

Goal Force: Teaching Video Models To Accomplish Physics-Conditioned Goals

Nate Gillman¹ Yinghua Zhou¹ Zitian Tang¹ Evan Luo¹ Arjan Chakravarthy¹
Daksh Aggarwal¹ Michael Freeman² Charles Herrmann¹ Chen Sun¹

Brown University¹ Cornell University²

{nate_gillman, yinghua_zhou, zitian_tang, chensun}@brown.edu

Abstract

Recent advancements in video generation have enabled the development of “world models” capable of simulating potential futures for robotics and planning. However, specifying precise goals for these models remains a challenge; text instructions are often too abstract to capture physical nuances, while target images are frequently infeasible to specify for dynamic tasks. To address this, we introduce *Goal Force*, a novel framework that allows users to define goals via explicit force vectors and intermediate dynamics, mirroring how humans conceptualize physical tasks. We train a video generation model on a curated dataset of synthetic causal primitives—such as elastic collisions and falling dominos—teaching it to propagate forces through time and space. Despite being trained on simple physics data, our model exhibits remarkable zero-shot generalization to complex, real-world scenarios, including tool manipulation and multi-object causal chains. Our results suggest that by grounding video generation in fundamental physical interactions, models can emerge as implicit neural physics simulators, enabling precise, physics-aware planning without reliance on external engines. We release all datasets, code, model weights, and interactive video demos at our project page, <https://goal-force.github.io/>.

1. Introduction

The past two years have witnessed a paradigm shift in video generation, evolving from coarse, rudimentary clips to near-photorealistic sequences [1, 5, 8]. This progress has sparked considerable interest in leveraging these models as “world models” for robotics and planning. One of the most exciting possibilities for using “world models” in planning involves generating a video that transitions from a current state (an initial frame) towards a specified goal state [19, 27]. Consider a soccer player at the start of a game: the initial frame shows the ball at midfield, and the objective is to score. Existing approaches predominantly rely on text or static im-

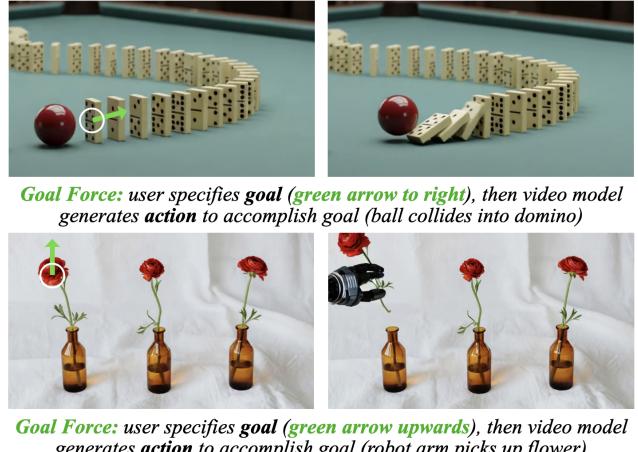


Figure 1. Given a force-conditioned task, **goal force** enables video models to generate the antecedent action to accomplish the task.

ages to define these goals. However, for complex physical tasks involving multi-step dynamics, these modalities often prove insufficient. Text is frequently too abstract; a soccer player’s intent is rarely just to “shoot at the goal,” but rather to strike the ball with specific force and precision. Conversely, specifying a goal via a target image is often overly burdensome or infeasible—potentially requiring a user to render the exact lighting of a ball entering the net.

In contrast, humans approach tasks differently than through abstract text or pixel-perfect images alone. We often decompose long, abstract tasks into concrete sub-goals that, particularly in sports, possess distinct physical properties like spatial location, dynamics, and motion. When taking a penalty kick, a soccer player does not focus merely on the static end state of the ball in the net, nor do they simply rely on the abstract concept of scoring. Instead, they aim to impart a specific trajectory and velocity—a “goal force”—onto the ball. This paper proposes a method that aligns with this intuition: defining goals through desired forces and intermediate dynamics. By specifying these goal

