Attention**.

Here t denotes the diffusion denoising timestep, which is distinct from the latent video timestep index τ used below. In the pretrained backbone, text tokens are injected into visual tokens through text-video cross-attention. In CompWorld, we apply this inherited layer to fused text tokens and refer to it as **Fused-Text Cross-Attention**.

3.2 ACTION TOKENS WITH NOT-PROVIDED PLACEHOLDERS

Let $\mathcal{M}$ denote the set of supported action types, including keyboard, mouse, camera-pose, and robot-arm actions. For each supported action type m, an action type-specific encoder $\mathcal{E}_{a,m}$ maps the raw action stream $\mathbf{A}_m$ into action tokens $\mathbf{U}_m \in \mathbb{R}^{(n+1)\times w\times d}$, where w is the action-window length aligned to each latent timestep. Raw actions are grouped according to the latent video timesteps following the sliding-window design of GameFactory (Yu et al., 2025b), so each latent timestep attends to the corresponding window of w action tokens. Discrete keyboard actions are represented

as multi-hot vectors and projected into action tokens, while continuous mouse, camera-pose, and robot-arm trajectories are encoded from their raw action vectors by action type-specific encoders.

Different training videos may annotate different subsets of the supported action types. To keep the input format fixed, CompWorld assigns one action slot to each supported action type and keeps the slot order unchanged for all training and inference examples. If an action type is available, its slot is filled with encoded action tokens; otherwise, it holds a learned "not provided" placeholder:

$$\mathbf{U}_m = \begin{cases} \mathcal{E}_{a,m}(\mathbf{A}_m), & \text{if action type } m \text{ is provided,} \\ \mathrm{Repeat}(\mathbf{u}_m^{\mathrm{miss}}, n+1, w), & \text{otherwise.} \end{cases}$$

The placeholder explicitly represents the absence of an action type. With this completed slot structure, videos annotated with different action types can train the same text-aligned action space.

At inference, the same fixed slots can contain any available subset of action types, including multiple action streams that were not paired together in the training annotations.

3.3 Action-to-Text Fusion

After filling the action slots, CompWorld concatenates the action tokens from all supported action types at each latent timestep in a fixed slot order:

$$\mathbf{U}_\tau = \mathrm{Concat}_{m\in\mathcal{M}}(\mathbf{U}_{m,\tau}) \in \mathbb{R}^{|\mathcal{M}|w\times d},$$

where $\tau \in \{0,\dots,n\}$ indexes the latent video timestep. We write $\mathbf{U} = \{\mathbf{U}_\tau\}_{\tau=0}^{n}$ for the temporally ordered action-token collection. CompWorld then fuses each $\mathbf{U}_\tau$ into the pretrained text space. In the ℓ-th DiT block, Action-to-Text Fusion lets text tokens attend to the action tokens at the corresponding latent timestep. Text tokens serve as queries and action tokens as keys and values:

$$\mathbf{Y}_\tau^{\mathrm{fuse},(\ell)} = \mathrm{CrossAttn}_{\mathrm{A2T}}^{(\ell)}(\mathbf{Y}, \mathbf{U}_\tau, \mathbf{U}_\tau) + \mathbf{Y}.$$

Stacking over latent timesteps gives $\mathbf{Y}^{\mathrm{fuse},(\ell)} \in \mathbb{R}^{(n+1)\times L\times d}$, so each latent timestep receives a different fused text condition. The residual connection preserves the pretrained text representation, while the attention output adds information from the provided action streams. Because text tokens attend to action tokens, actions modify the text condition before it is used by the video generation model, rather than being injected as a separate control branch. The condition therefore remains text-like, but is adjusted by keyboard, mouse, camera-pose, or robot-arm controls.

Fused-Text Cross-Attention then uses the fused text tokens in place of the original text tokens at the corresponding latent timestep. Let $\mathbf{H}_\tau^{(\ell)}$ denote the visual tokens at latent timestep τ in DiT block ℓ:

$$\mathbf{H}_{\mathrm{out},\tau}^{(\ell)} = \mathrm{CrossAttn}_{\mathrm{FTC}}^{(\ell)}(\mathbf{H}_\tau^{(\ell)}, \mathbf{Y}_\tau^{\mathrm{fuse},(\ell)}, \mathbf{Y}_\tau^{\mathrm{fuse},(\ell)}).$$

Visual tokens are thus conditioned through the pretrained text-space interface, and because every action type modifies the same text space, several action types can be fused jointly at inference. To preserve this interface during action-conditioned training, we keep the Fused-Text Cross-Attention weights fixed; the action type-specific encoders, learned "not provided" placeholders, Action-to-Text Fusion layers, and the remaining backbone blocks are trainable.

Baseline Injection Strategies. For comparison, we consider two common ways to inject the same action tokens. FrameInject adds action tokens to visual tokens before the backbone 3D attention,

$$\tilde{\mathbf{H}}_\tau^{(\ell)} = \mathbf{H}_\tau^{(\ell)} + P(\mathbf{U}_\tau),$$

where P projects the action tokens at the corresponding latent timestep into the visual stream. TokenConcat concatenates text and action tokens at each latent timestep before cross-attention,

$$\mathbf{C}_{\mathrm{concat},\tau} = [\mathbf{Y}; \mathbf{U}_\tau].$$

CompWorld differs from both by first fusing action tokens into text tokens, so the video generation model is still conditioned through Fused-Text Cross-Attention. Appendix B provides the action-encoder and FrameInject broadcasting details. For these baselines, the inherited backbone, including the same cross-attention layer, is trained together with the added action-injection modules.

