FORCEVID3D: PHYSICAL VIDEO GENERATION WITH CAMERA-SPACE 3D FORCE CONTROL

Anonymous authors
Paper under double-blind review



Figure 1: **Showcase of ForceVid3D under camera-space 3D force control.** Top: counterfactual generations from the same scene under different 3D force directions, including toward- and away-from-camera components. Bottom: representative scene-dependent responses to localized point forces and distributed wind, including off-center rigid-body motion, pendulum motion, planar sliding, flexible deformation, and plume advection.

ABSTRACT

Force-controlled video generation allows users to specify external forces while the model generates scene-dependent responses. Existing direct force-control methods typically rely on image-plane 2D representations, which do not fully express camera-space 3D forces, while 2D observations provide limited cues about responsive regions and their local geometry. We introduce **ForceVid3D**, a framework for generating videos from an initial image and an explicit camera-space 3D force. Our Projective 3D Force Map exposes the image-space geometry of camera-space 3D forces by combining force support, direction, and magnitude with projection cues. We further introduce Depth-Grounded Response Affinity Supervision (D-RAS), which combines response-affinity and local relative-depth objectives to capture responsive regions and their geometric context. To support learning and evaluation across diverse physical responses, we construct ForceVid3D-Dataset, a 60K-pair force–video dataset spanning seven response groups, and introduce ForceVid3D-Bench for force-to-response evaluation. Under comparable 2D force-control settings, ForceVid3D achieves consistent improvements in force adherence and physical consistency over prior force-conditioned methods while maintaining competitive visual quality. Dedicated 3D evaluations further demonstrate camera-space 3D force control, while controlled ablations support the contributions of the projective representation and D-RAS.

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.

1 INTRODUCTION

Controllable video generation supports text, trajectory, drag, motion-field, and camera interfaces (Yin et al., 2023; Wu et al., 2024; Geng et al., 2025; Wang et al., 2024). Force control differs from motion-target interfaces in that it specifies an external cause rather than a desired pixel motion: the visible response must emerge from scene geometry, support, material, contact, and the pretrained model's physical prior. Recent methods support point forces, wind, continuous inputs, and force-conditioned goals (Gillman et al., 2025; Wang et al., 2026; Gillman et al., 2026). However, direct methods typically encode forces in the image plane, while richer 3D approaches often rely on simulation, reconstructed scenes, or intermediate trajectories (Liu et al., 2024; Zhang et al., 2024; Wang et al., 2025; Liu et al., 2026). We study a direct setting in which an initial image and an explicit camera-space 3D force are mapped to a video response without reconstructing a 3D scene or running a physics simulator at inference.

This task poses two coupled challenges. First, the force representation must preserve 3D information in image-aligned features; simply appending a depth component leaves its view-dependent projection implicit. Second, force location alone does not identify responsive content, which depends on geometry, depth, support, attachment, and object constraints only partially visible in one image. A plausible video may therefore move in the wrong direction or activate the wrong region. These ambiguities are related but distinct: one concerns how the intervention should be represented under the camera, while the other concerns which scene content should respond. We therefore treat force geometry and response grounding as separate design problems.

How can camera-space 3D forces be represented for image-aligned video generation? Under perspective projection, a depth component has a location-dependent image-plane effect. Our **Projective 3D Force Map** preserves force support, direction, and magnitude while adding first-order projection cues, without prescribing a target trajectory or optical flow.

How can a 3D force be grounded in responsive scene content? The application region may overlap different surfaces or cover only part of an object. Our **Depth-Grounded Response Affinity Supervision (D-RAS)** uses a frozen-feature probe to identify controller levels where response association is decodable. Response-affinity and local relative-depth supervision then ground these features in responsive content and geometry. Auxiliary heads are removed at inference.

We construct **ForceVid3D-Dataset**, with 60K rendered force–video pairs across seven response groups, from rigid and constrained motion to aerodynamic transport and plume advection. **ForceVid3D-Bench** evaluates force-to-response generation on unseen images, complemented by controlled synthetic cases with simulator references for trajectory and physical analysis. Evaluation covers force adherence, including direction and grounding; physical consistency; and visual quality.

Under comparable 2D force-control settings, ForceVid3D achieves consistent improvements in force adherence and physical consistency over prior force-conditioned methods while maintaining competitive visual quality. Separate 3D evaluations demonstrate camera-space control, and matched ablations support the Projective 3D Force Map and D-RAS. Our contributions are:

- We introduce ForceVid3D, a framework for direct video generation under explicit camera-space 3D force control.
- We propose a Projective 3D Force Map that preserves the original 3D force condition while exposing its perspective geometry to an image-aligned video generator.
- We introduce D-RAS, training-time supervision that combines response affinity with local relative depth to associate 3D force interventions with responsive scene content and its geometric context.
- We construct a 60K-pair force–video dataset spanning diverse physical responses and ForceVid3D-Bench, an unseen-image benchmark for evaluating force-to-response generation.

2 RELATED WORK

Controllable video generation. Video generators support structural maps, dense motion, trajectories, drag, appearance, and camera motion (Chen et al., 2023; Wang et al., 2023; Yin et al., 2023; Wu et al., 2024; Geng et al., 2025; Wang et al., 2024; He et al., 2025). These interfaces typically specify desired motion, spatial correspondence, or viewpoint changes. Force conditioning instead specifies

an external intervention, allowing the response to emerge from its interaction with the observed scene.

Force- and physics-conditioned video generation. Recent methods support point forces and wind (Gillman et al., 2025), continuous force inputs (Wang et al., 2026), and force-conditioned goals (Gillman et al., 2026), but most direct interfaces remain image-plane controls. PhyParam (Li et al., 2026) conditions rigid-body video generation on object masks, 3D forces, and physical parameters. ForceVid3D combines a projective camera-space force representation with response-affinity and relative-depth supervision, supporting localized and distributed interventions across rigid motion, flexible deformation, and plume advection. Other approaches use rigid-body simulation, reconstructed 3D scenes, or generated 3D trajectories before video synthesis (Liu et al., 2024; Zhang et al., 2024; Wang et al., 2025; Liu et al., 2026). General video benchmarks assess perceptual quality or physical commonsense (Huang et al., 2024; Bansal et al., 2025; Meng et al., 2024; Motamed et al., 2026), but not adherence to a specified force. ForceVid3D directly conditions generation on camera-space 3D forces and evaluates force-to-response behavior without explicit scene reconstruction or physics simulation at inference.

3 METHOD

3.1 OVERVIEW AND PROBLEM SETUP

ForceVid3D generates a scene-dependent video from an initial image, text, and a camera-space 3D force (Figure 2). The Projective 3D Force Map represents the intervention (Section 3.2), while D-RAS grounds it in responsive content and local geometry (Section 3.3).

Task formulation. Given an initial frame I_0, text condition c, and force condition

$$\mathcal{F} = (\mathcal{R}, \mathbf{d}, m, t_s, t_e), \tag{1}$$

our goal is to generate an RGB video V depicting the scene response induced by $\mathcal{F}$. Here, $\mathcal{R}$ denotes the image-space support where the intervention is introduced, $\mathbf{d} = (d_x, d_y, d_z)$ is a unit direction in camera coordinates, $m \in [0, 1]$ is the normalized magnitude, and $[t_s, t_e)$ is the active interval. We use $+x$ for image right, $+y$ for image down, and $+z$ away from the camera. Local forces act at a point or compact region, whereas distributed forces such as wind may act over an extended region. The condition specifies the intervention rather than a future trajectory, optical flow, or target displacement.

Architecture. The force condition is rasterized over space and time at the video latent resolution. A 10-block control branch, initialized from the pretrained backbone, processes the force condition with the noisy video latent and injects residual features into the corresponding frozen backbone blocks through zero-initialized projections (Zhang et al., 2023). The Projective 3D Force Map provides the force representation, while D-RAS supplies training-time supervision over responsive content and local relative depth.

Training stages. We first train the force controller with supervised flow matching, including an additional object-region term for masked simulator samples, while keeping the video backbone frozen. We then use frozen-feature response probes to select D-RAS levels and continue controller training with the joint flow-matching and D-RAS objectives; all D-RAS modules are removed at inference.

3.2 PROJECTIVE 3D FORCE MAP

Direct 3D force encoding. We first extend pixel-aligned force conditioning (Gillman et al., 2025; Wang et al., 2026) to camera-space 3D directions. At image location q, the support channel $\rho(q)$ is a Gaussian centered at the application point for a local force and is spatially uniform for wind. The camera-space unit direction $\mathbf{d} = (d_x, d_y, d_z)$ and normalized magnitude $m \in [0, 1]$ are broadcast spatially, with $\widetilde{m} = 2m - 1$:

$$C_{\text{direct}}(q) = [\rho(q), d_x, d_y, d_z, \widetilde{m}]. \tag{2}$$

The condition is active only over the specified force interval $[t_s, t_e)$.