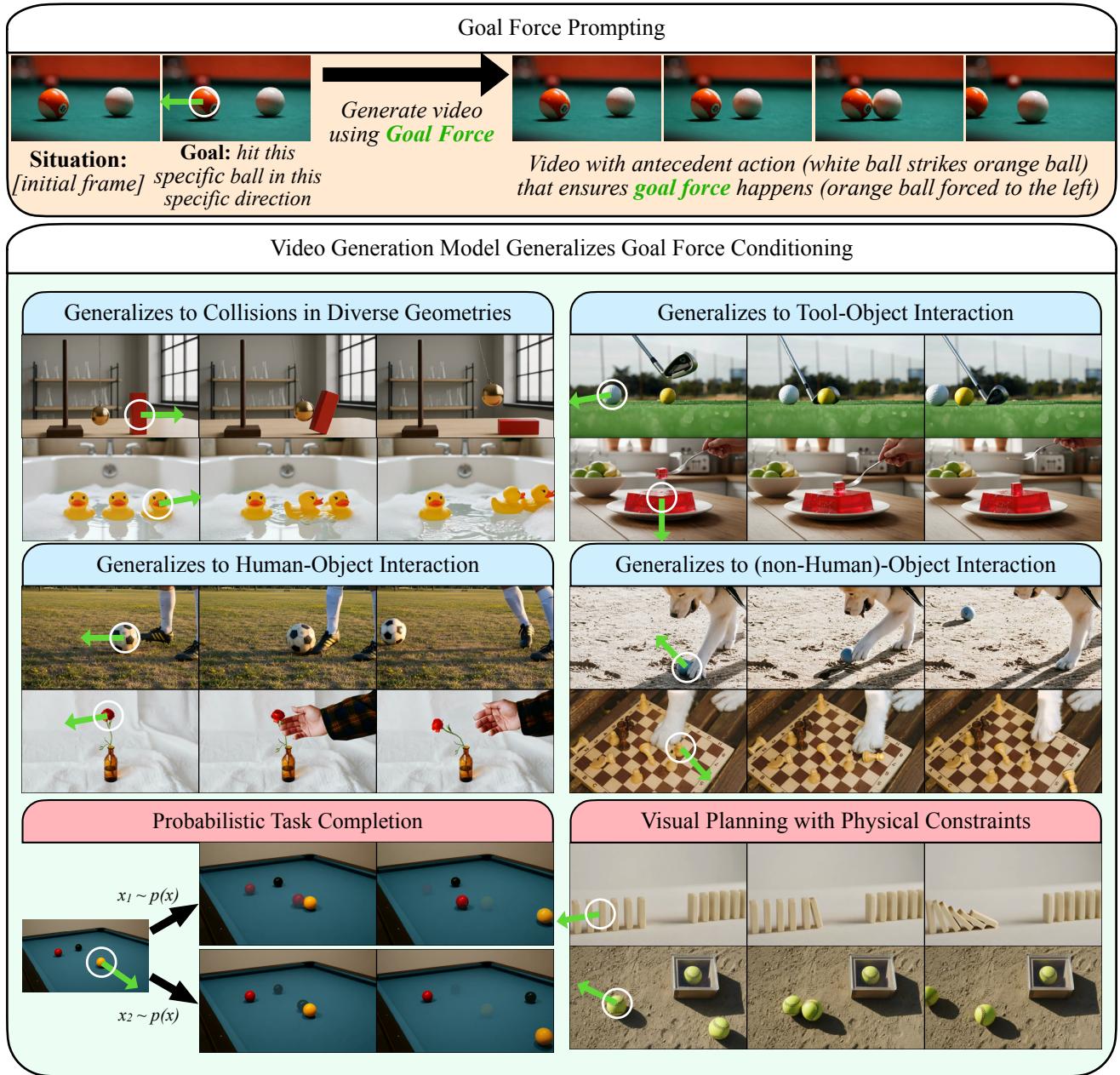


Figure 2. **Goal Force**: A user provides an input image and a **goal force**, and the model generates a video containing a force that locally causes the goal force. Our model generalizes to diverse objects and interactions and enables visual planning, respecting the physical properties of the objects and their environments.

forces, rather than limiting users to static endpoints or requiring direct, low-level scene manipulation, we offer a mechanism that is both precise enough for physics-based planning and intuitive for human users.

To accomplish this, we introduce a framework that conditions video generation on explicit goal force vectors. We curate a dataset of paired videos and “goal forces,” adapting a state-of-the-art open-source video model to accept these forces as a control signal. Our training strategy relies on

the hypothesis that learning fundamental physical interactions can bootstrap complex reasoning. We train the model on simple, synthetic examples of causal primitives, such as elastic collisions and falling dominos. Crucially, we find that this grounded training enables *non-trivial generalization* to highly diverse scenarios (Figure 2).

Our empirical results demonstrate that the model learns to propagate forces through time and space, handling chains of events where one object exerts force on another, which in

turn influences a third. Remarkably, this capability extends to zero-shot tool usage; for instance, the model can infer how to use a golf club to impart the desired force onto a ball, and to pick up a rose via its stem as opposed to its petals (Figure 1), despite only being trained on simpler collision data. This suggests the model is not merely memorizing patterns but acting as an implicit neural physics simulator.

Our main contributions are as follows:

1. We propose *Goal Force*, a new task and model which teaches video models to plan a causal chain of physical interactions to achieve a specified goal force. This moves beyond prior direct-force methods and changes how goals can be specified in world models.
2. We propose a training paradigm with a novel multi-channel control signal (for goal forces, direct forces, and mass) that teaches the model to act as an implicit neural physics simulator, requiring no simulator at inference.
3. We demonstrate powerful out-of-domain generalization: despite training only on simple synthetic data (*e.g.*, balls, dominos), our model leverages the base video model’s rich prior to generate complex, physically-plausible scenarios involving tool use, human-object interaction, and intricate multi-object collisions.

We release training and evaluation code, model weights, synthetic training data, and benchmark datasets at our project page, <https://goal-force.github.io/>.

2. Related Works

Video generative models: In recent years, video generation models have achieved remarkable progress in visual fidelity and the plausible rendering of complex dynamics [5, 7, 18, 22, 46]. The introduction of models like Sora [8] highlighted the potential for using large-scale generative models as “world simulators” capable of rendering diverse physical phenomena. This progress has been mirrored in open-source efforts [51, 60, 61], which are increasingly approaching the quality of closed-source systems. While these models serve as powerful video priors, they are typically conditioned on text or images and lack interfaces for fine-grained, precise control over physical actions or interactions, which is a gap our work aims to address.

Controllable video generation: To address the need for greater control, many methods have been proposed. A significant portion of this research focuses on controlling the camera perspective [21, 47, 69]. Another major direction is motion control, which uses various paradigms like drag-based editing [57, 62], trajectory specification [12, 41, 67], or optical flow guidance [32, 42, 45]. A limitation of many of these techniques [62, 67] is their reliance on densely specified, complete trajectories, which makes them unsuitable for predictive tasks where the full motion is unknown. Prior work like Motion Prompting [16] allows for sparse

trajectory inputs, but this still specifies motion rather than its underlying cause. More recently, Force Prompting [17] introduced direct physical control by specifying a force vector. However, all these methods focus on direct, immediate interventions. Our work, *Goal Force*, moves beyond this by enabling the model to reason about and plan a causal chain of forces: for example, hitting ball A in order to achieve a desired goal force on ball B.

Physics simulators and hybrid approaches: There is a long history of attempting to model physics from video. Early work [6, 33] focused on extracting intuitive physical properties, such as the modal bases of vibrating objects, but these methods struggle to represent general motion. An alternative research line incorporates explicit physics simulation [2, 11, 24, 29, 36, 37, 52, 55, 58, 66, 70]. While physically accurate, these approaches generally require access to 3D geometry, which is often unavailable. Hybrid models represent a compromise, as they combine physics simulators for dynamics with generative models for appearance [34, 38, 48]. A key limitation is that these models are constrained by the capabilities of their internal simulator (*e.g.*, rigid bodies only) and require it at inference time. More recent works have removed this dependency on internal simulators [44, 54] and can learn better representations of physical properties [25, 63], but these works focus on local physical properties rather than causal interactions. Concurrent works have also explored using simulated data to fine-tune models for freefall [30] or learning 3D trajectories [53]. Our approach differs fundamentally: we do not use any physics simulator at inference time. Instead, we train the generative model itself to act as an approximate “neural simulator” that can reason about and plan causal interactions to achieve a specified goal.

Interactive world models: The concept of a “world model” [20, 56] that can learn to simulate and interact with an environment has gained significant traction. To date, investigations have largely concentrated on video game environments [9, 10, 26, 50]. While some recent studies have begun to explore real-world applications [1, 4, 19, 31, 64], the forms of interaction are typically limited to text prompts or camera navigation. In contrast, our work introduces a new, physically-grounded form of interaction. By allowing a user to specify a goal force, we push the model to reason about physical cause-and-effect and plan the antecedent actions (like tool use or multi-object collisions) necessary to achieve that goal, representing a step towards more capable and physically-aware interactive world models.