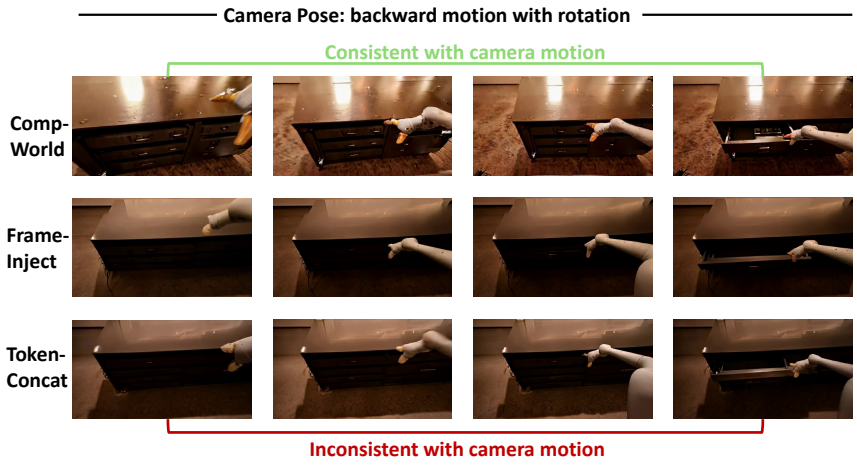

Figure 4: Qualitative compositional action control. In an RT-1-style arm+camera composition, CompWorld follows the commanded camera motion after training on separate action-type annotations, while FrameInject and TokenConcat keep the viewpoint nearly static.

4 Experiments

4.1 Experimental Settings

Compared Methods. We compare CompWorld with the standard injection baselines FrameInject and TokenConcat (Sec. 3.3), which share its backbone and action tokens, including the not-provided placeholders, so the comparison isolates how action tokens condition the backbone. We additionally compare with GameFactory (Yu et al., 2025b), a published action-conditioned video generator reimplemented on the same data and backbone following its original design (discrete actions via cross-attention, continuous actions via channel concatenation).

Implementation Details. We build all methods on top of an internal 1B-parameter text-to-video Diffusion Transformer (Wan2.1-1.3B for Table 4, with its settings in Appendix B). Videos are generated at 640×352 resolution with 77 frames, except RT-1, which is generated and evaluated at its native 320×256 resolution. The causal 3D VAE uses a temporal compression ratio of 4, mapping each clip to 20 latent timesteps. We use no extra context window; for evaluation, the first frame is the visual condition. Training uses batch size 64 on 8 NVIDIA A100 GPUs; each 10,000-step run takes about 40 hours, and all results use the 10,000-step checkpoint. At inference, we apply classifier-free guidance (Ho & Salimans, 2022) to the text condition and use 50 diffusion sampling steps. Unless otherwise stated, baselines share the same backbone and optimization setup.

Evaluation Protocol. We evaluate on held-out clips. To measure viewpoint action following, we reconstruct the camera trajectory of each generated video with GLOMAP (Pan et al., 2024) and define Cam as the L2 distance between the reconstructed and reference trajectories (the ground truth, or the commanded trajectory when no ground truth exists), computed on normalized relative camera matrices; lower is better. Cam is the shared camera-following metric across all settings with viewpoint motion; RT-1 per-subset omits it because its recorded camera is largely static. We additionally report PSNR and LPIPS on paired frames for reconstruction fidelity, and FID and FVD for image and video distributional quality; higher PSNR and lower LPIPS, FID, and FVD are preferred. Since arm+camera has no ground-truth video and thus no paired metrics, we compensate by decomposing its Cam into rotation error (RotErr) and scale-normalized translation error (TransErr).

4.2 Compositional Action Control Generalization

Keyboard + camera pose composition. This setting directly tests compositional action control: keyboard and camera-pose controls must work together at inference, while paired keyboard-camera annotations are unavailable during training. We construct a converted benchmark that trains on two



Figure 5: Qualitative results under separate action-control settings with CompWorld.

separate sources: GF-Keyboard from GameFactory/GF-Minecraft (Yu et al., 2025b) for keyboard control, and Context-as-Memory (Yu et al., 2025a) for camera poses. At evaluation, we use GF-Keyboard+Mouse clips and convert mouse pitch/yaw into camera-pose inputs, producing paired keyboard and camera-pose controls that never appear together in training. Their recorded videos provide ground truth for Cam and the paired metrics, so we start here. The protocol stresses cross-source composition: the camera branch is trained on Context-as-Memory poses but evaluated on camera-like mouse motion in GF-Minecraft, while the keyboard branch stays in-domain.

Table 1 reports the benchmark: CompWorld attains the best Cam, PSNR, LPIPS, and FID, outperforming FrameInject, TokenConcat, and GameFactory on control adherence, paired fidelity, and image quality. Its FVD is higher than those of TokenConcat and GameFactory, which both follow the joint command less closely (higher Cam). Because FVD compares video distributions without reference to the command, it does not penalize such deviations: GameFactory reaches the lowest FVD despite the worst Cam (31.08). The w/o Fix variant obtains a lower FVD but worse Cam, paired metrics, and FID, so updating Fused-Text Cross-Attention trades control and fidelity for FVD (Sec. 4.4).

Table 1: Compositional action control without paired keyboard-camera annotations.