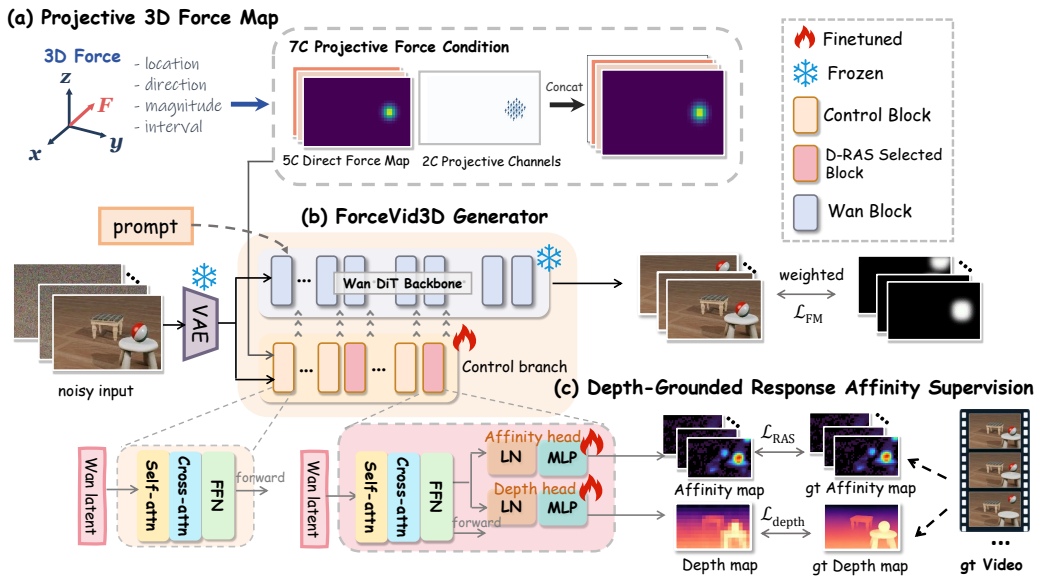
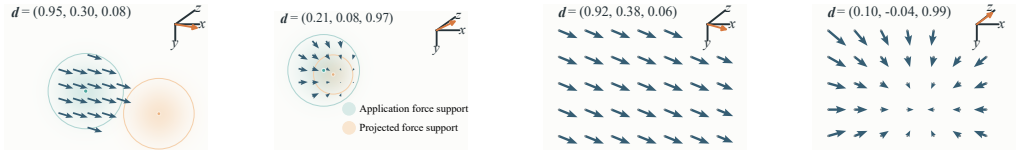


Figure 2: **ForceVid3D overview.** (a) The Projective 3D Force Map augments a direct camera-space force encoding with image-space projection cues. (b) A trainable control branch processes the force condition and injects force-conditioned residuals into a frozen video backbone. (c) D-RAS applies independent response and relative-depth heads to selected controller levels and averages their predictions during training. All D-RAS modules are removed at inference. $\mathcal{L}_{\text{FM}}$, $\mathcal{L}_{\text{RAS}}$, and $\mathcal{L}_{\text{depth}}$ denote the flow-matching, response-affinity, and relative-depth training losses, respectively.



(a) Point, image-plane-dominant **(b)** Point, depth-dominant **(c)** Wind, image-plane-dominant **(d)** Wind, depth-dominant

Figure 3: Projective cues for point/wind forces under image-plane/depth-dominant directions.

Projective cues. Although C_{direct} preserves the 3D force vector, it leaves its image-space geometry implicit. For image coordinate $q = (u, v)$, we use normalized camera coordinates $x(q) = (u - c_x)/f_x$ and $y(q) = (v - c_y)/f_y$, where (c_x, c_y) is the principal point and (f_x, f_y) are the focal lengths. Under perspective projection, an infinitesimal 3D displacement along $\mathbf{d}$ induces a normalized image displacement proportional to

$$U_{\mathcal{F}}(q) = \big(d_x - x(q)d_z,\ d_y - y(q)d_z\big), \tag{3}$$

up to a positive depth-dependent scale. Image-plane directions therefore produce approximately translation-like cues, whereas directions with a depth component produce spatially varying cues under perspective projection. We use $U_{\mathcal{F}}$ only as a geometric conditioning signal rather than a prediction of the resulting motion.

Final force condition. The two projective channels are gated by the force support and concatenated with the direct encoding, while the direct direction and magnitude channels remain spatially broadcast. Let $\widetilde{U}_{\mathcal{F}}(q)$ denote the support-gated version of $U_{\mathcal{F}}(q)$. The resulting seven-channel representation is

$$C_{\text{proj}}(q) = [C_{\text{direct}}(q), \widetilde{U}_x(q), \widetilde{U}_y(q)]. \tag{4}$$

The same temporal gate is applied to all seven channels. Figure 3 visualizes representative local and distributed force conditions. Details of support rasterization, temporal binning, projective-channel gating, and camera-intrinsic handling are provided in Appendix A.4.

3.3 DEPTH-GROUNDED RESPONSE AFFINITY SUPERVISION

The Projective 3D Force Map exposes intervention geometry but does not determine which visible content should respond. Building on the spatial semantics encoded by diffusion features (Baranchuk et al., 2021; Tang et al., 2023), D-RAS supervises intermediate controller representations with response association and local geometry.

Multi-level response prediction. Frozen-feature response probes select a small set of controller levels $\mathcal{S}$. For each selected level ℓ, independent response and relative-depth heads operate on its first-frame token slice $H_0^{(\ell)}$. Their predictions are averaged across levels:

$$a = \frac{1}{|\mathcal{S}|}\sum_{\ell\in\mathcal{S}} a^{(\ell)}, \qquad \hat{D} = \frac{1}{|\mathcal{S}|}\sum_{\ell\in\mathcal{S}} \hat{D}^{(\ell)}, \qquad \hat{A} = \sigma(a), \tag{5}$$

where $a^{(\ell)}$ and $\hat{D}^{(\ell)}$ are produced by level-specific heads. Probe calibration and selected levels are detailed in Appendix B.2.

Response and depth supervision. We use two first-frame-aligned supervision targets. Response affinity A^* associates first-frame visible content with the subsequent force response, using simulator instance masks and temporal appearance changes relative to the initial frame. Crucially, A^* is defined on the initial-frame surface: future response evidence is associated back to the corresponding visible content rather than used as a future motion or occupancy target. Relative depth provides local geometric context around the responsive surface. From initial depth D_0, we define

$$D^*(q) = \mathrm{clip}\left(\frac{\log D_0(q) - \mu_{\mathcal{R}^+}}{s_{\Omega_D} + \epsilon}, -\kappa, \kappa\right), \tag{6}$$

where $\mu_{\mathcal{R}^+}$ is the median log depth of responsive surface $\mathcal{R}^+$ and s_{Ω_D} is a robust scale over the surface and its valid-depth neighborhood. This removes absolute scene scale while retaining local front–back structure; Appendix B.1 gives the exact construction.

Training objective. Weighted binary cross-entropy (WBCE) addresses the sparsity of responsive tokens, while soft Dice loss encourages spatial overlap:

$$\begin{aligned}\mathcal{L}_{\text{D-RAS}} = \lambda_A[\mathcal{L}_{\text{WBCE}}(a, A^*) + \mathcal{L}_{\text{Dice}}(\hat{A}, A^*)] \\ + \frac{\lambda_D}{|\Omega_D|}\sum_{q\in\Omega_D} \text{SmoothL1}\big(\hat{D}(q), D^*(q)\big).\end{aligned} \tag{7}$$

The complete objective is $\mathcal{L} = \mathcal{L}_{\text{base}} + \mathcal{L}_{\text{D-RAS}}$, where $\mathcal{L}_{\text{base}}$ is shared flow matching (Lipman et al., 2022) with an additional object-region term for masked simulator samples. D-RAS modules are removed at inference.

4 FORCEVID3D DATASET AND BENCHMARK

4.1 FORCEVID3D-DATASET

Data generation. We construct **ForceVid3D-Dataset**, a synthetic training set containing 59,635 force–video pairs across seven physical response groups spanning rigid, constrained, contact-driven, and distributed dynamics. As illustrated in Figure 4(a), each sample starts from a predefined response template, followed by sampling scene assets, environments, layouts, physical properties, and a camera-space force condition. The resulting dynamics are produced with response-specific simulation, using PyBullet for modes that require rigid-body physics and procedural dynamics for the remaining modes, and rendered in Blender 5.2 as 49-frame videos at 10 FPS and resized to 640×384 for training. Within each response group, we vary object geometry, scene configuration, physical properties, camera settings, and force parameters to obtain diverse realizations of the same response mode. Each rendered video is paired with its force metadata and a text condition generated from the first frame. Figure 4(c) summarizes the dataset composition, with detailed generation settings and annotations provided in Appendix A.

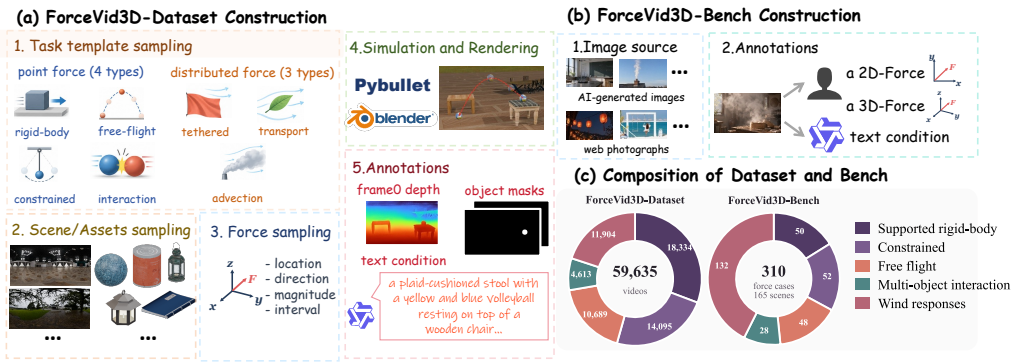


Figure 4: **Construction and composition of ForceVid3D-Dataset and ForceVid3D-Bench.** (a) Training samples combine response templates, scenes, assets, and camera-space forces with simulation and rendering; each video includes depth, masks, and text. (b) Web and AI-generated images receive image-plane and camera-space 3D force annotations. (c) Response-mode composition of the 59,635-video dataset and the 165-image benchmark with 310 evaluation cases.