Planning with videos: Video models have been applied to solve decision-making problems in robotic applications [35, 40]. A video generative model can serve as reward functions [15, 23], dynamics models [49, 59], and pixel-based planners [3, 28, 71]. For example, UniPi [14] and Adapt2Act [39] employ text-conditioned video gener-

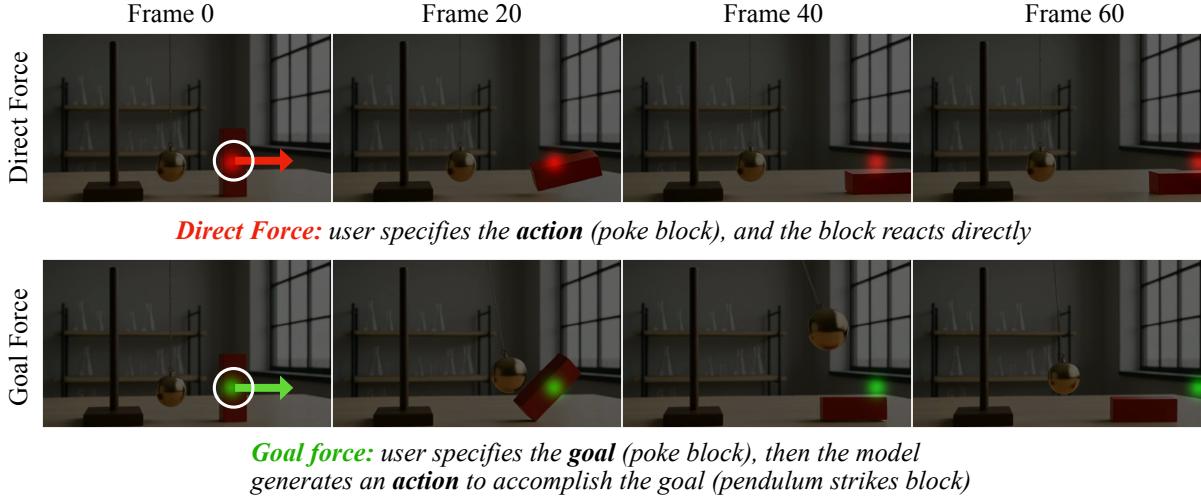


Figure 3. **Force prompt and goal force prompt result in different behaviors.** With a **direct force** applied to the red block (top), the effect is directly materialized (i.e. the block falls over). The force in this case is encoded in the red channel of the control signal as a moving Gaussian blob. In contrast, with a **goal force** applied to the red block (bottom), the model must find the antecedent motion to achieve the goal force (i.e. the pendulum swings to knock over the block). The force in this case is encoded in the green channel of the control signal as a moving Gaussian blob. We visualize the control signal overlaid on top of the video via alpha blending.

ative models to predict visual plans that depict future outcomes, which are then converted into robotic actions with inverse dynamics models. With our introduced framework, such visual planners can take goal forces, in addition to text, to specify the desired goals.

3. Method: Prompting with Goal Force

Our method reframes force-conditioned video generation from specifying a *direct force* (e.g., [17, 38, 66]) to declaring a desired *goal force*. Given a starting frame ϕ and a text prompt τ , the user specifies a “goal force” on a target object (make ball B move right). The model’s task is to generate a video v that synthesizes a physically-plausible *antecedent causal chain* (ball A striking ball B) to achieve that goal.

We achieve this by training a video generative model to act as an implicit neural physics planner. The core of our approach is a novel training paradigm built on a multi-channel physics control signal and a curriculum of synthetic data.

3.1. Multi-Channel Physics Control Signal

We introduce a 3-channel physics control tensor $\tilde{\pi} \in \mathbb{R}^{f \times 3 \times h \times w}$, where f is the number of frames, h and w are the spatial dimensions, and each of the 3 channels encodes a specific physical property. This tensor $\tilde{\pi}$ is the spatial-temporal encoding of the abstract user prompt.

Channel 0: Direct Force. Encodes an immediate, direct force (the “cause”). Following [17], we represent this as a “moving Gaussian blob” video, where the blob’s trajectory and duration are affinely proportional to the force vector (location, angle, and magnitude).

Channel 1: Goal Force. Encodes the desired *outcome* (the “effect”) on a target object. This channel uses the *same* moving Gaussian blob representation to specify the desired force (and resulting motion) on the *target* object. We visualize the practical difference between a Goal force and a Direct force in Figure 3.

Channel 2: Mass. Encodes privileged physical information, such as relative object mass. We represent this as a *static* Gaussian blob in this channel, centered on the object, with a radius affinely proportional to its mass. The mass signal is optional, and offers an interface for users to provide more fine-grained, object-level physical properties, when they are available. When not provided, Goal Force can instead resort to the physical priors encoded in video generative models themselves, a behavior referred to as “mass understanding” in [17].

Force and Mass Normalization. We note that force and mass values are not calibrated to an absolute physical scale. Instead, they follow an intuitive, relative scale normalized *within* each synthetic dataset (dominos, balls, plants). Our model learns this relative concept, as the Gaussian blob encoding is also defined proportionally to the value range of a given domain. This allows the model to generalize the *idea* of force (e.g., “small poke” vs. “large poke”) without requiring a unified, absolute scale.

3.2. Goal Reaching via Implicit Planning

We train the model on a synthetic dataset of simple causal chains (colliding balls, falling dominos) and complex dynamics (swaying flowers), generated using Blender and PhysDreamer [66]. This dataset contains three scenarios:

Goal Force is a “*how-to*” machine, generating the antecedent action in order to accomplish the force-conditioned task (**arrow**).

Prior methods are “*what if*” machines, directly accomplishing the force-conditioned task (**arrow**) without any implicit planning.