Method	Cam↓	PSNR↑	LPIPS↓	FID↓	FVD↓
FrameInject	28.07	15.54	0.6307	91.48	843.31
TokenConcat	27.91	15.72	0.6237	89.66	730.16
GameFactory	31.08	15.96	0.6132	87.75	**686.56**
CompWorld	**26.99**	**16.62**	**0.6074**	**84.74**	781.96
CompWorld w/o Fix	27.63	15.64	0.6290	91.82	720.84
CompWorld w/o Placeholder	28.85	16.22	0.6105	91.42	804.63
CompWorld w/o Text-Align	29.60	15.96	0.6169	98.35	838.36

Robot-arm + camera pose composition. This setting tests compositional control across two truly disjoint action types: robot-arm and camera-pose controls must compose at inference, while no paired arm-camera training data is available. We construct 300 clips per method from 50 held-out RT-1 scenes and 6 commanded camera trajectories (5 non-static, 1 static). Prompts contain only the RT-1 instruction, so camera following cannot come from text alone. Since no ground-truth video exists, Cam, RotErr, and TransErr (Sec. 4.1) are measured against the commanded trajectory and paired metrics are omitted; means aggregate over clips that GLOMAP registered for every method. Unlike in RT-1 per-subset, the camera here is explicitly commanded.

Table 2: Robot-arm plus camera-pose compositional control on RT-1 scenes with commanded camera trajectories.

Method	RotErr↓	TransErr↓	Cam↓	FID↓	FVD↓
FrameInject	1.403	10.016	10.469	80.15	754.31
TokenConcat	1.336	11.904	12.351	74.21	745.32
GameFactory	1.225	12.936	13.249	75.74	708.90
CompWorld	**1.150**	**9.601**	**9.940**	**74.08**	**660.10**
CompWorld w/o Text-Align	1.635	13.517	14.023	74.57	666.52

Table 2 reports the arm+camera benchmark: CompWorld attains the lowest RotErr, TransErr, and Cam (9.940 vs. 10.469 for the next-best FrameInject), so its reconstructed camera trajectory best matches the commanded one under simultaneous arm control. CompWorld also achieves the lowest FID and FVD, outperforming FrameInject, TokenConcat, and GameFactory on every metric (see Fig. 4). Appendix E further tests held-out joint keyboard commands within one action type, where CompWorld again leads FrameInject and TokenConcat.

4.3 Comparison Results

Datasets. We evaluate on three action-control subsets: P2P (Yue et al., 2026) (keyboard-and-mouse), SpatialVID (Wang et al., 2025b) (camera pose), and RT-1 (Brohan et al., 2022) (robot arm). We use 8 P2P games for consistent action semantics, resize P2P and SpatialVID to 640×352 and keep RT-1 at 320×256, segment each source into fixed-length clips, and hold out 5% for testing.

Together, these sources test whether one shared interface can be learned from videos annotated with different action types across domains. We report full per-subset results on the internal 1B backbone (Table 3) and Wan2.1-1.3B (Table 4), with qualitative examples in Fig. 5; Appendix D summarizes the merged evaluation over the three subsets on both backbones.

Quantitative Analysis. Beyond the compositional benchmarks, the per-subset results test whether CompWorld serves as a shared action-conditioning interface across heterogeneous action types. On the internal 1B backbone (Table 3), CompWorld attains the best value on every metric across P2P, SpatialVID, and RT-1, beating FrameInject, TokenConcat, and GameFactory. The same interface transfers to Wan2.1-1.3B (Wang et al., 2025a) (Table 4), where CompWorld again leads on most cells across the three subsets, indicating that the text-aligned design is not tied to one video generator. On both backbones, the merged summary in Appendix D shows CompWorld achieving the best value on all five metrics, from camera following (Cam) to video quality (FVD).

Table 3: Internal 1B per-subset evaluation on P2P, SpatialVID, and RT-1.

Method	P2P (Keyboard+Mouse)					SpatialVID (Camera Pose)					RT-1 (Arm)			
	Cam↓	PSNR↑	LPIPS↓	FID↓	FVD↓	Cam↓	PSNR↑	LPIPS↓	FID↓	FVD↓	PSNR↑	LPIPS↓	FID↓	FVD↓
FrameInject	17.74	12.09	0.5280	74.36	1047.40	18.94	11.34	0.6155	87.12	1108.42	17.27	0.3233	60.34	684.54
TokenConcat	15.01	12.96	0.5143	69.50	936.77	8.30	13.66	0.5196	56.01	639.24	18.58	0.2889	54.88	553.80
GameFactory	15.43	13.45	0.5090	75.22	1064.80	10.34	12.91	0.5270	62.84	768.70	18.44	0.3020	56.06	560.30
CompWorld	**14.26**	**13.46**	**0.5086**	**63.77**	**935.85**	**6.17**	**15.26**	**0.4564**	**54.93**	**576.28**	**19.55**	**0.2687**	**52.45**	**481.05**
CompWorld w/o Fix	14.35	12.07	0.5266	73.87	1054.95	7.22	13.45	0.5127	73.83	899.14	19.11	0.2824	56.97	601.18

Table 4: Wan2.1-1.3B per-subset evaluation on P2P, SpatialVID, and RT-1.

