OptiWorld: Optimal Control for Video World Generation under Physical Constraints

Yu Yuan
Purdue University

Jianhao Yuan
University of Oxford

Xijun Wang
Purdue University

Daiqing Li
Sixteen Miles Labs

Liu He
Purdue University

Lu Ling
Purdue University

Stanley H. Chan
Purdue University

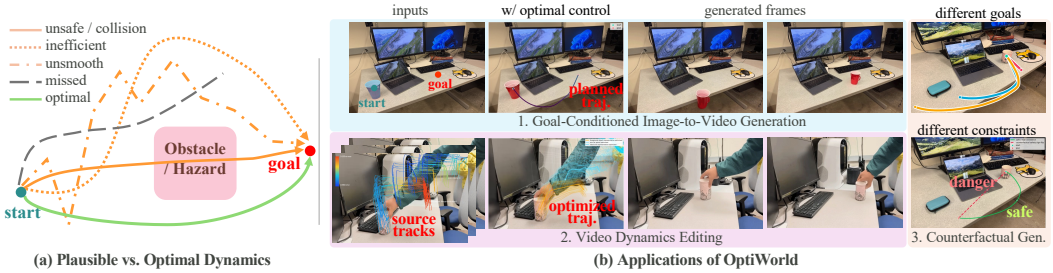


Figure 1: **Left:** without optimal control, a video generator may produce motions that look plausible but are unsafe, not smooth, or inefficient. **Right:** OptiWorld addresses these failures by introducing optimal control at video inference time: it plans a physically preferable trajectory before rendering. Zoom in for details of the planned/optimized trajectory.

Abstract

Video generation models are becoming a scalable form of world models, but they mainly generate plausible motion rather than proactively control or optimize the underlying dynamics. As a result, an object in the generated video may follow trajectories that are unsafe, not smooth, inefficient, or physically inconsistent. In this work, we propose **OptiWorld**, a framework that brings classical optimal control into video generation at inference time. OptiWorld first extracts a compact, task-relevant world state, then plans an optimal trajectory under physical constraints, and finally renders the video conditioned on this trajectory. We formulate planning as a geometric problem on a continuous manifold, which converts 3D geometry and task-dependent physical constraints into a unified planning geometry. By adding this optimal-control layer, OptiWorld generates videos with preferable dynamics, demonstrating strong potential in multiple tasks including goal-conditioned image-to-video generation, video dynamics editing, and counterfactual generation.

1 Introduction

World models predict future world states from current observations and control inputs [1, 2]. Recent video generation models offer a scalable pixel-space form of world modeling: given an initial frame and a condition (often specified by prompts or actions), they can synthesize realistic future frames [3–11]. This capability has enabled applications in robot learning [12, 13, 11, 14], generative gaming [9, 15, 16, 6, 17], and world dynamics simulation [4, 3, 18, 8, 10, 19].

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

However, as video generation models move toward video world models, they still lack physical consistency and controllability [20–27]. A video world model should make future motion physically valid and controllable, not just visually realistic. Recent work has added physical priors to video generation [28–32, 26, 33–35]. These methods make motion more physically plausible. However, they do not further optimize the motion under physical constraints. As shown in Figure 1, consider generating a video where a cup full of water is moved from one side of a laptop to the other. There are many possible paths from the start to the goal. Some paths may reach the goal but are unsafe, not smooth, or inefficient.

In this paper, we propose **OptiWorld**, a framework that brings classical **optimal control** [36] into video generation at inference time. We observe that recent 3D-aware video generation models [37–44] already have strong world rendering ability, but they rely on a predefined input trajectory to specify how the world should evolve. Therefore, the key missing piece for video world models is not flawless visuals, but proactive world understanding and optimal planning before generation [27]. OptiWorld fills this gap with a three-stage pipeline: understanding, planning, and generation.

The understanding stage extracts a compact, task-relevant world state from the initial frame and the goal in a zero-shot manner. It uses generic vision foundation models to obtain geometric [45] and semantic [46] information, and a vision-language model (VLM) [47] to reason about high-level object relations in the world. The planning stage is the core optimal-control component because it decides which motion the video model should render. This is challenging in open visual scenes: the planner must jointly model continuous 3D geometry and task-dependent physical constraints, including goal reaching, safety, smoothness, and efficiency. Traditional rule-based planners are often tied to fixed state spaces and hand-designed rules, while directly optimizing these heterogeneous constraints as separate penalties can be brittle and highly non-convex [48–53]. We therefore use a Riemannian manifold as a natural way to convert world information into continuous geometry, and formulate optimal planning as a geometric shortest-path problem. For example, unsafe regions become long or uphill, while goal- and efficiency-favored directions become short or downhill; the resulting geodesic-like trajectory can then be used as video guidance. Finally, the generation module renders the video based on the planned trajectory and the initial frame.