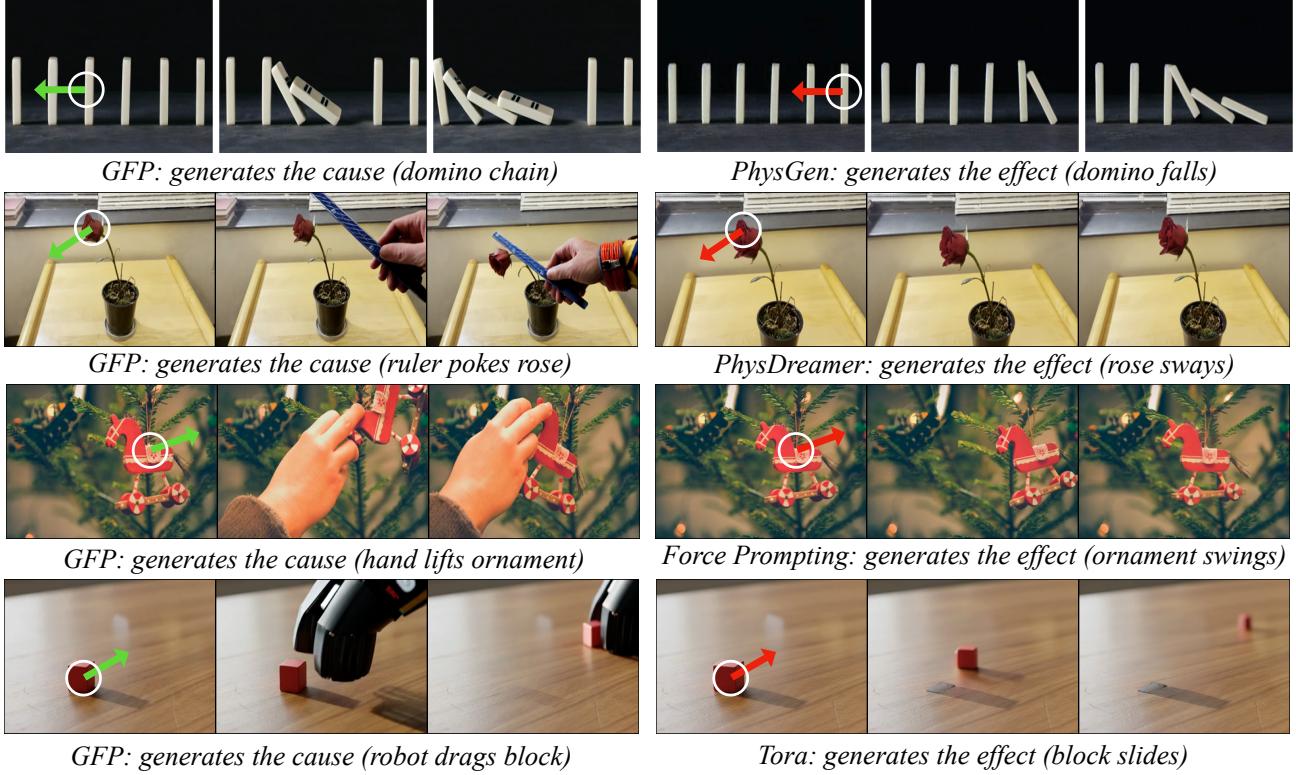


Figure 4. In prior methods (right), the user provides a force, and the model directly applies the force to the target object. In our method (left), the user provides a goal force, and the model generates the causes that achieve the desired effect on the target object. The top three methods (PhysGen [38], PhysDreamer [66], and Force Prompting [17]) all accept forces as conditioning; the fourth method, Tora [67], accepts trajectories rather than forces, so we condition on an acceptable trajectory.

- **Dominoes (3k videos):** Generated in Blender, these videos show a line of dominos where a direct force on one initiates a chain reaction, linking the “cause” to a “goal force” on a downstream domino.
- **Rolling Balls (6k videos):** Blender scenes of multiple balls. A direct force is applied to a “projectile” ball, which is aimed to either collide with a “target” ball (4.5k videos) or miss it (1.5k videos).
- **PhysDreamer Carnation (3k videos):** Videos of a flower swaying after being poked, generated with PhysDreamer [66], a method that integrates 3D Gaussians and a physics simulator. This component teaches the model complex, non-rigid dynamics from a direct force.

Full data generation details are in Appendix 7.2.

This synthetic data for the ball collisions and domino collisions provides ground-truth pairs of (direct force, resulting goal force). Our key training strategy is to ran-

domly mask the causal information. For each training video, we provide *either* the direct force (in Ch 0) *or* the goal force (in Ch 1), zeroing out the other. (And in scenes without collisions, namely 1/4 of the ball scenes and all the plant scenes, we only provide the direct force in Ch 0.) This forces the model to learn the physical reasoning:

- **Goal → Plan:** Given a goal force, the model must infer and generate the antecedent direct-force event.
- **Action → Outcome:** Given a direct force, the model must simulate the resulting collision and secondary force.

The mass channel (Ch 2) is also randomly masked during training. This teaches the model to leverage privileged physics information when available but also to rely on its internal, learned physics prior to estimate properties (like mass) from appearance when it is not. The text prompt’s role is to set the semantic context (*e.g.*, “a pool table”) and guide the model toward a plausible distribution of videos. It

Benchmark	Two Object Collision			Multi Object Collision			Human Object Interaction			Tool Object Interaction		
	Force Adh.	Real. Motion	Visual Qual.	Force Adh.	Real. Motion	Visual Qual.	Force Adh.	Real. Motion	Visual Qual.	Force Adh.	Real. Motion	Visual Qual.
Text-only, zero-shot	57%	57%	49%	60%	63.3%	76.7%	77.8%	53.3%	55.6%	96.7%	56.7%	70%
Text-only, fine-tuned	66%	46%	49%	60%	86.7%	66.7%	57.8%	47.8%	45.6%	70%	46.7%	50%

Table 1. **Human study comparing Goal Force method to text-only baselines.** Numbers indicate the percentage of human pairwise preferences for Goal Force Prompting over each text-only baseline on each benchmark dataset. The proposed model consistently yields superior goal force adherence against both baselines, with minimal degradation of motion realism and visual quality.

does not, however, specify the low-level causal plan, such as which ball should strike another. This ambiguity is intentional: it forces the model to leverage its internal prior to plan a valid antecedent action, constrained only by the specific objective of the goal force prompt.

3.3. Architecture and Training Details

We build our model on Wan2.2 [51], a Mixture-of-Experts diffusion model. We use a ControlNet [65] module to condition on our physics signal $\tilde{\pi}$. We fine-tune this ControlNet only for the high-noise expert, as this expert is primarily responsible for global structure and low-frequency dynamics [13], which aligns with our physics-planning task. The ControlNet module clones the first 10 DiT layers from the pretrained Wan2.2, fine-tuning them and feeding their outputs to the frozen base model via zero-convolutions. We encode the goal force prompt π using the frozen Wan2.2 encoder and pass the result through a randomly initialized patch embedding layer before feeding it to the ControlNet DiT layers. We fine-tune the model for 3,000 steps with an effective batch size of 4 (1 per device on 4 NVIDIA 80GB A100s), which completes in under 48 hours. We use videos of 81 frames at 16 FPS during training and inference.

4. Experimental Comparisons

4.1. Comparison to Text-Only Baselines

To evaluate Goal Force, we first compare to baselines that use text-only conditioning. We create a new benchmark of 25 challenging scenes curated from permissively licensed web sources {Pexels, Pixabay, Unsplash} as well as generative models {Nano-banana, GPT-Image-1}. We then conduct a 2AFC human study ($N = 10$) on Prolific, comparing our full Goal Force model against those baselines.

Baselines. We compare against two models:

1. *Text-only (Zero-shot):* Wan2.2 base model, prompted with a text suffix, e.g., “...a golf ball rolls across the grass, colliding with another ball. The secondary object is moved with very strong force to the left.”.
2. *Text-only (Fine-tuned):* Our ControlNet architecture finetuned on our synthetic data, but with the physics control signal zeroed out, relying only on the text suffixes provided during training.