4.2 ForceVid3D-Bench

Benchmark construction. We construct **ForceVid3D-Bench** from 165 unseen images collected from web photographs and AI-generated scenes, as shown in Figure 4(b). Human annotators specify image-plane and camera-space 3D forces, including the application location and force direction. The resulting benchmark contains 310 force-conditioned evaluation cases, including 156 image-plane conditions and 154 camera-space 3D conditions, across the 165 images. The two control settings are not paired one-to-one across images: some images contain only image-plane or only camera-space 3D conditions, while others contain both. Text conditions are generated from the input image using the same vision-language annotation pipeline. The benchmark further annotates responsive content and explicit non-responsive distractors, together with task-specific motion and physical constraints used during evaluation.

Evaluation protocol. ForceVid3D-Bench evaluates force adherence, physical consistency, and visual quality. Image-plane cases support comparison with prior force-conditioned methods, while camera-space 3D cases evaluate 3D control. Detailed annotation and evaluation protocols are provided in Appendix C.

5 Experiments

5.1 Experimental Setup

Implementation details. We build ForceVid3D on the publicly available Wan2.2-TI2V-5B backbone (Wan et al., 2025) with a 10-block force-control branch, while keeping the pretrained video backbone frozen. We generate 49-frame videos at 640×384 resolution and 10 FPS, and use the 15K-step checkpoint as the final model. Base SFT mixes 90% force-labelled synthetic samples with 10% unlabelled real wind clips, which use a learned null-force condition. Full training and inference settings are provided in Appendix F.

Baselines. We compare with Force Prompting (Gillman et al., 2025), StreamForce (Wang et al., 2026), and Goal Force (Gillman et al., 2026) using their released implementations and checkpoints, retaining each method's native backbone, force encoding, and inference schedule. All comparable methods receive the same initial image, text condition, and image-plane force. Force Prompting and StreamForce support point forces and wind, whereas Goal Force is evaluated only on point forces. Since the baselines lack an equivalent camera-space 3D interface, we compare them with ForceVid3D in the shared 2D setting and report 3D results separately. Appendix F.1 lists the corresponding configurations.

Table 1: **Comparison under shared image-plane force control.** The upper block evaluates methods supporting both point forces and wind; the lower block provides the point-force comparison including Goal Force, whose released interface does not support wind. All scores are reported on a 0–100 scale. Bold marks the best result within each comparable block.

	Force Adherence				Physical Consistency	Visual Quality	
Method	**F-Align ↑**	**Proj.-Resp.@45 ↑**	**M^+ IoU ↑**	**H-Force ↑**	**H-Physics ↑**	**Aesth. ↑**	**Imaging ↑**
Shared point-force and wind evaluation							
Force Prompting	68.82	47.44	82.87	62.18	49.04	58.47	65.90
StreamForce	70.63	68.59	81.32	75.32	48.88	**58.84**	**68.81**
ForceVid3D	**84.21**	**89.74**	**83.14**	**93.59**	**75.16**	58.07	66.44
Point-force evaluation including Goal Force							
Force Prompting	63.71	36.67	72.88	48.89	36.39	58.65	68.18
StreamForce	64.09	70.00	70.88	69.44	33.61	**58.96**	**70.22**
Goal Force	75.37	75.56	73.43	73.06	67.50	55.92	68.19
ForceVid3D	**78.82**	**93.33**	**74.26**	**94.72**	**71.39**	57.35	67.18

Evaluation. ForceVid3D-Bench contains 165 images and 310 cases: 156 image-plane conditions for comparison and 154 camera-space 3D conditions for capability evaluation. *Force Adherence* includes F-Align, which combines task-specific response success and response grounding; Proj.-Resp.@45, a response-aware projected-direction success measure; M^+ IoU, the overlap between detected response motion and the annotated responsive region; and H-Force, a human rating of whether the intended content responds consistently with the prescribed force. *Physical Consistency* includes H-Physics, a human rating of whether the generated motion remains physically coherent with respect to scene constraints, contact, continuity, and object identity; and, when simulator references are available, P-ADE, the object-scale-normalized average error between generated and reference centroid paths. *Visual Quality* uses the Aesthetic and Imaging dimensions of VBench (Huang et al., 2024). For fair comparison with prior force-conditioned methods, the shared 2D comparisons and ablations evaluate only the prescribed image-plane force direction and do not constrain the depth component of the generated response. Full camera-space 3D directional consistency is evaluated separately in Section 5.4. Appendix D provides the metric definitions, validity rules, and aggregation details.

5.2 Comparison with Prior Methods in the Shared 2D Setting

Table 1 compares ForceVid3D with prior force-conditioned methods in their shared 2D control space. ForceVid3D achieves the best result on every reported force-adherence and physical-consistency metric in both comparison settings. In the shared point-force and wind evaluation, F-Align rises from the best baseline result of 70.63 to 84.21, while H-Force rises from 75.32 to 93.59. The point-force evaluation including Goal Force shows the same lead on the core control and physics metrics. Aesthetic and Imaging remain competitive with prior methods. Because backbones, training data, and native samplers differ, this comparison evaluates complete systems; component attribution is provided by the matched ablations.

Figure 5 illustrates three complementary effects. ForceVid3D follows the upward force while maintaining stable apparent depth, localizes the horizontal response without structural breakage or motion leakage, and produces coherent flexible motion under distributed wind. The baselines instead show weaker or ambiguous displacement, deformation, or under-response, complementing the aggregate evidence for directional adherence, response localization, and physical coherence.

5.3 Ablation Studies in the Shared 2D Setting

We evaluate the contributions of the Projective 3D Force Map and D-RAS under the same image-plane protocol used for the main comparison. The representation variants use matched 5K training schedules. For the objective study, both branches continue from the same 8K SFT checkpoint to 11K with identical data exposure, optimization, evaluation cases, and sampling seed; they differ only in whether D-RAS is enabled.

Projective 3D Force Map. Replacing the direct encoding with the projective representation improves F-Align, projected-response success, and human judgments of force adherence and physical

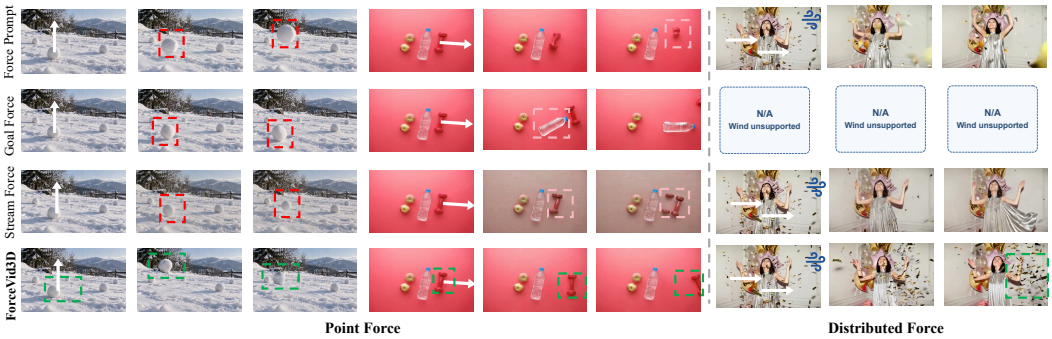


Figure 5: **Qualitative comparison under shared image-plane forces.** Each example fixes the input scene and force condition across methods. Representative point-force and wind responses are shown over time; arrows indicate the prescribed force direction.

Table 2: **Component ablations under image-plane force control.** A rows compare force representations under matched training. B rows continue from the same 8K SFT checkpoint to 11K, with or without D-RAS. All scores are reported on a 0–100 scale; bold marks the better result within each pair.

			Force Adherence				Physical Consistency	Visual Quality	
ID	Representation / Objective	Train interval	F-Align ↑	Proj.-Resp.@45 ↑	M^+ IoU ↑	H-Force ↑	H-Physics ↑	Aesth. ↑	Imaging ↑
Force representation									
A1	Direct-5C + SFT	0–5K	75.79	74.36	81.59	78.53	55.29	**59.51**	**66.29**
A2	Projective-7C + SFT	0–5K	**78.35**	**76.28**	**82.11**	**83.33**	**59.46**	59.49	66.06
Training objective: matched branches from the same 8K SFT checkpoint									
B1	Projective-7C + SFT	8K–11K	80.08	78.85	82.23	85.58	64.58	**59.21**	**66.10**
B2	Projective-7C + D-RAS	8K–11K	**81.79**	**83.33**	**82.26**	**86.54**	**67.95**	59.14	65.86

consistency, while response grounding and visual quality change only slightly. These matched results support the projective representation under the same 2D evaluation protocol used for the main comparison.

D-RAS. Under matched continuation, D-RAS improves F-Align, projected-response success, and H-Physics while maintaining nearly unchanged response-region overlap (M^+ IoU: 82.23 versus 82.26). These results support the joint auxiliary objective under matched training exposure.

5.4 Camera-Space 3D Force Control

We evaluate camera-space 3D control on unseen images and simulator-referenced cases: unseen images assess final-model capability, while paired simulator references enable controlled component attribution.

Unseen-image 3D capability. Table 3 evaluates the final model on camera-space 3D conditions in ForceVid3D-Bench. The point-force and wind cases cover localized and distributed interventions on unseen images, testing force adherence and response localization under explicit camera-space directions. Because the compared baselines do not expose an equivalent 3D force interface, this experiment evaluates ForceVid3D's 3D control capability rather than a direct cross-method ranking.

Across the full human-rated 3D set, H-Force reaches 88.92 for point forces and 67.80 for wind, providing direct evidence of final-model control on both localized and distributed interventions. The simulator-referenced analysis below then isolates the contributions of the proposed components.

Controlled synthetic 3D ablations. For the controlled analysis, Z-subset Ang. measures projected angular error on depth-containing direct responses, FS@45 measures directional success within 45° over direct-response cases, and R-Act. IoU measures agreement in responsive-region activation. Exact applicability sets and denominators are provided in Appendix D.

Table 3: **Camera-space 3D force-control capability.** Point and wind are reported separately. Automatic response and visual-quality metrics use the automatically evaluable subsets (27 point and 5 wind cases), while human ratings cover all 88 point and 66 wind cases and assess the full prescribed 3D direction.