By explicitly incorporating optimal control, OptiWorld improves the motion in generated videos, making it safer, smoother, more efficient, and more physically consistent. We demonstrate its effectiveness across multiple tasks including goal-conditioned image-to-video generation, video dynamics editing, and counterfactual generation. Our core contributions are:

- We introduce optimal control into video generation, achieving safer, smoother, more efficient, and physically consistent motions via proactive world understanding and optimal planning prior to generation.
- We model heterogeneous physical constraints as a continuous geometric manifold, enabling optimal control for open-scene video motion planning.

2 Related Work

Video Generation Models. Recent progress in video generation has been largely driven by advances in diffusion models [54, 55] and the scaling [56] of Diffusion Transformer [57]. Building on these developments, modern video models generate frames from images, videos, text, or actions at scale [3, 4, 58, 59, 12, 7, 5]. These approaches achieve strong visual quality and generalization, making them promising for applications such as content generation, robot data synthesis, and interactive world simulation [3, 4, 18, 8, 10, 19, 9, 15, 16, 6, 17, 12, 13, 11, 14]. However, these models primarily learn appearance-level **motion** from data, without explicit modeling or control of underlying **dynamics**. As a result, generated videos may appear realistic but still violate physical principles and lack precise controllability, especially in out-of-distribution scenarios [20–24, 26, 33, 60, 61].

Motion-controlled Video Generation. A large body of work improves controllability by introducing additional motion-related conditions, such as camera poses, motion trajectories, optical flow, or other user-specified signals [62–70, 37–44]. While these methods enhance controllability, the control signals are typically derived from predefined inputs, and do not fundamentally address physical consistency of the generated motion.

Table 1: **World-modeling view of OptiWorld.** A concise comparison by observation model and dynamics model.

Component		Optimal Control	Video Generation (I2V)	OptiWorld
Observation model	Input	observation y_t	frame y_0	frame + goal (y_0, g)
	Modeling	observer $\mathcal{O}_{\text{oc}}$	encoder E_{VAE}	understanding $\mathcal{E}_\phi$
	Output	state x_t	token z_0	world state $\mathcal{W}$
Dynamics model	Input	state/control/goal (x_t, u_t, g)	token + condition (z_0, c)	world state + goal $(\mathcal{W}, g)$
	Modeling	dynamics function f	DiT $p_\theta(\cdot \mid z_0, c)$	planner $\mathcal{P}_\eta$
	Output	controls $u^\star_{0:T-1}$ / states $x^\star_{1:T}$	tokens $z_{1:T} \to \hat{y}_{1:T}$	optimal path $\tau^\star \to \hat{y}_{1:T}$

Physics-aware Video Generation. Recent works incorporate physical priors into video generation via three main paradigms: post-hoc simulation [71–75], simulation-guided generation [76–80], and learned priors or evaluators for improving physical plausibility [81, 28–32, 82, 33–35, 83]. The closest to our work is VLIPP [29], which uses GPT-4o [84] to generate coarse motion trajectories for guiding diffusion models. However, prior methods focus on enforcing plausibility rather than explicitly optimizing dynamics under multiple objectives such as safety, smoothness, and efficiency.

3 Preliminaries

A world model needs two basic components: an *observation model* that abstracts information from observations, and a *dynamics model* that describes how the world evolves over time. From this view, optimal control and video generation are two different forms of world modeling. Table 1 compares their observation models, dynamics models, and how OptiWorld combines them.

3.1 Optimal Control as Explicit World Modeling

Classical optimal control is a model-based paradigm that assumes an explicit world model of the dynamical system. Its observation model turns sensor observations into a latent world state, and the dynamics model predicts how the state evolves under controls [36]. Let y_t denote the observation, x_t the world state, and u_t the control. A standard formulation is

$$x_t = \mathcal{O}_{\text{oc}}(y_t), \qquad x_{t+1} = f(x_t, u_t), \qquad y_t = h(x_t), \tag{1}$$

where $\mathcal{O}_{\text{oc}}$ estimates a planning state from an observation, h maps states to observations, and f is the controlled dynamics. Given a goal g, optimal control chooses the controls by solving

$$u^\star_{0:T-1} = \arg\min_{u_{0:T-1}} \phi(x_T, g) + \sum_{t=0}^{T-1} \ell(x_t, u_t; g), \quad \text{s.t. } x_{t+1} = f(x_t, u_t),\ r_i(x_t, u_t) \le 0. \tag{2}$$

Here ϕ is a terminal cost (Mayer) term, ℓ is a running cost (Lagrange) term, and r_i represents constraints such as collision avoidance and safety. The strength of optimal control is explicit decision-making: it selects controls or states according to clear goals and constraints. Its limitation is that it assumes the planning state and dynamics are already available.