Method	P2P (Keyboard+Mouse)					SpatialVID (Camera Pose)					RT-1 (Arm)			
	Cam↓	PSNR↑	LPIPS↓	FID↓	FVD↓	Cam↓	PSNR↑	LPIPS↓	FID↓	FVD↓	PSNR↑	LPIPS↓	FID↓	FVD↓
FrameInject	16.14	10.82	0.6348	109.74	1496.54	13.13	10.86	0.6186	86.30	1000.70	12.61	0.4046	84.62	1143.68
TokenConcat	15.11	11.24	0.5658	89.18	1314.09	13.28	11.40	0.5831	67.56	954.82	15.39	0.3216	63.20	805.25
GameFactory	14.66	11.25	0.5530	84.69	1223.60	15.04	11.37	**0.5670**	68.60	882.20	**17.13**	**0.2950**	51.82	**706.70**
CompWorld	**14.44**	**11.98**	**0.5356**	**81.84**	**1063.38**	**12.78**	**11.79**	0.5713	**62.52**	**861.18**	16.62	0.2987	**49.90**	710.57
CompWorld w/o Fix	15.20	11.14	0.5666	91.76	1283.63	13.83	11.34	0.5866	72.75	1044.59	14.65	0.3431	67.55	867.84

4.4 Ablation Study

Effect of aligning with the pretrained text space. To isolate whether text-space alignment matters—as opposed to any shared latent sufficing—we compare against CompWorld w/o Text-Align, which routes action tokens through a parameter-matched cross-attention over a randomly initialized latent instead of the pretrained text-conditioning pathway. Cam degrades from 26.99 to 29.60 on keyboard+camera (Table 1) and from 9.940 to 14.023 on arm+camera (Table 2), with RotErr and TransErr degrading similarly, while FID and FVD on arm+camera stay close (74.57/666.52 vs. 74.08/660.10). There, the drop is on control adherence, not visual quality: alignment with the pretrained text space, not a shared latent alone, drives the gains.

Effect of preserving the pretrained text-space interface. Keeping Fused-Text Cross-Attention fixed lets action conditioning enter through the pretrained text-space interface. This ablation tests

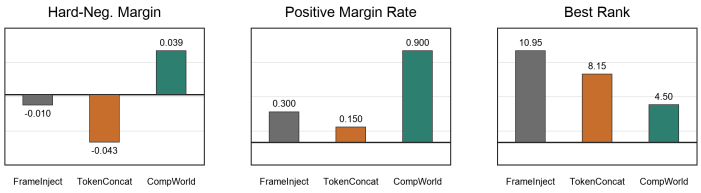

Figure 6: **Interface semantic alignment analysis.** Bars report semantic text-pooling metrics; CompWorld yields positive margins and a lower best rank, while FrameInject and TokenConcat sit at negative or near-zero margins.

whether the pretrained interface should remain a stable conditioning pathway or adapt together with the new action modules. CompWorld w/o Fix updates this interface: it lowers FVD on keyboard+camera (720.84 vs. 781.96, Table 1) but worsens every metric on the per-subset tables (Tables 3, 4); for example, SpatialVID FVD rises from 576.28 to 899.14 on the internal backbone, so we keep it frozen.

Effect of not-provided placeholders. The "not provided" placeholder gives each absent action slot an explicit state in the shared interface. Removing it degrades every metric in Table 1, raising Cam from 26.99 to 28.85 and FID from 84.74 to 91.42, so representing unavailable action types explicitly beats leaving them implicit.

4.5 Interface Semantic Alignment Analysis

We further probe whether the action effect consumed by the denoising backbone is semantically aligned with the text-conditioning interface, comparing the three conditioning interfaces that share our action tokens (Sec. 4.1). Such alignment is what text-space fusion is meant to provide: the effect of each action on generation should be organized through the text-conditioning channel. The probe diagnoses semantic accessibility in each method's consumed interface, not raw action decodability or an extra action-following metric: it asks whether pretrained text anchors rank the intended action effect above confusable alternatives. On the keyboard and camera-pose cases from Sec. 4.2, we run text-anchor retrieval and counterfactual sensitivity on each interface (details in Appendix F). Fig. 6 and Table 5 show that CompWorld achieves higher Top-3 retrieval, a lower best-anchor rank, and positive hard-negative margins and CF Gain.

Table 5: **Action-effect interface comparison.**

Method	Top-3↑	Best Rank↓	Hard-Neg.↑	Pos.↑	CF Gain↑
FrameInject	0.200	10.95	-0.0096	0.300	-0.0152
TokenConcat	0.200	8.15	-0.0428	0.150	-0.0112
CompWorld	**0.550**	**4.50**	**0.0391**	**0.900**	**0.0106**

5 Conclusion

We presented CompWorld, which projects heterogeneous action streams into a text-aligned unified action space using the pretrained text space as a shared interface prior, so keyboard, mouse, camera-pose, and robot-arm controls can be composed at inference even when trained on separate sources. CompWorld improves compositional action control on both compositional benchmarks, action following and video quality on separate action-type subsets across two backbones, and yields a more semantically aligned conditioning interface. These results position the pretrained text space as an interface prior for compositional action control, beyond conventional language prompts.

Limitations and broader impact. CompWorld reduces the need for paired multi-action supervision but not for richer interactive data. Because its shared interface prior depends on the pretrained text-to-video model, fine-grained high-frequency controls remain hard; on arm+camera, RT-1 has no executable environment, so our metrics report trajectory consistency rather than task success. Future work should study richer paired data and action-specific adherence metrics. Controllable video world models can support useful simulators but may also enable deceptive interactive media.