	Force Adherence				Physical Consistency	Visual Quality	
Force	F-Align ↑	Proj.-Resp.@45 ↑	M^+ IoU ↑	H-Force ↑	H-Physics ↑	Aesth. ↑	Imaging ↑
Point	74.17	59.26	66.95	88.92	66.48	55.56	66.85
Wind	78.38	80.00	96.88	67.80	54.92	58.49	68.10

Table 4: **Controlled synthetic ablations under camera-space 3D forces.** All scores except angular error and P-ADE are reported on a 0–100 scale; bold marks the better result within each matched pair.

			Force Adherence					Physical Consistency		Visual Quality	
ID	Representation / Objective	Train interval	F-Align ↑	Z-subset Ang. ↓	FS@45 ↑	R-Act. IoU ↑	H-Force ↑	P-ADE ↓	H-Physics ↑	Aesth. ↑	Imaging ↑
A1	Direct-5C + SFT	0–5K	28.37	46.61	72.34	88.32	72.92	**0.690**	60.42	47.49	65.91
A2	Projective-7C + SFT	0–5K	**31.03**	**31.13**	**89.36**	**92.16**	**82.29**	0.812	**69.10**	**49.73**	**68.90**
B1	Projective-7C + SFT	8K–11K	27.50	43.64	85.11	92.17	87.50	0.770	78.82	**48.93**	68.43
B2	Projective-7C + D-RAS	8K–11K	**31.91**	**33.09**	**89.36**	**94.15**	**89.58**	**0.707**	**79.17**	48.76	**68.64**

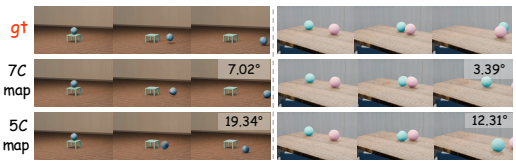


Figure 6: **Illustrative comparison of force representations.** Values denote projected angular error (lower is better). Projective-7C reduces it from 19.34° to 7.02° and from 12.31° to 3.39°, illustrating improved directional control in two depth-containing 3D cases.

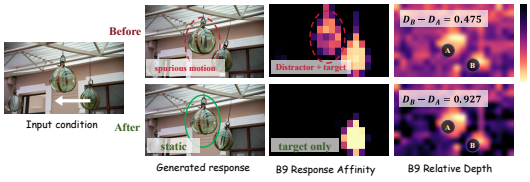


Figure 7: **Illustrative D-RAS failure correction.** From the 8K SFT initialization to the 11K D-RAS checkpoint, distractor affinity is suppressed, target–distractor depth separation improves, and the spurious response is removed. For visualization, B9 diagnostic readouts are shown at latent frame 5 of the first denoising step.

On 72 simulator-referenced cases, Table 4 shows that Projective-7C reduces depth-sensitive projected angular error from 46.61° to 31.13° and raises FS@45 from 72.34 to 89.36. P-ADE changes from 0.690 to 0.812, indicating that improved directional alignment does not necessarily translate to better full-path agreement. These results support the role of position-dependent projection cues in depth-sensitive directional control. Figure 6 illustrates two representative depth-containing cases.

Under matched continuation from the same SFT checkpoint, D-RAS improves F-Align, depth-sensitive angular error, FS@45, response-activation agreement, and P-ADE, while leaving visual quality largely unchanged. Together, the controlled results support complementary effects of the projective representation on directional control and D-RAS on response evolution. Figure 7 shows an illustrative 8K-SFT-to-11K-D-RAS failure correction.

6 Conclusion

We introduced **ForceVid3D** for camera-space 3D force control, combining a projective force representation with D-RAS for response and depth grounding. Together with a 60K force–video dataset and ForceVid3D-Bench, the framework achieves strong force adherence and physical consistency in the shared 2D setting, while dedicated evaluations validate camera-space 3D control. Together, these results suggest that exposing projective force geometry and supervising response-aware intermediate representations provides a practical route to direct 3D physical control without explicit scene reconstruction at inference.

AI USE STATEMENT

Generative AI tools were used to assist writing, literature retrieval, research execution, and synthetic-data generation. All AI-assisted content was reviewed and verified by the authors, who take responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

The appendix provides details of the dataset construction, model training, evaluation protocols, and implementation. Code, data, and model checkpoints will be released.

REFERENCES

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. In *International Conference on Learning Representations*, volume 2025, pp. 102075–102121, 2025.

Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video diffusion models with motion prior and reward feedback learning. *arXiv preprint arXiv:2305.13840*, 2023.

Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Carl Doersch, Yusuf Aytar, Michael Rubinstein, Chen Sun, Oliver Wang, Andrew Owens, and Deqing Sun. Motion prompting: Controlling video generation with motion trajectories. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–12. IEEE, 2025.

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. *Advances in Neural Information Processing Systems*, 38:103174–103201, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20077–20087, 2026.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for video diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 100433–100464, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Yanxun Li, Hao Wen, Bingze Song, Jiashu Zhu, Aiming Hao, Chubin Chen, Jintao Chen, Jiahong Wu, Xiangxiang Chu, and Miao Wang. Learning explicit physical parameter control and benchmarking for video generation. *arXiv preprint arXiv:2607.18924*, 2026.

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

Wei Liu, Ziyu Chen, Zizhang Li, Yue Wang, Hong-Xing Yu, and Jiajun Wu. Realwonder: Real-time physical action-conditioned video generation. *arXiv preprint arXiv:2603.05449*, 2026.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 948–958. IEEE, 2026.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in neural information processing systems*, 36: 1363–1389, 2023.

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

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. *Advances in Neural Information Processing Systems*, 38:167907–167932, 2025.

Hanhui Wang, Yiming Xie, Haiwen Feng, Zhaoyang Lv, Shenlong Wang, and Huaizu Jiang. Streaming video generation with streaming force control. *arXiv preprint arXiv:2606.07508*, 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. *Advances in Neural Information Processing Systems*, 36:7594–7611, 2023.

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*, pp. 1–11, 2024.

Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. Draganything: Motion control for anything using entity representation. In *European Conference on Computer Vision*, pp. 331–348. Springer, 2024.

Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824. IEEE, 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.

A Dataset Details

We construct **ForceVid3D-Dataset**, comprising 59,635 (approximately 60K) rendered force–video pairs with explicit 3D force annotations. The dataset is organized to cover distinct *physical response modes* under external forces, with semantic and visual variation within each mode.

A.1 Simulation Pipeline

Each training sample is generated by: (1) sampling a scene and physical assets; (2) sampling object properties and initial states; (3) applying a local or distributed 3D force; (4) simulating the resulting dynamics; and (5) rendering the video together with force metadata and auxiliary annotations.

We implement scene construction and rendering in Blender 5.2 and use PyBullet for response modes that require rigid-body simulation. Videos contain 49 frames at 10 FPS.

A.2 PHYSICAL RESPONSE MODES

The dataset contains two force families and seven response-mode groups:

- **Localized point force:** supported rigid-body motion (sliding, rolling, spinning, and toppling), free-flight motion, constrained motion (including pendular and suspended motion), and multi-object interaction through collision or momentum transfer;
- **Distributed wind force:** anchored or tethered motion, free aerodynamic transport, and fluid or plume advection.

Within each family, we randomize object geometry, appearance, camera configuration, force direction, magnitude, location, and physical parameters. This variation provides multiple realizations of similar response modes across semantically different objects and scenes.

A.3 DATASET STATISTICS

Table 5 reports the number of videos in each physical response mode. The seven groups are mutually exclusive and contain 59,635 synthetic force–video pairs in total. Figure 8 provides representative examples from the localized and distributed force families.

Table 5: **Composition of ForceVid3D-Dataset.**

Response Mode	Number of Videos
Supported Rigid-Body	18,334
Free-Flight	10,689
Constrained	14,095
Multi-Object Interaction	4,613
Anchored / Tethered Wind	11,034
Free Aerodynamic Transport	499
Fluid / Plume Advection	371
Total	**59,635**

To improve generalization to real-world scenes, base SFT additionally uses 6,136 unlabelled real wind clips through the NULL-FORCE condition; these clips are not part of ForceVid3D-Dataset. Simulated videos are resized to 640×384 using Lanczos interpolation without cropping, and the real clips are preprocessed to the same resolution, length, and frame rate.

A.4 FORCE ANNOTATION AND COORDINATE CONVENTION

Each physical intervention is represented as

$$\mathcal{F} = (\mathcal{R}, \mathbf{d}, m, t_s, t_e), \tag{8}$$

where $\mathcal{R}$ denotes the spatial support of the force, $\mathbf{d} = (d_x, d_y, d_z)$ is a unit direction in camera coordinates, m is the normalized force magnitude, and $[t_s, t_e)$ denotes the temporal application interval on the RGB-frame axis.

Camera coordinates. The horizontal and vertical camera axes correspond to the image-plane x and y directions, while z denotes the optical axis. The exact sign convention used in our implementation is summarized in Table 6.

For local forces, $\mathcal{R}$ is specified by an image-space application point and rasterized into a Gaussian support map ρ with standard deviation 1.5 latent cells. For distributed wind, $\rho = 1$ throughout the active interval. The direction and magnitude are constant within each annotated intervention, and all force channels are zero outside it. The first VAE latent bin corresponds to RGB frame 0, and latent bin $k > 0$ covers frames $4k-3, \ldots, 4k$; a bin is active whenever its covered frames intersect $[t_s, t_e)$. The annotations describe the physical intervention only and do not contain future trajectories.

The direction and magnitude channels remain spatially broadcast within an active latent bin. The two projective channels are exposed over the full image for wind and gated to $\rho(q) \geq 0.05$ for a local point force. The same temporal gate is applied to all seven channels.