Human Study for Generalization. Our benchmark spans four categories of increasing generalization from our training data: (1) two-object collisions (cantaloupes, pendulum striking object, pool ball, rubber duck toys in water, bars of soap, soccer balls, softballs), (2) multi-object collisions (ball colliding with domino, golf balls, tennis balls) (3) human-object interaction (hand interacting with ornaments, toy car; we also include in this category a dog interacting with a ball, and a cat knocking over a chess piece), and (4) tool-object interaction (golf club hitting golf ball, and a fork touching a dome of jello). Participants evaluated videos on three axes: *Goal Force Adherence* (Does the video accomplish the specified goal?), *Realistic Motion*, and *Visual Quality*.

Table 1 compares the performance of Goal Force against the text-only baselines. These results demonstrate that our model outperforms both baselines on goal force adherence, demonstrating that the text prompt is not sufficient, confirming that the explicit physics control signal is critical for solving the task. The results also demonstrate that this goal force adherence is achieved with minimal degradation of visual quality and motion realism. Despite training only on synthetic balls, dominos, and a single flower, our model generalizes effectively, enabling complex, out-of-domain interactions like tool use and human-object planning, as visualized in Figure 2.

4.2. Comparison to Prior Methods

The Goal Force prompting task is new, and prior force-conditioned methods (e.g., PhysGen [38], PhysDreamer [66], and Force Prompting [17]) are not designed to solve it. These models can only simulate a *direct* force (the cause), not plan the *antecedent* action required to achieve a *goal* force (the effect). As shown qualitatively in Figure 4, when given a goal force prompt, those prior methods misinterpret it as a direct, non-causal poke on the target object. Similarly, motion-conditioned models like ToRA [68] can follow a specified trajectory but fail to adhere to causality, often moving the target object before an antecedent event (like a hand) arrives. While prior methods cannot perform Goal Force prompting, our model is still capable of performing direct Force Prompting (FP). A qualitative comparison against these prior works is provided in Figure 4.

Table 2. Visual planning accuracy across scenes. Our model achieves a high success rate in selecting a physically valid force initiator across diverse, complex scenarios.

Scene	# Valid	# Success	% Accuracy
Dominos 1	50	50	100.00
Pool 1	22	12	54.55
Pool 2	49	48	97.96
Duckie 1	40	34	85.00
Duckie 2	37	24	64.86
Duckie 3	41	38	92.68

5. Goal Force Enables Visual Planning

We now evaluate a core claim of our work: that Goal Force enables a form of visual planning. We test this by analyzing three key properties of the generated plans: their physical accuracy, their diversity, and their awareness of privileged physics information such as mass.

5.1. Visual Plans are Accurate

We first test if the model’s visual planning adheres to physical constraints. We create a benchmark of scenes containing “natural blockers” (Figure 5), where distractor objects are physically constrained from initiating the goal force. A successful plan requires the model to identify and select a valid, unconstrained object to execute the causal chain.

For each scene, we generate 50 videos. To isolate the planning logic from the base video diffusion model’s artifacts, we filter out trials exhibiting stochastic visual degradation (*e.g.*, object hallucination) prior to analysis. We define **accuracy** as the percentage of valid trials where the goal force is initiated by the correct, unconstrained object, rather than by a distractor or through spontaneous, non-causal motion.

Results. We report accuracy for each scene in Table 2. A random baseline achieves at most 33.3% accuracy given our distractor design. The model demonstrates strong physical reasoning across most of the scenes. In the pool example (Fig. 5, top), a stick blocks the orange ball. Our model correctly selects the white ball as the initiator in 98% of valid trials. On the rubber duckie benchmark (Fig. 5, bottom), it selects the correct initiator. We observe that most failure cases involve the target object moving spontaneously, rather than the model choosing an incorrect, constrained initiator. We also observe this trend of physically grounded visual planning generalizes to other natural scenarios, including the ones shown in Figure 2.

5.2. Visual Plans are Diverse

Beyond accuracy, we test if our model produces a *diverse* set of valid plans rather than suffering from mode collapse.

Table 3. Diversity metric ($\delta(p)$) scores for the 5-domino task. Higher is better (Max: 1.0). Our model (0.6577) shows significant diversity compared to the deterministic baseline (0.3900).

Distribution (p)	Score ($\delta(p)$)
Our Model (Goal Force)	0.6577
Reference: $\text{Unif}\{0..4\}$ (Max diversity)	1.0000
Reference: $\text{Unif}\{0..3\}$	0.8920
Reference: $\text{Unif}\{0..2\}$	0.7635
Reference: $\text{Unif}\{0..1\}$	0.6042
Reference: $\text{Unif}\{0\}$ (Deterministic)	0.3900

We design a multi-modal task: a line of six dominos where the goal is to topple the rightmost (sixth) domino block. This goal can be achieved by initiating a chain reaction from any of the five preceding dominos. A deterministic model would repeatedly target the same domino, whereas we hypothesize Goal Force will sample from the full distribution of valid plans.

A non-diverse or deterministic model would exhibit mode collapse, targeting the same domino repeatedly. We hypothesize that our Goal Force model will instead sample from a diverse distribution of valid initial actions. To quantify this, we propose a diversity metric $\delta(p)$ based on the Jensen-Shannon Divergence (JSD). Let $\hat{p}(x)$ be the empirical probability mass function (PMF) over the set of the $N = 5$ targetable dominos, $S = \{0, 1, 2, 3, 4\}$. We define our **diversity metric** as:

$$\delta(p) = 1 - \text{JSD}(\hat{p} \parallel \text{Unif}(S)). \quad (1)$$

This metric is normalized to provide an interpretable score. A perfectly diverse model sampling uniformly from all 5 dominos ($\hat{p} = \text{Unif}(S)$) achieves the maximum score of $\delta(p) = 1.0$. Conversely, a fully deterministic model exhibiting complete mode collapse (*i.e.*, \hat{p} is a Dirac delta function on a single domino) yields the baseline score of $\delta(p) \approx 0.39$.

Results. We present our findings in Table 3. Across 26 random seeds, our model achieves a diversity score of 0.6577, significantly higher than the deterministic baseline (0.3900) and distinct from distributions with collapsed support (*e.g.*, $\text{Unif}\{0, 1\}$). This demonstrates that our model successfully explores a multi-modal distribution of valid plans rather than collapsing to a single solution.

5.3. Visual Plans Leverage Privileged Physics

Next, we test if the model’s visual plans can use privileged mass information provided in the control signal to help guide their plans. Our experiments focus on ball collision. In this setting, a physically-grounded plan must account for mass; for example, achieving a specific goal force on a heavier target requires a stronger impact.

Goal Forces Enable Visual Planning With Respect To Physical Constraints



*In order to move the red ball in the direction of the **goal force prompt**, it must be hit with the white ball rather than the orange ball, because the path from the orange ball is blocked by the pool stick.*



*In order to move the rubber duck in the direction of the **goal force prompt**, it must be hit by the center duck, because the path from the other duck is blocked by a concrete barrier.*

Figure 5. **Given a goal force prompt, the model chooses the physically correct way to execute it.** Top: even though there exist multiple plausible initiators, the model correctly selects the initiator to achieve the desired force on the target. Bottom: With multiple plausible rubber ducks that could initiate the force, the model selects the initiator that is not blocked by a physical barrier.