AI USE STATEMENT

In this work, we used generative AI tools to edit the paper for readability. We have not used generative AI tools for any task with required disclosure, such as proposing hypotheses, designing or implementing methods and experiments, or interpreting results, and the remaining required-disclosure tasks, such as writing proofs, are not applicable to this work. We have reviewed all AI-assisted work: every edited passage was checked by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The anonymous source code for data processing, training, and evaluation in the public Wan2.1-1.3B setting is available in the supplementary material; since our main results use an internal backbone, we additionally report comparisons on that public backbone (Wang et al., 2025a) in Table 4 and Appendix D. The backbone, optimization setup, and metric definitions behind all reported numbers are given in Sec. 4.1, with the Wan2.1-1.3B settings in Appendix B, and the training procedure, the action type-specific encoders, and the dataset sources and licenses are given in Appendices A–C. The two non-standard protocols, the held-out action-combination splits and the interface alignment analysis, are specified in Appendix E and Appendix F, respectively.

REFERENCES

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in Atari. In *Thirty-eighth Conference on Neural Information Processing Systems*, 2024.

Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Xiao Fu, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints, 2024. URL https://arxiv.org/abs/2412.07760.

Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. *arXiv preprint arXiv:2405.04233*, 2024.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *International Conference on Machine Learning*, 2024.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7310–7320, 2024.

Decart, Julian Quevedo, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer. https://oasis-model.github.io/, 2024.

Google DeepMind. Veo 2: Our state-of-the-art video generation model. https://deepmind.google/technologies/veo/veo-2/, 2024.

Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. *arXiv preprint arXiv:2412.14169*, 2024.

Ria Doshi, Homer Walke, Oier Mees, Sudeep Dasari, and Sergey Levine. Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation. In *Conference on Robot Learning*, 2024.

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

Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *arXiv preprint arXiv:2109.13396*, 2021.

Xiao Fu, Xian Liu, Xintao Wang, Sida Peng, Menghan Xia, Xiaoyu Shi, Ziyang Yuan, Pengfei Wan, Di Zhang, and Dahua Lin. 3dtrajmaster: Mastering 3d trajectory for multi-entity motion in video generation. In *ICLR*, 2025.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *arXiv preprint arXiv:2405.17398*, 2024.

Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang, and Chuang Gan. Adaworld: Learning adaptable world models with latent actions. *arXiv preprint arXiv:2503.18938*, 2025.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision*, pp. 330–348. Springer, 2025.

William H Guss, Brandon Houghton, Nicholay Topin, Phillip Wang, Cayden Codel, Manuela Veloso, and Ruslan Salakhutdinov. Minerl: A large-scale dataset of minecraft demonstrations. *arXiv preprint arXiv:1907.13440*, 2019.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in Neural Information Processing Systems*, 31, 2018.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. In *International Conference on Machine Learning*, 2023.

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. Vima: General robot manipulation with multimodal prompts. *arXiv preprint arXiv:2210.03094*, 2(3):6, 2022.

Seung Wook Kim, Jonah Philion, Antonio Torralba, and Sanja Fidler. Drivegan: Towards a controllable high-quality neural simulation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5820–5829, 2021.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

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 arXiv:2506.17201*, 2025.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, 2022.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, pp. 360–378. Springer, 2024.

Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, Tong He, Jiangmiao Pang, Yu Qiao, and Kaipeng Zhang. Yume-1.5: A text-controlled interactive world generation model. *arXiv preprint arXiv:2512.22096*, 2025.

Antonio Montanaro, Luca Savant Aira, Emanuele Aiello, Diego Valsesia, and Enrico Magli. Motioncraft: Physics-based zero-shot video generation. *Advances in Neural Information Processing Systems*, 37:123155–123181, 2024.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

NVIDIA. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Robotics: Science and Systems*, 2024.

Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes Lutz Schönberger. Global Structure-from-Motion Revisited. In *European Conference on Computer Vision (ECCV)*, 2024.

Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, Stephen Spencer, Jessica Yung, et al. Genie 2: A large-scale foundation world model. `https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/`, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Team Seawead, Ceyuan Yang, Zhijie Lin, Yang Zhao, Shanchuan Lin, Zhibei Ma, Haoyuan Guo, Hao Chen, Lu Qi, Sen Wang, et al. Seaweed-7b: Cost-effective training of video generation foundation model. *arXiv preprint arXiv:2504.08685*, 2025.

The Movie Gen Team. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025a.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, et al. Spatialvid: A large-scale video dataset with spatial annotations. *arXiv preprint arXiv:2509.09676*, 2025b.

Lirui Wang, Xinlei Chen, Jialiang Zhao, and Kaiming He. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. In *Advances in Neural Information Processing Systems*, 2024a.

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 *Advances in Neural Information Processing Systems*, 2023.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024b.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, 2024c.

Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *International Conference on Learning Representations*, 2024a.

Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024b.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos. In *International Conference on Learning Representations*, 2025.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025a.

Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11590–11599, 2025b.

Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

Yuguang Yue, Irakli Salia, Samuel Hunt, Chris Green, Wenzhe Shi, and Jonathan J Hunt. Scaling behavior cloning improves causal reasoning: An open model for real-time video game playing. *arXiv preprint arXiv:2601.04575*, 2026.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T Freeman. Physdreamer: Physics-based interaction with 3d objects via video generation. In *European Conference on Computer Vision*, pp. 388–406. Springer, 2024.

Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Fei Kang, Biao Jiang, Zedong Gao, Eric Li, Yang Liu, and Yahui Zhou. Matrix-Game: Interactive world foundation model. *arXiv preprint arXiv:2506.18701*, 2025.

Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K. Wong. Uni-ControlNet: All-in-one control to text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024a.

Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. Robodreamer: Learning compositional world models for robot imagination. *arXiv preprint arXiv:2404.12377*, 2024b.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Irasim: Learning interactive real-robot action simulators. *arXiv preprint arXiv:2406.14540*, 2024.