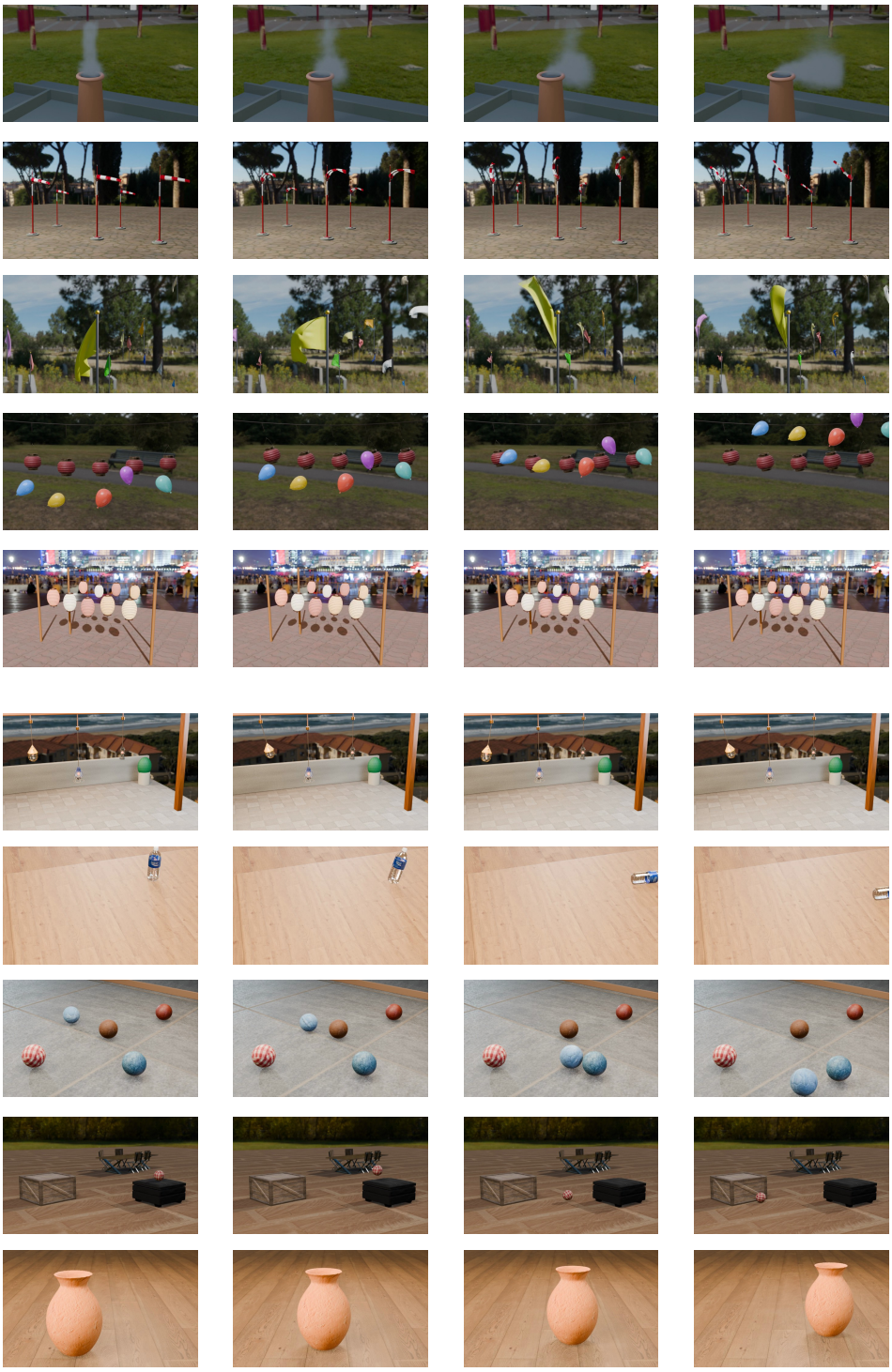


Figure 8: **Diversity of ForceVid3D-Dataset.** The top five rows show distributed wind-force responses, and the bottom five rows show localized point-force responses.

We use the camera intrinsics stored with each simulation. If they are unavailable, the encoder uses fixed nominal intrinsics with $f_x = f_y = 888.89$ pixels at the 640×384 input resolution and the image center as principal point. The intrinsics are scaled consistently to the latent grid. An image coordinate (u, v) maps to continuous latent coordinates $x_l = uW_l/W - 0.5$ and $y_l = vH_l/H - 0.5$.

Table 6: **Camera-coordinate convention used by the force encoder.**

Direction	Meaning
$+x$	Image right
$-x$	Image left
$+y$	Image down
$-y$	Image up
$+z$	Forward along the optical axis (away from the camera)
$-z$	Toward the camera

B Depth-Grounded Response Affinity Supervision Details

B.1 D-RAS Target Construction

D-RAS requires a soft response-affinity target and a local relative-depth target. Training uses simulator examples with instance masks and valid first-frame depth. An instance is responsive when its visible state changes under the intervention, and its initial-frame mask defines the surface to which the future response is associated. Benchmark-image masks are used for evaluation and benchmark annotation, not as D-RAS training targets.

Given a ground-truth video V and initial frame I_0, we first compute

$$\Delta I_t(x,y) = \frac{1}{C}\sum_{c=1}^{C}\left|V_t^{(c)}(x,y) - I_0^{(c)}(x,y)\right|. \tag{9}$$

For simulator object i with future mask M_t^i, we compute a normalized response energy and soft selection weight,

$$r_i = \frac{\sum_{t,x} M_t^i(x)\Delta I_t(x)}{\sum_{t,x} M_t^i(x) + \epsilon}, \qquad w_i = \mathrm{clip}\left(\frac{r_i - 0.035}{0.08}, 0, 1\right). \tag{10}$$

The response-affinity target is the area-downsampled weighted union of the corresponding initial-frame instance masks,

$$A^* = \mathcal{D}_{12\times 20}\left[\max_i w_i M_0^i\right]. \tag{11}$$

No spatial dilation or Gaussian blur is applied to A^*; a one-token dilation is used only to form the local support of the relative-depth loss. All level-specific response and depth heads read the I_0 token slice, so the target has one temporal reference plane rather than a repeated future target.

The target intentionally describes *response association* rather than future velocity, optical flow, trajectory, or exact future occupancy. For example, the same object may remain responsive under multiple force directions even though its resulting motion changes substantially.

For valid initial-frame depth D_0, we compute the log-depth target in Equation 6. The center is the median log depth of the responsive surface, and the scale is the median absolute deviation over the surface plus a narrow context ring, lower-bounded by 0.10 log-depth units. We supervise depth only on this local domain Ω_D, excluding invalid renderer pixels.

Averaging the level-specific response logits gives a, with $\hat{A} = \sigma(a)$. Let Ω_A denote the response-target token grid. We use

$$\mathcal{L}_{\mathrm{WBCE}}(a, A^*) = -\frac{1}{|\Omega_A|}\sum_{q\in\Omega_A}\left[w_+ A^*(q)\log\hat{A}(q) + \big(1 - A^*(q)\big)\log\big(1-\hat{A}(q)\big)\right], \tag{12}$$

$$\mathcal{L}_{\mathrm{Dice}}(\hat{A}, A^*) = 1 - \frac{2\langle\hat{A}, A^*\rangle + \epsilon}{\|\hat{A}\|_1 + \|A^*\|_1 + \epsilon}, \tag{13}$$

where the positive-class weight is $w_+ = 8.0$. The weighted cross-entropy addresses the sparsity of responsive tokens, while the soft Dice term measures their spatial overlap. The aggregated depth prediction uses SmoothL1 over valid pixels in Ω_D. Table 7 summarizes the target hyperparameters.

Table 7: **D-RAS target hyperparameters.**

Parameter	Value
Soft threshold	0.035
Soft transition width	0.08
Response-target dilation	0 tokens
Response-target blur	None
Spatial target resolution	12×20 tokens
Temporal target resolution	one I_0 token plane
Depth clipping κ	5
Depth scale floor	0.10 log-depth units
Depth context-ring width	1 token
Positive-class weight	8.0

B.2 Layer-Wise Response Probe Analysis

B.2.1 Probe Architecture and Training

We freeze the complete ForceVid3D generator and attach an independent lightweight probe to each of the 10 ControlNet blocks. For hidden feature $H^{(l)}$, the response probe produces logits

$$a^{(l)} = W_l \, \mathrm{LN}(H^{(l)}) + b_l. \tag{14}$$

Only the probe parameters are optimized. The underlying ControlNet features are detached during probe training. We train independent response and relative-depth probes for all ten levels for 1,000 updates with AdamW, learning rate 3×10^{-4}, and weight decay 10^{-4}, using a fixed held-out set of 80 rigid point-force and 80 pendulum examples. Response probes use the objective above; depth probes use SmoothL1 on the same valid relative-depth target used by D-RAS. Probe calibration uses diffusion timesteps $\tau \geq 950$ (out of 1,000).

The probe uses the *dynamic view*: the ground-truth future is encoded and corrupted at high diffusion noise, while the first-frame condition is retained. Table 8 reports decodability under this input protocol.

B.2.2 Layer-Wise Decodability

Table 8 reports the layer-wise probe results on the frozen 8K SFT controller.

Table 8: **Layer-wise D-RAS probe on the frozen 8K SFT controller.** Dynamic-view probe: response Dice is higher-is-better and relative-depth MAE is lower-is-better. Scores are task-macro means over 80 rigid and 80 pendulum validation samples.

Metric	B0	B1	B2	B3	B4	B5	B6	B7	B8	B9
Response Dice ↑	0.5089	0.5177	0.5231	0.5294	0.5322	0.5356	0.5381	0.5437	0.5481	**0.5493**
Depth MAE ↓	1.1031	0.9438	0.8596	0.7850	0.7595	0.7464	0.7184	0.6502	0.5916	**0.5685**

Probe performance generally improves with controller depth, with B9 strongest on both readouts. We select B9 together with the intermediate B4 to provide supervision at two controller depths. Block indices are zero-based.