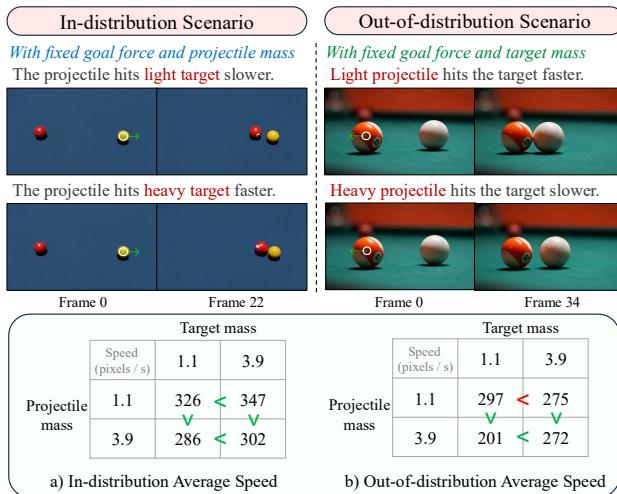


Figure 6. **Visual plans take advantage of mass information.** We test goal force prompting on in-distribution (left) and out-of-distribution (right) scenarios. In both scenarios, our model can adjust the moving speed of the projectile accordingly when the object masses are changed to cause the desired force magnitude. The direction of the “ $<$ ” sign indicates the desired numerical relationship; **green** indicates satisfaction, **red** indicates violation.

We design a ball collision task with a fixed goal force magnitude, varying the projectile and target masses. In Figure 6, we test our model on two such scenarios. One is in-distribution with a scene and viewpoint similar to our training data. The other features an out-of-distribution background, viewpoint, lighting, and ball size. We expect the model to learn two principles: (1) if projectile mass is constant, a heavier target requires a faster projectile; (2) if target

mass is constant, a heavier projectile can move slower.

To quantitatively measure the ball collision, we use Faster R-CNN [43] to detect the positions of the two balls. Then we determine the collision time and compute the projectile’s moving speed accordingly. We generate 15 videos for each combination of masses and average the speed over the samples. We observe that in the in-distribution scenario, the projectile’s speeds satisfy all four desired speed magnitude relationships. In the out-of-distribution scenario, our results satisfy three of them, while the fourth is very close. This demonstrates the model’s capability in leveraging privileged physics information for visual planning.

6. Conclusion

We introduce Goal Force, a paradigm that shifts generative video control from specifying a direct force (the cause) to declaring a desired goal force (the effect). We demonstrate that by training on simple, synthetic causal primitives, a video model can learn to function as an implicit neural physics planner. This enables the model to reason backward from a user-defined goal and generate a physically plausible, antecedent causal chain to achieve it. Our key finding is that this planning capability generalizes to complex, out-of-domain scenarios involving tool use and human-object interactions. This work represents a step toward interactive world models that can not only simulate a physical reaction but also reason about and plan the actions required to achieve a desired physical outcome.

Acknowledgements: We would like to thank Bill Freeman, Calvin Luo, David Fleet, and Miki Rubinstein for useful discussions. This material is based upon work partially

supported by the U.S. National Science Foundation under Cooperative Agreement No. 2433429. Our research was conducted using computational resources at the Center for Computation and Visualization at Brown University.

References

- [1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. [1](#) [3](#)
- [2] Luca Savant Aira, Antonio Montanaro, Emanuele Aiello, Diego Valsesia, and Enrico Magli. Motioncraft: Physics-based zero-shot video generation. *arXiv preprint arXiv:2405.13557*, 2024. [3](#)
- [3] Anurag Ajay, Seungwook Han, Yilun Du, Shuang Li, Abhi Gupta, Tommi Jaakkola, Josh Tenenbaum, Leslie Kaelbling, Akash Srivastava, and Pulkit Agrawal. Compositional foundation models for hierarchical planning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. [3](#)
- [4] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. *arXiv preprint arXiv:2412.03572*, 2024. [3](#)
- [5] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. [1](#) [3](#)
- [6] Peter W. Battaglia, Jessica B. Hamrick, and Joshua B. Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013. [3](#)
- [7] 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. [3](#)
- [8] Tim Brooks, Bill Peebles, Connor Homes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024. [1](#) [3](#)
- [9] 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 *Forty-first International Conference on Machine Learning*, 2024. [3](#)
- [10] Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. *arXiv preprint arXiv:2411.00769*, 2024. [3](#)
- [11] Hsiao-yu Chen, Edith Trechselk, Tuur Stuyck, Petr Kadlecak, Ladislav Kavan, Etienne Vouga, and Christoph Lassner. Virtual elastic objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. [3](#)
- [12] Tsai-Shien Chen, Chieh Hubert Lin, Hung-Yu Tseng, Tsung-Yi Lin, and Ming-Hsuan Yang. Motion-conditioned diffusion model for controllable video synthesis. *arXiv preprint arXiv:2304.14404*, 2023. [3](#)
- [13] Sander Dieleman. Diffusion is spectral autoregression. Blog post, 2024. Accessed: YYYY-MM-DD. [6](#)
- [14] Yilun Du, Sherry Yang, Pete Florence, Fei Xia, Ayzaan Wahid, brian ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B. Tenenbaum, Leslie Pack Kaelbling, Andy Zeng, and Jonathan Tompson. Video language planning. In *International Conference on Learning Representations (ICLR)*, 2024. [3](#)
- [15] Alejandro Escontrela, Ademi Adeniji, Wilson Yan, Ajay Jain, Xue Bin Peng, Ken Goldberg, Youngwoon Lee, Danijar Hafner, and Pieter Abbeel. Video prediction models as rewards for reinforcement learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. [3](#)
- [16] Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Carl Doersch, Yusuf Aytar, Michael Rubinstein, et al. Motion prompting: Controlling video generation with motion trajectories. *arXiv preprint arXiv:2412.02700*, 2024. [3](#)
- [17] 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, 2025. [3](#) [4](#) [5](#) [6](#) [1](#)
- [18] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duvval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023. [3](#)
- [19] Ziyu Guo, Xinyan Chen, Renrui Zhang, Ruichuan An, Yu Qi, Dongzhi Jiang, Xiangtai Li, Manyuan Zhang, Hongsheng Li, and Pheng-Ann Heng. Are video models ready as zero-shot reasoners? an empirical study with the MME-CoF benchmark. 2025. [1](#) [3](#)
- [20] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. [3](#)
- [21] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation, 2024. [3](#)
- [22] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Grishchenko, 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. [3](#)
- [23] Tao Huang, Guangqi Jiang, Yanjie Ze, and Huazhe Xu. Diffusion reward: Learning rewards via conditional video diffusion. *arXiv preprint arXiv:2312.14134*, 2023. [3](#)
- [24] Tianyu Huang, Haoze Zhang, Yihan Zeng, Zhilu Zhang, Hui Li, Wangmeng Zuo, and Rynson WH Lau. Dreamphysics: Learning physical properties of dynamic 3d gaussians with video diffusion priors. *arXiv preprint arXiv:2406.01476*, 2024. [3](#)
- [25] Sihui Ji, Xi Chen, Xin Tao, Pengfei Wan, and Hengshuang Zhao. PhysMaster: Mastering physical representation for video generation via reinforcement learning, 2025. [3](#)
- [26] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video