A Training Algorithm

Algorithm 1 highlights how a video with only a subset of action annotations is completed into fixed action slots before entering the shared text-space interface.

Algorithm 1: Training with Disjoint Action Sets and Placeholder Completion

Input: Dataset $\mathcal{D}$ of videos, prompts, heterogeneous action annotations and action-window length w; action type set $\mathcal{M}$; learnable placeholder embedding $\mathbf{u}_m^{\text{miss}}$ for each $m \in \mathcal{M}$

while *not converged* **do**

- Sample $(\mathbf{X}, c, \{\mathbf{A}_m\}_{m\in\mathcal{S}})$ from $\mathcal{D}$;
- Encode video and text: $\mathbf{Z}_0 \leftarrow \mathcal{E}_v(\mathbf{X})$, $\mathbf{Y} \leftarrow \mathcal{E}_{\text{text}}(c)$;
- **foreach** $m \in \mathcal{M}$ **do**
  - **if** $m \in \mathcal{S}$ **then**
    - Align and encode $\mathbf{A}_m$: $\mathbf{U}_m \leftarrow \mathcal{E}_{a,m}(\mathbf{A}_m)$;
  - **else**
    - Use placeholder tokens: $\mathbf{U}_m \leftarrow \mathrm{Repeat}(\mathbf{u}_m^{\text{miss}}, n+1, w)$;
- For each latent timestep τ, concatenate action tokens in fixed slot order: $\mathbf{U}_\tau \leftarrow \mathrm{Concat}_{m\in\mathcal{M}}(\mathbf{U}_{m,\tau})$;
- Fuse actions into the pretrained text-space interface: $\mathbf{Y}_\tau^{\text{fuse}} \leftarrow \mathrm{CrossAttn}_{\text{A2T}}(\mathbf{Y}, \mathbf{U}_\tau, \mathbf{U}_\tau) + \mathbf{Y}$ for each τ;
- Sample $t \sim \mathcal{U}(1, T)$ and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, then obtain $\mathbf{Z}_t$;
- Optimize the denoiser using Fused-Text Cross-Attention with $\{\mathbf{Y}_\tau^{\text{fuse}}\}_{\tau=0}^n$:
- $\mathcal{L} = \left\| \epsilon_\theta(\mathbf{Z}_t, t, \{\mathbf{Y}_\tau^{\text{fuse}}\}_{\tau=0}^n) - \epsilon \right\|_2^2$;

B Implementation Details for Action Interfaces

Action encoders. Each action type produces one token sequence $\mathbf{U}_m \in \mathbb{R}^{(n+1)\times w\times d}$, matching the notation in Sec. 3.2. For latent timestep τ, the encoder consumes the local raw-action window aligned to the corresponding video interval and preserves the window dimension as w action tokens. Keyboard states are represented as multi-hot vectors and projected to w tokens with dimension d. Continuous mouse, camera-pose, and robot-arm actions use the same temporal windowing, with their raw action vectors mapped to w text-dimensional action tokens by the corresponding action type-specific projection head. If action type m is not provided, its slot is filled by repeating the learned placeholder $\mathbf{u}_m^{\text{miss}}$ over all $(n+1)\times w$ latent-window positions.

FrameInject broadcasting. FrameInject uses the same temporally aligned action tokens as CompWorld, but adds them to the visual stream instead of first fusing them into text tokens. Let $\mathbf{H}_{\tau,s}^{(\ell)}$ denote the visual token at latent timestep τ and spatial index s. For each τ, FrameInject concatenates the action tokens from all fixed slots at the same temporal index and applies a layer-specific projection P_ℓ, including the flattening or pooling over the slot and window tokens:

$$\mathbf{b}_\tau^{(\ell)} = P_\ell(\mathrm{Concat}_{m\in\mathcal{M}}\, \mathbf{U}_{m,\tau})\,.$$

The projected action vector is then broadcast only over the spatial tokens of the same latent timestep:

$$\tilde{\mathbf{H}}_{\tau,s}^{(\ell)} = \mathbf{H}_{\tau,s}^{(\ell)} + \mathbf{b}_\tau^{(\ell)}.$$

Thus, FrameInject keeps the same fixed action slots and latent-timestep alignment as CompWorld, while differing only in where the action information is injected into the video backbone.

Wan2.1-1.3B setting. The comparisons in Table 4 and Appendix D keep the resolutions and first-frame conditioning of Sec. 4.1 but follow the native clip length of Wan2.1-1.3B: all methods are trained and evaluated on 81-frame clips (21 latent timesteps). Training uses AdamW with a constant learning rate of 2×10^{-5} and a per-GPU batch size of 1 on 8 GPUs, and all results use the 10,000-step checkpoint. At inference, we use classifier-free guidance scale 5.0 and 50 sampling steps.

C Dataset Details

This appendix summarizes the datasets used in our experiments, focusing on the action annotations each source contributes and how each source supports the disjoint heterogeneous-control setting.

GameFactory / GF-Minecraft. GameFactory (Yu et al., 2025b) provides Minecraft gameplay videos with keyboard and mouse metadata, including movement keys, special control states, mouse-related pitch/yaw changes, and player position. Its randomized action collection is useful for studying rare or held-out keyboard/mouse combinations. We use GF-Minecraft in the compositional benchmark as the keyboard-control source and in the held-out combination analysis as a within-dataset stress test.