B.3 D-RAS Architecture and Training

B.3.1 Level-Specific Heads and Prediction Aggregation

For each selected controller level $\ell \in \mathcal{S}$, independent response and relative-depth heads operate on the 3,072-dimensional first-frame feature:

$$a^{(\ell)} = W_A^{(\ell)} \, \mathrm{LN}(H_0^{(\ell)}) + b_A^{(\ell)}, \qquad \hat{D}^{(\ell)} = W_D^{(\ell)} \, \mathrm{LN}(H_0^{(\ell)}) + b_D^{(\ell)}. \tag{15}$$

The heads do not share parameters across levels. Their predictions are aggregated as

$$a = \frac{1}{|\mathcal{S}|} \sum_{\ell \in \mathcal{S}} a^{(\ell)}, \qquad \hat{D} = \frac{1}{|\mathcal{S}|} \sum_{\ell \in \mathcal{S}} \hat{D}^{(\ell)}. \tag{16}$$

Response probabilities are computed as $\hat{A} = \sigma(a)$ after averaging the level-specific logits. For the selected B4 and B9 levels, D-RAS therefore uses four independent lightweight heads in total: one response head and one relative-depth head at each level. Each head is initialized from the corresponding calibrated probe, and the heads and controller are then jointly optimized.

Flow matching retains its standard timestep distribution, while D-RAS is activated for sampled timesteps $\tau \geq 950$. All D-RAS heads are training-only and are removed at inference.

B.3.2 Matched Continuation Ablation

Training is organized in two stages: force-conditioned SFT from step 0 to 8K, followed by a second-stage continuation to 15K for the final checkpoint. Table 2 reports the two continuations as B1 and B2. Both continue from the same 8K checkpoint to 11K with the same data, optimization, evaluation cases, and random seed; they differ only in whether D-RAS is enabled. Their comparison therefore isolates the training objective. The pretrained video backbone remains frozen. Table 9 lists the complete continuation configuration.

Table 9: **Second-stage training and matched D-RAS ablation configuration.**

Setting	Value
Common initialization	8K-step SFT checkpoint
Final continuation	8K–15K (7,000 updates)
Matched ablation continuation	8K–11K (3,000 updates)
Ablation branches	SFT / D-RAS
Selected controller blocks	B4 and B9 (zero-based)
Per-level prediction heads	Independent response and depth heads
Prediction aggregation	Mean across B4/B9 predictions
Response aggregation	Mean logits before sigmoid
D-RAS loss activation	$\tau \geq 950$ (of 1,000)
λ_A / λ_D	1×10^{-4} / 5×10^{-5}
Loss-weight ramp	250 updates from step 8K
Controller learning rate	5×10^{-6} cosine to 1×10^{-6}
Response/depth-head learning rate	1×10^{-4}

C Benchmark Protocol Details

C.1 Benchmark and Controlled Diagnostics

ForceVid3D-Bench is constructed from 165 unseen images collected from web photographs and AI-generated scenes. It contains 310 force-conditioned evaluation cases: 156 image-plane conditions (90 point force and 66 wind) and 154 camera-space 3D conditions (88 point force and 66 wind). The two control settings are not paired one-to-one across images: some images contain only image-plane or only camera-space 3D conditions, while others contain both. Each annotation specifies the force family, application location or spatial support, camera-space direction, magnitude, and active interval.

For response grounding, annotators label responsive content as M^+ and explicit non-responsive distractors as M^-. Unannotated background is not automatically treated as negative. Point-force and distributed-wind cases retain separate task labels, and image-plane and camera-space 3D conditions are reported separately so that 2D-only methods are compared only in their shared control space.

We additionally use a controlled synthetic evaluation set of 72 cases for matched component analysis and ablation. These cases provide paired simulator futures, response regions, depth, trajectories, and task-specific physical constraints. Results on this set are reported separately from the unseen-image system comparison.

C.2 Motion Extraction and Valid Frames

For point-force cases, annotated first-frame masks are propagated across frames using SAM2. ForceVid3D-Bench point-force response detection normalizes centroid displacement by the square root of the initial mask area, using the median of the first three centroids as the baseline. Response onset requires two consecutive frames beyond 0.15 object scales. A PCA/SVD axis over a ten-frame response window is oriented toward the largest excursion.

For distributed-wind responses, background correspondences estimate a homography to frame zero, with affine ECC and identity fallbacks. Farneback flow is computed on the aligned frames and summarized inside the responsive region. The compensated frames provide the basis for regional motion analysis. For the controlled synthetic evaluation set, motion overlap is computed from independently normalized generated and simulator-reference frame differences under the fixed simulator camera.

Trajectory comparison uses the common visible prefix and terminates when the tracked object exits the frame or its mask becomes unavailable. Short or zero-length trajectories do not define P-ADE. Metric-specific applicability rules are given together with their definitions in Appendix D.

C.3 Aggregation

Metrics are computed per evaluation case and then macro-averaged over applicable cases. When multiple generated outputs correspond to the same case, they are first aggregated within that case. Point-force and wind cases are retained as separate task categories where appropriate, and image-plane and camera-space 3D evaluations are reported separately.

C.4 Evaluation Set Composition

Table 10 summarizes the retained ForceVid3D-Bench evaluation set. Metric applicability depends on the semantics of each case; an inapplicable measurement is excluded, whereas an applicable non-response is treated as a failure by response-based metrics.

Table 10: **Composition of the ForceVid3D-Bench evaluation set.**

Control setting	Point force	Wind	Total
Image-plane force	90	66	156
Camera-space 3D force	88	66	154
Total	**178**	**132**	**310**

D Evaluation Metrics and Human Study

D.1 Controlled Diagnostic Metrics

Task-aware force alignment. For available components $q_1, \ldots, q_K \in [0, 1]$, where K is the number of included components, we define their harmonic mean as

$$H(q_1, \ldots, q_K) = \begin{cases} 0, & \text{if any } q_k = 0, \\ \dfrac{K}{\sum_{k=1}^{K} q_k^{-1}}, & \text{otherwise.} \end{cases} \tag{17}$$

Per-case force alignment is

$$\text{F-Align}_i = H(D_i, G_i, 1 - L_i), \tag{18}$$

where D_i denotes task-aware response success, G_i measures agreement with the expected response, and L_i measures background motion leakage. An unavailable leakage component is omitted. We compute the harmonic mean per case before averaging across cases.

In the controlled synthetic evaluation set, D_i is the 45° directional-success indicator for direct responses, ground-truth motion-energy recall for pendular responses, and secondary-object transfer success for collisions. Let $E_t^{\text{gen}}(q)$ and $E_t^{\text{gt}}(q)$ denote generated and simulator-reference motion

energy, computed from Gaussian-smoothed adjacent-frame RGB differences and independently normalized per video using P70/P95. We define

$$G_i = \frac{\sum_{t,q} \min(E_t^{\text{gen}}(q), E_t^{\text{gt}}(q))}{\sum_{t,q} \max(E_t^{\text{gen}}(q), E_t^{\text{gt}}(q)) + \epsilon}, \tag{19}$$

$$L_i = \frac{\sum_{t,q \in S} E_t^{\text{gen}}(q)}{\sum_{t,q} E_t^{\text{gen}}(q) + \epsilon}. \tag{20}$$

Here S is the background support outside the union of simulator instance supports and reference motion above 0.1, expanded by seven pixels. Pendular recall uses the same energy intersection as G_i, divided by total reference energy. The controlled synthetic evaluation set uses paired motion references, whereas ForceVid3D-Bench uses annotated regions, so their absolute F-Align values are not directly comparable.

Projected direction and success. For direct responses, robust early-response fitting yields unit image-plane directions $\hat{\mathbf{v}}_i$ and $\mathbf{v}_i^*$ from generated and simulator-reference trajectories. The projected angular error is

$$\theta_i = \arccos\big(\text{clip}(\hat{\mathbf{v}}_i^\top \mathbf{v}_i^*, -1, 1)\big). \tag{21}$$

FS@45 is the percentage of applicable direct-response cases with a detected response and $\theta_i \le 45^\circ$:

$$\text{FS@45} = \frac{100}{|\mathcal{C}_{\text{dir}}|} \sum_{i \in \mathcal{C}_{\text{dir}}} \mathbf{1}[\text{response detected}]\mathbf{1}[\theta_i \le 45^\circ]. \tag{22}$$

The direct-response set $\mathcal{C}_{\text{dir}}$ contains 47 cases and excludes pendular and collision responses. Non-response counts as failure. *Z-subset Ang.* is the mean projected angular error, in degrees, on the 16 direct-response cases with $|d_z| \ge 0.6$. It measures image-plane response direction under depth-containing forces, rather than a reconstructed 3D angle.

Region-activation grounding. R-Act. IoU measures agreement in the activation of annotated responsive regions, including primary targets and simulator-responsive secondary objects. For a tracked or annotated region R_t, define its mean motion energy and temporally pooled activation as

$$e_t(R) = \frac{1}{|R_t|} \sum_{q \in R_t} E_t(q), \qquad a(R) = \max_t \frac{1}{3} \sum_{\tau=t}^{t+2} e_\tau(R). \tag{23}$$

We use windows with visible support; when no complete window is valid, we average visible frames, or assign zero if none exist. Generated and reference activations are computed separately, giving

$$\text{IoU}_R = \frac{\min(a_{\text{gen}}(R), a_{\text{gt}}(R))}{\max(a_{\text{gen}}(R), a_{\text{gt}}(R)) + \epsilon}. \tag{24}$$

R-Act. IoU averages these scores over valid positive regions within each case and then over cases. It measures region-level activation agreement, not pixel-mask overlap or exact response timing, and is distinct from G_i in Equation 19.