- generation from world model: A physical law perspective. 2025. 3
- [27] Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B Tenenbaum. Learning to act from actionless videos through dense correspondences. *arXiv preprint arXiv:2310.08576*, 2023. 1
- [28] Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B. Tenenbaum. Learning to act from actionless videos through dense correspondences. In *International Conference on Learning Representations (ICLR)*, 2024. 3
- [29] Simon Le Cleac'h, Hong-Xing Yu, Michelle Guo, Taylor Howell, Ruohan Gao, Jiajun Wu, Zachary Manchester, and Mac Schwager. Differentiable physics simulation of dynamics-augmented neural objects. *IEEE Robotics and Automation Letters*, 2023. 3
- [30] Chenyu Li, Oscar Michel, Xichen Pan, Sainan Liu, Mike Roberts, and Saining Xie. Pisa experiments: Exploring physics post-training for video diffusion models by watching stuff drop. *arXiv preprint arXiv:2503.09595*, 2025. 3
- [31] Yichen Li and Antonio Torralba. MultiModal action conditioned video generation. 2025. 3
- [32] Yaowei Li, Xintao Wang, Zhaoyang Zhang, Zhouxia Wang, Ziyang Yuan, Liangbin Xie, Yuexian Zou, and Ying Shan. Image conductor: Precision control for interactive video synthesis. *arXiv preprint arXiv:2406.15339*, 2024. 3
- [33] Zhengqi Li, Richard Tucker, Noah Snavely, and Aleksander Holynski. Generative image dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3
- [34] Zizhang Li, Hong-Xing Yu, Wei Liu, Yin Yang, Charles Hermann, Gordon Wetzstein, and Jiajun Wu. Wonderplay: Dynamic 3d scene generation from a single image and actions. *arXiv preprint arXiv:2505.18151*, 2025. 3
- [35] Junbang Liang, Ruoshi Liu, Ege Ozguroglu, Sruthi Sudhakar, Achal Dave, Pavel Tokmakov, Shuran Song, and Carl Vondrick. Dreamitate: Real-world visuomotor policy learning via video generation. *arXiv preprint arXiv:2406.16862*, 2024. 3
- [36] Jiajing Lin, Zhenzhong Wang, Shu Jiang, Yongjie Hou, and Min Jiang. Phys4dgen: A physics-driven framework for controllable and efficient 4d content generation from a single image. *arXiv preprint arXiv:2411.16800*, 2024. 3
- [37] Fangfu Liu, Hanyang Wang, Shunyu Yao, Shengjun Zhang, Jie Zhou, and Yueqi Duan. Physics3d: Learning physical properties of 3d gaussians via video diffusion. *arXiv preprint arXiv:2406.04338*, 2024. 3
- [38] Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision (ECCV)*, 2024. 3, 4, 5, 6, 1
- [39] Calvin Luo, Zilai Zeng, Yilun Du, and Chen Sun. Solving new tasks by adapting internet video knowledge. *arXiv preprint arXiv:2504.15369*, 2025. 3
- [40] Robert McCarthy, Daniel C.H. Tan, Dominik Schmidt, Fernando Acero, Nathan Herr, Yilun Du, Thomas George Thuruthel, and Zhibin Li. Towards generalist robot learning from internet video: A survey. *arXiv preprint arXiv:2404.19664*, 2024. 3
- [41] Koichi Namekata, Sherwin Bahmani, Ziyi Wu, Yash Kant, Igor Gilitschenski, and David B Lindell. Sg-i2v: Self-guided trajectory control in image-to-video generation. *arXiv preprint arXiv:2411.04989*, 2024. 3
- [42] Muyao Niu, Xiaodong Cun, Xiantao Wang, Yong Zhang, Ying Shan, and Yinqiang Zheng. Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. *arXiv preprint arXiv:2405.20222*, 2024. 3
- [43] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015. 8
- [44] David Romero, Ariana Bermudez, Hao Li, Fabio Pizzati, and Ivan Laptev. Learning to generate object interactions with physics-guided video diffusion. 2025. 3
- [45] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3
- [46] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 3
- [47] Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. *arXiv preprint arXiv:2411.04928*, 2024. 3
- [48] Xiyang Tan, Ying Jiang, Xuan Li, Zeshun Zong, Tianyi Xie, Yin Yang, and Chenfanfu Jiang. Phymotion: Physics-grounded dynamics from a single image. *arXiv preprint arXiv:2411.17189*, 2024. 3
- [49] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024. 3
- [50] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024. 3
- [51] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3, 6

- [52] 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. 2025. 3
- [53] Chen Wang*, Chuhao Chen*, Yiming Huang, Zhiyang Dou, Yuan Liu, Jiatao Gu, and Lingjie Liu. Physctrl: Generative physics for controllable and physics-grounded video generation. In *NeurIPS*, 2025. 3
- [54] Akihisa Watanabe, Jiawei Ren, Li Siyao, Yichen Peng, Erwin Wu, and Edgar Simo-Serra. SimDiff: Simulator-constrained diffusion model for physically plausible motion generation. 2025. 3
- [55] Jiajun Wu, Ilker Yildirim, Joseph J Lim, Bill Freeman, and Josh Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2015. 3
- [56] Jiajun Wu, Erika Lu, Pushmeet Kohli, Bill Freeman, and Josh Tenenbaum. Learning to see physics via visual de-animation. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 3
- [57] 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*, 2024. 3
- [58] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *CVPR*, 2024. 3
- [59] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 1(2):6, 2023. 3
- [60] Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. LongLive: Real-time interactive long video generation. 2025. 3
- [61] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenqi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3
- [62] 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. 3
- [63] Guanqi Zhan, Xianzheng Ma, Weidi Xie, and Andrew Zisserman. Inferring dynamic physical properties from video foundation models. 2025. 3
- [64] Jiahua Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal M Patel, Paul Pu Liang, Daniel Khashabi, Cheng Peng, Rama Chellappa, Tianmin Shu, Alan Yuille, Yilun Du, and Jieneng Chen. World-in-world: World models in a closed-loop world. 2025. 3
- [65] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 6
- [66] 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*. Springer, 2024. 3, 4, 5, 6, 1, 2
- [67] Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. *arXiv preprint arXiv:2407.21705*, 2024. 3, 5
- [68] Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2063–2073, 2025. 6
- [69] Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024. 3
- [70] Licheng Zhong, Hong-Xing Yu, Jiajun Wu, and Yunzhu Li. Reconstruction and simulation of elastic objects with spring-mass 3d gaussians. In *European Conference on Computer Vision*, 2024. 3
- [71] Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, D. Y. Yeung, and Chuang Gan. Robodreamer: Learning compositional world models for robot imagination. *arXiv preprint arXiv:2404.12377*, 2024. 3