3.2 Video Generation as Implicit World Modeling

Video generative models can also be considered implicit world models. Their observation model maps pixels into visual tokens and back to pixels, while their dynamics model evolves those tokens in latent space. Given an initial frame y_0 and a generic condition c, a latent video generator can be written as

$$z_0 = E_{\text{VAE}}(y_0), \qquad z_{1:T} \sim p_\theta(z_{1:T} \mid z_0, c), \qquad \hat{y}_{1:T} = D_{\text{VAE}}(z_{1:T}). \tag{3}$$

Here E_{VAE} and D_{VAE} are the visual encoder and decoder, z_t denotes visual tokens, and c can be a text prompt, action, or goal condition. The latent dynamics p_θ is usually implemented by a Diffusion

Transformer (DiT) [57]. This resembles state-space modeling in optimal control: z_t can be interpreted as a latent state, while the DiT implicitly parameterizes transitions in this latent space.

The difference is that the dynamics are hidden inside the generator rather than exposed as a function with explicit costs and constraints. This gives video models strong scene generalization and high visual quality, but it also means they lack the decision layer of optimal control. At inference time, the model is usually asked to generate a plausible future, not to optimize which future should happen under safety, smoothness, energy, or goal constraints.

3.3 Integrating Optimal Control into Video World Generation

OptiWorld integrates these two views by introducing an explicit intermediate planning layer between optimal control and video generation. Its observation model converts pixels and goals into a planning-ready world state $\mathcal{W}$. Its dynamics model optimizes a trajectory τ before rendering the video. Let $J(\tau; \mathcal{W}, g)$ be the physical planning objective under world state $\mathcal{W}$ and goal g. The decomposition is

$$\mathcal{W} = \mathcal{E}_\phi(y_0, g), \qquad \tau^\star = \mathcal{P}_\eta(\mathcal{W}, g) = \arg\min_{\tau \in \mathcal{C}(\mathcal{W})} J(\tau; \mathcal{W}, g), \qquad \hat{y}_{1:T} = \mathcal{G}_\psi(y_0, \tau^\star, g), \quad (4)$$

Here $\mathcal{E}_\phi$ is the zero-shot understanding module, and $\mathcal{W}$ contains the start state x_0, goal state x_g, and physical constraints $\mathcal{C}$. $\mathcal{P}_\eta$ is the optimal planner, $\tau^\star$ is the optimized trajectory, and $\mathcal{G}_\psi$ renders the optimized trajectory into future frames $\hat{y}_{1:T}$. In this way, OptiWorld keeps the visual generation ability of video models while adding the explicit decision layer of optimal control.

4 Method

4.1 Problem Formulation

We study physically constrained video generation for rigid-object motion in 3D scenes. The paper focuses on two tasks. The first is *goal-conditioned image-to-video generation*: given an initial frame y_0 and a goal g specified by a point and language instruction, generate a video $\hat{y}_{1:T}$ where the selected object moves along a physically preferable trajectory. The second is *video dynamics editing*: given a source video, improve the object motion while preserving the scene identity and task intent. In both tasks, the generated motion should be safe, smooth, and efficient.

4.2 Overview

OptiWorld has three stages, shown in Figure 2. First, *understanding* converts the image and goal into a task-relevant 3D world state. Second, *planning* optimizes a 3D path in this world under goal-reaching, safety, smoothness, and efficiency constraints. Third, *generation* uses the optimized path as motion guidance for a controllable video model.

The motivation is practical: recent 3D-trajectory-conditioned generators already have strong rendering ability once a motion signal is given [37–44], but that motion is usually predefined by the input. OptiWorld therefore focuses on the missing steps before rendering: understanding the scene and planning a physically optimized 3D trajectory.

4.3 Understanding: Zero-Shot 3D World State Construction

Planning is difficult to perform directly on raw pixels when explicit geometric reasoning and constraints are required. It requires a structured scene representation that captures geometry for distance and motion reasoning, as well as semantic regions for obstacles, buffers, hazards, and targets. It should also encode task-level relations, since the same object can play different roles in different tasks. For example, a laptop is a high-risk region for a cup of water, but not necessarily for a plush toy. We therefore build a structured and compact world state:

$$\mathcal{W} = \{\mathcal{P}, \mathcal{S}, \mathcal{A}, \mathcal{T}\}. \quad (5)$$

Here $\mathcal{P}$ is the metric 3D point world, $\mathcal{S}$ contains semantic regions, $\mathcal{A}$ contains task attributes such as object role and semantic risk, and $\mathcal{T}$ contains optional 3D tracks for video editing.