Trajectory consistency. P-ADE measures image-plane path agreement after removing each path's initial offset and normalizing both paths by the initial reference-object scale $s_0 = \max(\sqrt{|M_0^{\text{gt}}|}, 1)$. Over the common visible prefix, the normalized paths $\hat{\mathbf{q}}$ and $\mathbf{q}^*$ are independently resampled at $K = 32$ uniformly spaced fractions of arc length:

$$\text{P-ADE} = \frac{1}{K} \sum_{k=1}^{K} \|\hat{\mathbf{q}}(u_k) - \mathbf{q}^*(u_k)\|_2, \qquad u_k = \frac{k-1}{K-1}. \tag{25}$$

This resampling removes traversal-speed differences while preserving path extent. P-ADE is undefined for fewer than two distinct visible positions; the A1 result uses 70 valid cases and the other rows use 72. Lower values indicate closer path agreement.

Physical consistency and visual quality. P-ADE provides the automatic path-consistency measure, complemented by H-Physics ratings. Aesthetic and Imaging Quality follow the VBench protocol described in Appendix D.4.

D.2 FORCEVID3D-BENCH METRICS

Force adherence and projected-direction proxy. ForceVid3D-Bench uses Equation 18 with spatial M^+ IoU as G_i. For ordinary direction-sensitive responses, D_i is response-aware projected-direction success. For multi-object interaction, constrained motion, anchored/tethered motion, and free flight, D_i is response detection because displacement is not a universal proxy for the applied force. The reported ForceVid3D-Bench F-Align therefore uses the two available components,

$$\text{F-Align}_i = H(D_i, G_i). \tag{26}$$

For the shared-2D case set $\mathcal{C}$, Proj.-Resp.@45 measures response-aware agreement with the projected force direction:

$$\text{Proj.-Resp.@45} = \frac{100}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \mathbf{1}[\text{response detected}]\mathbf{1}[\theta_i \leq 45^\circ]. \tag{27}$$

The reference direction is the annotated image-plane projection of the force, while the observed direction is estimated from target motion or regional flow. An undetected response contributes zero. The shared evaluation contains 156 image-plane cases, including 90 point-force cases used in the point-only comparison.

For camera-space 3D evaluation, we predefined the automatically evaluable subset by retaining cases whose prescribed force has a non-degenerate image-plane projection ($|d_z| \leq 0.95$) and whose visible response provides a valid image-space observable of the applied force. We exclude cases in which contact or support constraints, gravity-dominated projectile motion, or tether and return dynamics make the annotated displacement incompatible with the instantaneous force direction; these cases remain in the human evaluation. This protocol yields 27 point-force cases and 5 wind-force cases, on which we report automatic response and visual-quality metrics. Human ratings cover all 88 point-force and 66 wind cases and assess the full prescribed 3D direction.

Response grounding. ForceVid3D-Bench annotations define M^+ as content expected to respond during the clip and M^- as explicitly non-responsive distractor content. Unannotated background is not automatically treated as negative. For point-force cases, annotated first-frame response masks are propagated across frames, while regional wind annotations define the evaluation region.

After camera compensation and Gaussian smoothing ($\sigma = 1.2$), per-video motion energy is normalized using its P70 and P95 values and temporally filtered with a three-frame maximum moving average. Thresholding the resulting motion map at 0.15 gives the detected response region $\hat{M}$. We measure spatial response agreement as

$$\text{M}^+\text{IoU} = \frac{|\hat{M} \cap M^+|}{|(\hat{M} \cap (M^+ \cup M^-)) \cup M^+|}. \tag{28}$$

Thus, motion on explicitly annotated distractors is penalized, while unannotated background is not treated as negative support. M^+ IoU is the response-grounding metric reported in the ForceVid3D-Bench tables.

Physical consistency. Because an unseen image has no unique future trajectory, ForceVid3D-Bench does not compare generated motion against a pseudo-ground-truth video. H-Physics instead evaluates whether the realized motion is physically coherent for the given scene and intervention, including constraint preservation, contact behavior, motion continuity, and object identity. For example, pendular motion should preserve its fixed pivot, collision responses should follow contact, and generated objects should remain temporally coherent without duplication, splitting, or disappearance.

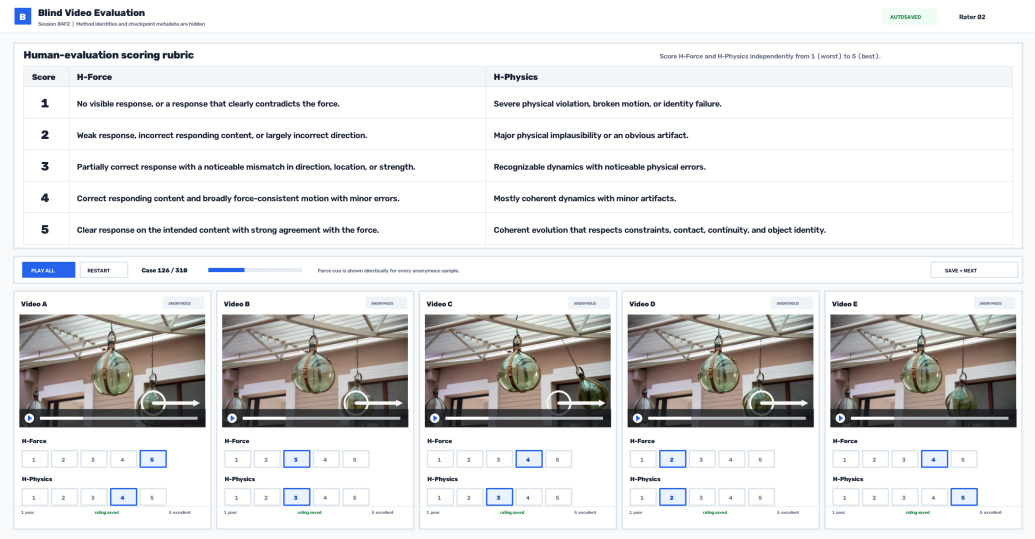


Figure 9: **Human-evaluation interface.** Candidate videos are displayed with the applied-force visualization and five-point rubric, with separate ratings for force adherence and physical consistency.

D.3 Human Evaluation and Aggregation

Evaluation procedure. Three annotators independently scored each output on a five-point scale. Within each case, candidate videos were shown together in randomized order with method identities hidden, together with the initial image and applied-force visualization. H-Force measures whether the intended content responds consistently with the intervention, while H-Physics assesses physical coherence in terms of constraints, contact, continuity, and object identity. For 2D conditions, unspecified depth motion is not penalized; for 3D conditions, the prescribed depth direction is also considered.

For each output i and human metric, we average the three ratings $r_{i,j} \in \{1, 2, 3, 4, 5\}$ and linearly map the result to $[0, 100]$:

$$\bar{r}_i = \frac{1}{3}\sum_{j=1}^{3} r_{i,j}, \qquad H_i = 100\,\frac{\bar{r}_i - 1}{4}, \qquad \bar{H} = \frac{1}{N}\sum_{i=1}^{N} H_i. \tag{29}$$

H-Force and H-Physics are scored independently, and final results are macro-averaged across cases.

Table 11: **Human-evaluation scoring rubric.**

Score	H-Force	H-Physics
1	No visible response, or a response that clearly contradicts the force.	Severe physical violation, broken motion, or identity failure.
2	Weak response, incorrect responding content, or largely incorrect direction.	Major physical implausibility or an obvious artifact.
3	Partially correct response with a noticeable mismatch in direction, location, or strength.	Recognizable dynamics with noticeable physical errors.
4	Correct responding content and broadly force-consistent motion with minor errors.	Mostly coherent dynamics with minor artifacts.
5	Clear response on the intended content with strong agreement with the force.	Coherent evolution that respects constraints, contact, continuity, and object identity.

Annotation interface. Annotators were shown the initial image, applied-force visualization, and candidate videos for the same condition, and assigned separate H-Force and H-Physics scores according to Table 11. The evaluation interface is shown in Figure 9.

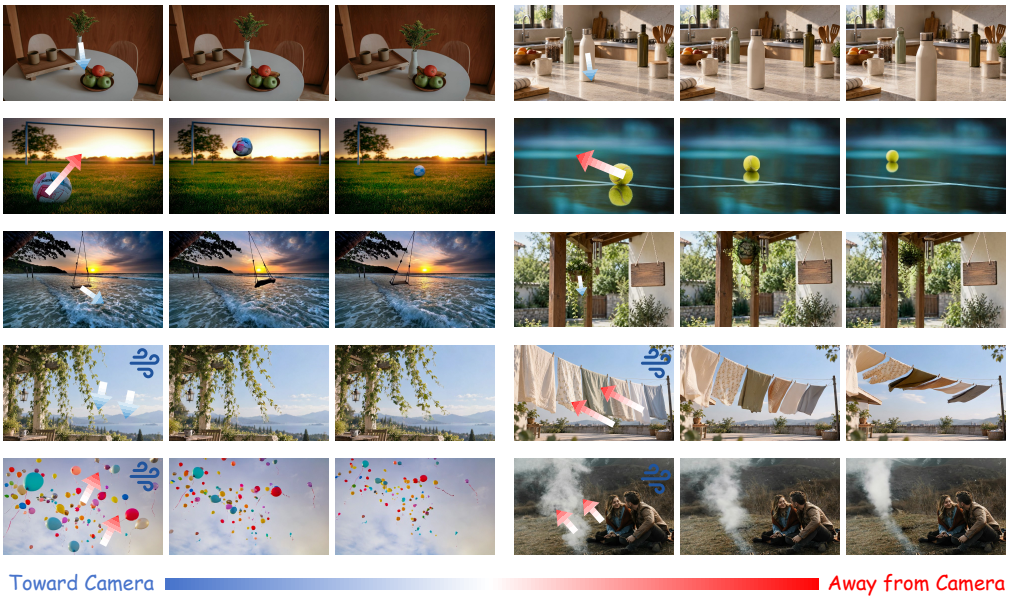


Figure 10: **Additional qualitative results of ForceVid3D.** Each triplet shows the initial frame with the prescribed 3D force followed by two generated frames. Arrow orientation indicates the projected image-plane direction, while color encodes the depth component: blue points toward the camera and red points away from it.