Goal Force: Teaching Video Models To Accomplish Physics-Conditioned Goals

Supplementary Material

7. Additional Experiment Details

7.1. Comparison to Prior Works: Direct Force Prompting Quantitative Comparison

We encode the goal force prompt in the second channel of the control signal, and we encode the direct force prompt (which is a similar task to PhysDreamer [66], Force Prompting [17], and PhysGen [38]) in the first channel of the control signal. In Table 4 we compare the “Direct Force Prompting” capability of our model to those three models via a 2AFC human study ($N = 10$) conducted on Prolific. We gathered two benchmarks: a PhysGen benchmark, consisting of four scenes highlighted on that work’s project page; as well as a PhysDreamer benchmark, consisting of three scenes highlighted on that work’s project page. We compare our model to PhysGen and Force Prompting on the PhysGen benchmark, and we compare our model to PhysDreamer and Force Prompting on the PhysDreamer benchmark. Note that PhysGen models rigid body mechanics, whereas PhysDreamer models oscillations.

7.2. Synthetic Data Generation

In this section, we provide an in-depth discussion of the methods and specific parameters utilized for generating our synthetic training data. This data is used to train the Goal Force model to act as an implicit neural physics planner.

For all synthetic datasets, a key step in creating the multi-channel control signal $\tilde{\pi}$ is the projection of 3D forces and object properties onto the 2D image plane. We use the camera’s parameters to map 3D force vectors and object world coordinates into 2D pixel coordinates, enabling us to accurately model physical interactions within the video frames.

7.2.1. Dominos Dataset

We generated 3k videos of domino chain reactions using Blender. The setup models a causal chain where an initial direct force on one domino results in a predictable goal force on a downstream target domino.

To ensure diversity and robustness, we randomized the following parameters per video:

- **Domino Count:** Uniformly sampled from $\text{Unif}\{3, \dots, 10\}$.
- **Scene Geometry:** Randomized placement and orientation of the domino line.
- **Causality:** Choice of the initial target domino and the direction of the chain reaction (i.e., hitting the domino in front or behind).
- **Visuals:** Randomized camera position, ground textures (from 42 Polyhaven options), and High Dynamic Range

Images (HDRIs) for lighting and background (from 50 Polyhaven options).

- **Force Magnitude:** Continuous values from $[0, 1]$, where 0 represents the minimum force required for the domino to topple and 1 represents a maximal, strong impulse.

Each video is accompanied by a JSON file that records the names of the initial and adjacent contact dominos, along with the complete 2D pixel coordinates for all dominos across every frame.

7.2.2. Rolling Balls Dataset

This dataset comprises 6k videos generated in Blender, split into two primary categories to capture both collision and non-collision causal interactions:

1. Collision Set (4.5k videos): A “projectile” ball, acted upon by an unseen point force, is aimed to collide with one specific “target” ball within a group of initially stationary “distractor” balls.
2. Non-Collision Set (1.5k videos): The projectile ball is aimed such that it misses the target ball.

For the Collision Set, we ensured a diverse range of physical scenarios by randomizing:

- **Ball Count:** $\text{Unif}\{3, \dots, 9\}$.
- **Physical Properties:** Ball colors, ball masses $\text{Unif}(1.0, 4.0)$ kg, and all ball positions.
- **Visuals:** Randomized camera position and ground textures.
- **Force Calculation:** To guarantee collision, a minimum required force is calculated based on the projectile mass, distance to the target, and a randomized collision time ($\text{Unif}(2.5, 4.5)$ seconds). This minimum force is scaled by $\text{Unif}(1.2, 1.6)$ to introduce physical variation.

The collision videos are evenly split between straight-on and indirect collisions. For both types, the script first calculates the precise angular window necessary for the projectile to hit the target.

- For straight-on collisions, the force is aimed directly at the center of this calculated angular window.
 - For indirect collisions, the force angle is randomly sampled within this window, resulting in an off-center strike.
- This mixed-collision approach helps the model learn diverse post-collision behaviors.

For the Non-Collision Set, we randomized: ball quantity ($\text{Unif}\{3, \dots, 5\}$), ball textures, positions, camera angle, ground textures, target ball selection, force angle ($[0, 360^\circ]$), and force magnitude ($[0, 1]$).

For all ball videos, a JSON file records initial 2D/3D coordinates and physics parameters. For the videos featuring indirect collisions, we also save the complete 2D pixel tra-

Visual Quality	Dominos	Pool balls	Stone tower	Wall toy	Orange Rose	White Rose	Tulip
Force Prompting	90.0%	80.0%	60.0%	50.0%	100.0%	80.0%	60.0%
PhysGen	60.0%	100.0%	100.0%	80.0%	—	—	—
PhysDreamer	—	—	—	—	50.0%	50.0%	50.0%

Force Adherence	Dominos	Pool balls	Stone tower	Wall toy	Orange Rose	White Rose	Tulip
Force Prompting	90.0%	90.0%	80.0%	90.0%	50.0%	60.0%	50.0%
PhysGen	90.0%	80.0%	80.0%	40.0%	—	—	—
PhysDreamer	—	—	—	—	40.0%	30.0%	60.0%

Table 4. **Human study comparing the Direct Force capability of the Goal Force method to prior works.** Numbers indicate the percentage of human pairwise preferences for Goal Force Prompting’s direct force capability (i.e. encoding the force in the first channel) over each baseline on each benchmark dataset. The results demonstrate that Goal Force achieves consistently higher visual quality, as well as superior force adherence against the majority of baselines. Notably, our method achieves these results without relying on physics simulators or 3D assets at inference, unlike PhysDreamer and PhysGen. We note that PhysGen models rigid body mechanics, whereas PhysDreamer models oscillations, so they can’t be directly compared to one another.

jectory of the target ball. For the non-collision videos, we save the final 2D trajectory angle of the projectile ball.

7.2.3. Plants Dataset

This dataset, generated using PhysDreamer [66] (which integrates 3D Gaussians and a physics simulator), focuses on non-rigid body dynamics. The videos show a plant (carnation) swaying after being subjected to a direct force. We randomized the following parameters:

- **Initial Conditions:** Camera position and initial object configuration.
- **Force Application:** Contact points, force angles, and force magnitudes in $[0, 1]$, where 0 is a gentle poke and 1 is a strong impulse.