Context-as-Memory. Context-as-Memory (Yu et al., 2025a) studies scene-consistent interactive long video generation with memory retrieval, and provides long videos with per-frame camera poses. In our compositional control benchmark, it is used as the camera-pose supervision source. Evaluation is then conducted on GF-Minecraft keyboard-and-mouse clips whose mouse motion is converted into camera-pose inputs, so the model is tested on keyboard and camera-pose composition without paired keyboard-camera annotations during training.

P2P. Pixels2Play (P2P) (Yue et al., 2026) contains human gameplay across many 3D games with frame-aligned keyboard and mouse actions. Its annotations include keyboard states, mouse movement and buttons, and frame-level text when available. We use P2P as the keyboard-and-mouse source in the separate action-control evaluation because it provides dense low-level action supervision aligned with visual observations.

SpatialVID. SpatialVID (Wang et al., 2025b) is a large-scale in-the-wild video dataset with spatial annotations, including camera poses, depth maps, dynamic masks, structured captions, and serialized camera-motion instructions. The pose annotations provide camera translation and orientation, while the instructions describe camera movements such as dolly, truck, pedestal, tilt, pan, roll, or stay. We use SpatialVID as the camera-pose source in the separate action-control evaluation.

RT-1. RT-1 (Brohan et al., 2022) is a real-world robotics dataset and model for large-scale language-conditioned robot control. Its demonstrations include natural language instructions and robot actions covering arm movement, gripper control, and base movement. In our experiments, RT-1 is used as the robot-arm control source, with its recorded robot actions providing supervision for the robot-arm action type in a domain that is visually and mechanically distinct from gameplay and camera-motion data.

Asset licenses and terms. All external datasets and model assets used in our experiments are credited above and used according to their providers' terms. P2P/Open Pixel2Play reports an MIT license; SpatialVID is released under CC-BY-NC-SA-4.0; Wan2.1-1.3B is released under Apache-2.0; and the Robotics Transformer code associated with RT-1 is released under the Apache license. For GameFactory/GF-Minecraft, Context-as-Memory, and RT-1 data access, we follow the terms specified by the corresponding project or dataset providers.

D Merged Quantitative Results

Table 6 reports the merged evaluation over P2P, SpatialVID, and RT-1. The main paper presents the full per-subset tables for the internal 1B and Wan2.1-1.3B backbones; this appendix table provides the compact aggregate view used for cross-backbone comparisons.

E Held-Out Action Combinations

Beyond composing different action types, we test whether CompWorld generalizes to held-out joint commands within one action type. This is a complementary within-dataset stress test rather than a replacement for the cross-type compositional benchmarks in Sec. 4.2. We use GF-Minecraft (Yu et al., 2025b) because its randomized keyboard-and-mouse collection reduces obvious human-play bias and supports removing specific action combinations.

We construct two video-level splits with different related-combination evidence. In the first, Forward+Left videos are removed while nearby WASD pairs such as Forward+Right, Backward+Left, and Backward+Right remain, testing transfer from observed combinations to the held-out one. In the

Table 6: Merged evaluation on the internal 1B model and Wan2.1-1.3B. Results aggregate P2P, SpatialVID, and RT-1; CompWorld improves aggregate fidelity and quality metrics on both backbones, suggesting that the action-to-text interface is not tied to one generator.

Backbone	Method	Cam↓	PSNR↑	LPIPS↓	FID↓	FVD↓
Internal 1B						
	FrameInject	18.41	13.56	0.4890	51.79	529.04
	TokenConcat	11.24	15.06	0.4409	41.48	370.28
	GameFactory	12.54	14.93	0.4460	46.01	447.40
	CompWorld	**9.69**	**16.09**	**0.4112**	**39.61**	**363.00**
	CompWorld w/o Fix	10.35	14.88	0.4406	48.08	471.48
Wan2.1-1.3B						
	FrameInject	14.50	11.43	0.5526	66.37	739.15
	TokenConcat	13.95	12.68	0.4901	51.87	612.16
	GameFactory	14.87	13.25	0.4710	48.10	531.00
	CompWorld	**13.37**	**13.47**	**0.4685**	**44.88**	**485.32**
	CompWorld w/o Fix	14.39	12.38	0.4988	54.64	642.20

second, all pairwise WASD combinations are filtered from training, so training contains no videos with sustained simultaneous WASD direction pairs.

To avoid treating incidental co-occurrences as combination supervision, a video contains Forward+Left only when more than 7 frames have that command; otherwise it is treated as not containing the target. This threshold keeps enough training data while removing sustained occurrences of the target combination. The test set contains clips with at least 70 Forward+Left frames.

Table 7 reports both splits. CompWorld leads FrameInject and TokenConcat on every metric in both, suggesting that the text-aligned interface reuses evidence from nearby combinations or single-action trajectories when the target joint command is unseen.

Table 7: Held-out action combination generalization on GF-Minecraft for the two splits described above.

(a) Similar WASD Pairs Observed

Method	Cam↓	PSNR↑	LPIPS↓	FID↓	FVD↓
FrameInject	10.12	19.79	0.4467	42.43	494.98
TokenConcat	10.24	20.00	0.4462	42.56	456.70
CompWorld	**10.05**	**20.04**	**0.4433**	**39.68**	**440.93**

(b) All WASD Pairs Filtered

Method	Cam↓	PSNR↑	LPIPS↓	FID↓	FVD↓
FrameInject	11.73	19.55	0.4520	46.62	533.62
TokenConcat	11.22	19.49	0.4553	44.39	516.72
CompWorld	**10.96**	**19.61**	**0.4478**	**43.62**	**454.57**