Scores are aggregated per output and then macro-averaged across cases. Inapplicable measurements are excluded according to predefined case semantics.

D.4 Visual Quality

We evaluate visual quality using the `aesthetic_quality` and `imaging_quality` dimensions of VBench (Huang et al., 2024), rather than its aggregate score. All scores are computed on clean generated videos without force overlays or diagnostic visualizations, using the official VBench implementation.

All methods are evaluated on the same predefined ForceVid3D-Bench cases. The shared 2D evaluation contains 90 point-force and 66 wind-force cases per applicable method; Goal Force is evaluated on the 90 point-force cases because its released interface does not support wind. The 3D visual-quality evaluation uses the same automatically evaluated subset as the response metrics, containing 27 point-force and 5 wind cases.

Frame-level scores are averaged within each video and then macro-averaged over evaluation cases. For methods supporting both point force and wind, the pooled 2D result in Table 1 is weighted by the corresponding case counts.

These metrics complement the force-adherence and physical-consistency measurements by assessing perceptual quality.

E Additional Qualitative Results and Failure Cases

E.1 Qualitative Results

Figure 10 presents additional ForceVid3D generations across diverse objects and scenes under camera-space 3D force control. The examples include localized point forces and distributed wind with varied 3D directions.

E.2 Failure Cases

Figure 11 illustrates four representative failure modes. *No response* occurs when the target exhibits little or no response to the applied force, which may arise when the corresponding response pattern is weakly represented in the pretrained motion prior or training data. *Incorrect response association* activates a nearby object or background region instead of, or in addition to, the intended content, reflecting the ambiguity of inferring responsive surfaces and object constraints from a single image. *Trajectory drift* occurs when the initial response is consistent with the prescribed force but subsequent motion deviates from a plausible continuation, potentially due to limitations in the inherited motion prior as the sequence evolves. Finally, *physical inconsistency* includes implausible contact transfer, excessive deformation, splitting, duplication, disappearance, or identity drift. Such failures can arise under complex contact or strong deformation, where the video prior does not explicitly enforce physical constraints.

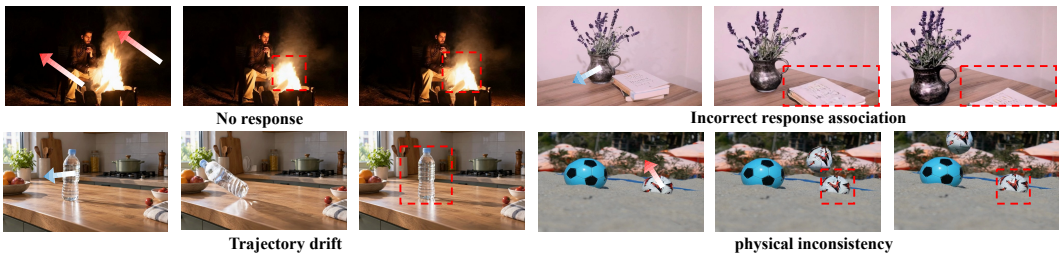


Figure 11: **Representative failure cases of ForceVid3D.** Each triplet shows the initial frame with the prescribed force followed by two generated frames. Top left: the flame remains largely upright despite the prescribed force. Top right: the book moves although the force targets the vase. Bottom left: the bottle tilts left in the first generated frame but is upright again in the second. Bottom right: duplication of the targeted ball produces two instances, including one near its initial position. Red dashed boxes highlight the affected regions.

F Implementation and Reproducibility Details

F.1 Baseline Details

We compare with Force Prompting (Gillman et al., 2025), StreamForce (Wang et al., 2026), and Goal Force (Gillman et al., 2026) using the authors' released implementations and pretrained checkpoints. Force Prompting uses the released CogVideoX-5B-I2V model and force ControlNet, StreamForce uses its Wan2.2-TI2V-5B causal force-control model, and Goal Force uses its Wan2.2-I2V-A14B high-noise-expert ControlNet. We retain each method's native force encoding and inference schedule; we do not replace a baseline backbone or retrain it on ForceVid3D-Dataset. Every comparable run receives the same initial image, text description, and image-plane force condition. Goal Force is reported only for point force because its released interface does not support wind.

Table 12: **Baseline configurations used in the ForceVid3D-Bench comparison.**

Method	Backbone	Force Interface	Evaluation Setting
Force Prompting	CogVideoX-5B-I2V	Point / Wind	2D force
StreamForce	Wan2.2-TI2V-5B	Point / Wind	2D force
Goal Force	Wan2.2-I2V-A14B	Point	2D force
ForceVid3D	Wan2.2-TI2V-5B	Point / Wind / 3D force	2D + 3D force

F.2 Model and Training Configuration

Table 13 summarizes the backbone, controller, and base-SFT optimization settings.

The controller is initialized by copying the first ten backbone Transformer blocks. A zero-initialized linear projection follows each control block and injects its residual into the corresponding backbone

Table 13: **Base SFT model and flow-matching configuration.**

Model setting	Value	Training setting	Value
Video backbone	Wan2.2-TI2V-5B (Wan et al., 2025)	Optimizer	AdamW
Resolution	640×384	Peak learning rate	1×10^{-5}
Frames / FPS	49 / 10	Weight decay	1×10^{-4}
Controller depth	10 blocks	Global batch size	8
Force representation	Projective, 7 channels	Optimizer steps	8,000
Trainable parameters	1.731B (controller)	Adam (β_1, β_2)	$(0.9, 0.95)$
Frozen modules	Backbone, VAE, text encoder	Gradient clipping	1.0
Training hardware	$4\times$ Ascend 910B3	Precision	BF16 / FP32
Per-device batch / accumulation	1 / 2	LR schedule	250-step warmup, cosine

block. The seven-channel force tensor has shape $7 \times 13 \times 24 \times 40$ for the default video setting. A spatial 2×2 PixelUnshuffle converts it to $28 \times 13 \times 12 \times 20$, after which a $1 \times 1 \times 1$ convolution maps it to the 3,072-dimensional token space. The controller receives the sum of this encoding and detached scene features from the backbone. During base SFT, 1.731B controller parameters are trainable while the pretrained video backbone remains frozen.

Base SFT updates the force stem, control blocks, residual projections, and force-condition embeddings while keeping the Wan backbone, VAE, and text encoder frozen. Trainable parameters and AdamW states are stored in FP32, while frozen weights and forward computation use BF16. The learning rate is warmed up linearly for 250 steps and then cosine-decayed from 10^{-5} to 5×10^{-6}. We use AdamW with $\epsilon = 10^{-8}$, gradient checkpointing, and activation offloading, without EMA.

The matched D-RAS branch continues from the 8K SFT checkpoint to 11K while retaining the same frozen backbone. Independent response and relative-depth heads operate on the probe-selected B4 and B9 features, and their predictions are averaged across levels, with response logits averaged before applying the sigmoid. The D-RAS heads are jointly optimized with the controller and removed at inference. The final ForceVid3D checkpoint continues training to 15K, while component attribution uses the matched B1/B2 continuation from 8K to 11K.

The base SFT and matched continuation mixture contains 90% force-labelled simulated videos and 10% unlabelled real wind clips. The latter use a learned NULL-FORCE embedding rather than an artificial numerical force. When exact simulation masks are available, the native flow-matching loss is augmented with a 0.2-weighted object-region term. This region is formed from the union of each object's first-frame and future masks, dilated by one latent pixel, smoothed with a Gaussian of $\sigma = 1$, and normalized per visible object and frame. Real and mask-free wind videos use the global flow-matching loss.

Inference. All ForceVid3D evaluations, including the final model on ForceVid3D-Bench and matched checkpoints on the controlled synthetic evaluation set, use 50 function evaluations (NFE) and a classifier-free guidance (CFG) scale of 5.0. For counterfactual cases, the initial frame, text prompt, force magnitude, application support, sampler configuration, and random seed are held fixed while only the force direction changes.

Compute. Base SFT and D-RAS are trained with PyTorch DDP through Accelerate on four Ascend 910B3 accelerators with 64 GiB HBM per device. Training uses BF16 forward computation with FP32 optimizer states. D-RAS heads are removed at inference, so SFT and D-RAS use the same generation architecture and sampler budget for a fixed force representation.

Reproducibility. We will release the synthetic-data pipeline, training and inference code, force annotations, benchmark evaluation code, model checkpoints, and experiment configurations.

G Limitations and Future Directions

Physical priors and response coverage. ForceVid3D inherits the motion and physical priors of its pretrained video backbone. It can therefore fail on response modes that are poorly represented in either the backbone or ForceVid3D-Dataset, especially complex contact and strong deformation. Expanding the coverage of interaction types and studying transfer across related response modes are important directions for future work.

No numerical-physics guarantee. The method generates visually plausible responses to explicit 3D forces but is not a numerical simulator. Its projective representation does not recover metric depth, mass, friction, or other latent physical parameters, and the generated dynamics are not guaranteed to satisfy Newtonian mechanics exactly. Incorporating estimated scene properties or simulator-derived constraints may improve physical calibration.

Interaction horizon and control scope. Our experiments focus on 49-frame videos with a single externally specified intervention. Extending the framework to long-horizon dynamics, sequential or simultaneous forces, torque, and closed-loop interactive control remains future work.