F Interface Semantic Alignment Analysis Details

This section gives the protocol for Sec. 4.5. The diagnostic evaluates whether the conditioning interface consumed by the denoising backbone is accessible to pretrained text anchors; it complements, but does not replace, direct action-following metrics. The analysis compares the representation interfaces consumed by the denoising backbone, rather than arbitrary intermediate embeddings. We intentionally evaluate each method in its native consumed interface without training an additional projector into text space; such a projector would measure post-hoc probeability rather than whether the conditioning signal is directly accessible to pretrained text-space semantics. Let c denote the text prompt, a an action input, and a_0 the corresponding idle or no-op action. For CompWorld, we measure the action effect as

$$\Delta_{\text{CompWorld}}(c, a) = \text{Pool}\left(F(c, a) - T(c)\right),$$

where $T(c)$ is the base text condition broadcast over latent timesteps and $F(c, a)$ is the timestep-specific fused text condition after Action-to-Text Fusion. For TokenConcat, we use the key-space effect of the appended action tokens,

$$\Delta_{\text{TokenConcat}}(a) = \text{Pool}\left(K(a) - K(a_0)\right),$$

where $K(\cdot)$ denotes the key projection consumed by the same cross-attention layer. For FrameInject, we instrument the visual-side action interface and compute

$$\Delta^{s}_{\text{FrameInject}}(a) = \text{Pool}\left(G_s(a) - G_s(a_0)\right),$$

where $s \in \{A, B, C\}$ indexes the traced stage. Stage A denotes the action-side interface before visual broadcast, stage B denotes the signal after visual addition, and stage C denotes the signal after the first temporal mixing operation. The main paper reports stage A as the direct FrameInject interface; the full A/B/C tracing is shown in Table 8.

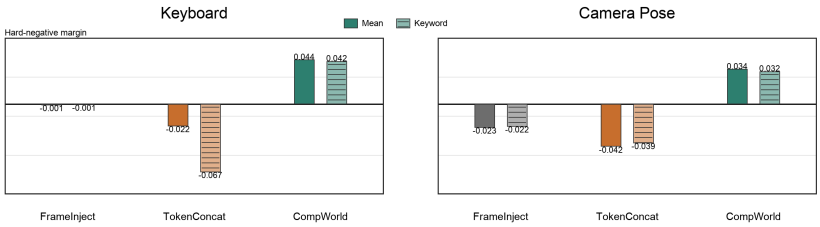

Figure 7: Action-effect retrieval breakdown by action type. We report the hard-negative margin under mean pooling and keyword-span pooling. CompWorld achieves positive margins for both keyboard and camera-pose actions, indicating more direct accessibility to pretrained text anchors.

Table 8: Full action-effect interface comparison. The main semantic readout and all-pooling robustness indicate that CompWorld exposes action effects more directly to pretrained text anchors; FrameInject stage C shows how the visual-side signal changes after temporal mixing. CF Cons. denotes CF Consistency.

Method	Interface	Pooling	Top-3↑	Best Rank↓	Hard-Neg. Margin↑	Pos. Rate↑	CF Cons.↑	CF Gain↑
Main semantic readout								
FrameInject	A: action interface	mean+keyword	0.200	10.95	-0.0096	0.300	0.500	-0.0152
FrameInject	B: after visual addition	mean+keyword	0.200	10.95	-0.0096	0.300	0.500	-0.0152
FrameInject	C: after temporal mixing	mean+keyword	0.100	7.90	-0.0154	0.250	0.500	-0.0172
TokenConcat	action-token key effect	mean+keyword	0.200	8.15	-0.0428	0.150	0.500	-0.0112
CompWorld	fused-minus-base text effect	mean+keyword	**0.550**	**4.50**	**0.0391**	**0.900**	0.500	**0.0106**
All-pooling robustness								
FrameInject	A: action interface	all	0.300	7.67	-0.0047	0.450	0.500	-0.0083
FrameInject	B: after visual addition	all	0.300	7.67	-0.0047	0.450	0.500	-0.0083
FrameInject	C: after temporal mixing	all	0.200	8.45	-0.0090	0.225	0.250	-0.0112
TokenConcat	action-token key effect	all	0.150	8.13	-0.0219	0.325	0.375	-0.0059
CompWorld	fused-minus-base text effect	all	**0.375**	**5.70**	**0.0196**	**0.650**	**0.625**	**0.0070**

Text-anchor retrieval. For each canonical action case, we construct a small set of short text anchors covering the positive action description, related hard negatives, and unrelated motions. We embed these anchors with the same text encoder used by the backbone and rank them by cosine similarity to the measured action effect. This probe evaluates semantic accessibility of each native conditioning interface, not direct action following or calibrated hidden-state quality. We report Top-3 hit rate, the rank of the best positive anchor, and Hard-Negative Margin, defined as the similarity of the best positive anchor minus the strongest hard negative. A positive margin means that the interface effect is closer to the intended action semantics than to the most confusable alternative. Positive Margin Rate (Pos.) is the fraction of action cases with a positive Hard-Negative Margin.

Counterfactual sensitivity. We further evaluate paired action changes such as left versus right, forward versus idle, and looking up versus looking down. For each pair, we compute the target-minus-opposite text-anchor margin before and after replacing the action with its counterfactual. CF Consistency measures the fraction of cases where this margin changes in the expected direction, and CF Gain measures the average target-over-counterfactual margin gain.

Paired metric sweep. In addition to the absolute summary, the analysis script reports paired CompWorld-over-baseline deltas on matched pooling, action type, and case items. We report the mean+keyword semantic pooling subset in the main paper because it most directly reads action-anchor content; all-pooling results are included as robustness checks.