We construct $\mathcal{W}$ with an agentic zero-shot pipeline. First, a metric geometry model estimates depth and 3D point maps from the input frame [45]. Second, a promptable segmentation model localizes

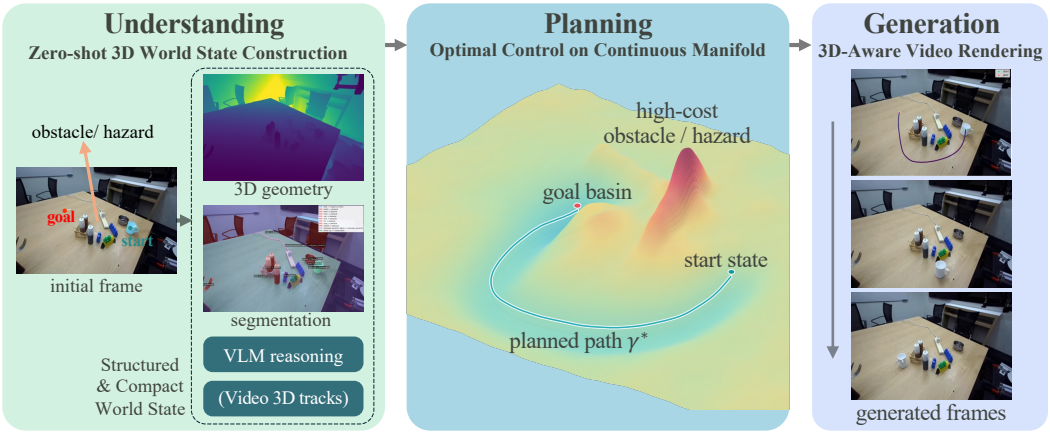


Figure 2: **OptiWorld pipeline.** Given an initial frame and goal, the understanding stage builds a compact 3D world state from geometry, segmentation, VLM reasoning, and 3D tracking (for video dynamics editing). The planning stage converts this world state into a continuous constraint-induced manifold, where hazards become high-cost regions, the goal becomes a basin, and the optimized path follows a smooth low-cost valley. The generation stage then renders detailed future frames from the first frame, prompt, and planned/optimized 3D trajectory.

task-relevant objects and regions [46]. Third, a VLM reasons about the goal and assigns task roles such as obstacle, buffer, hazard, and target [47]. These outputs are lifted into 3D to form object states, goal sets, signed distances, semantic maps, and task weights. For video dynamics editing, we additionally extract dense 3D point tracks [85]; these tracks become editable variables while anchoring the optimized motion to the source video. More details are provided in Appendix C.

This design keeps the understanding module general. Each component is built from strong vision foundation models with zero-shot scene understanding ability, so OptiWorld does not need to train a new world-understanding model for every object category, scene type, or task setting. The output is not a full physical simulator, but a planning-ready state that contains just enough geometry, semantics, and task logic for the optimal planner.

4.4 Planning: Optimal Control on a Continuous Manifold

Planning is the core of OptiWorld because it decides which motion the video model should render. A visually strong generator cannot fix a poor trajectory: if the guidance path cuts through a hazard, jitters, or takes an unnecessary detour, the generated video will inherit those dynamics.

One possible solution is to use a traditional rule-based planning pipeline: define hand-designed costs, run a graph search such as RRT [50], and refine the result with a model-predictive-control (MPC)-style numerical optimizer [52]. This works well in specific domains where the state space and rules are fixed, such as navigation or target tracking [48, 49]. However, video generation starts from open visual scenes. The planner must combine 3D geometry with task-dependent physical constraints, including goal reaching, safety, smoothness, and efficiency, all extracted or inferred from pixels. Encoding these signals as separate rules makes the modeling brittle and tied to particular scenarios, while directly optimizing them as independent penalties leads to a highly non-convex objective [53].

We therefore formulate planning as a geometric problem. A Riemannian manifold gives a natural way to convert world information into continuous geometry [53, 86, 87]. At each object state q, a positive-definite metric $G(q)$ defines the local squared travel cost

$$ds_G^2 = dq^\top G(q) dq. \tag{6}$$

When $G(q)$ is large near unsafe regions, a path through those regions becomes long even if it is short in Euclidean distance. Therefore, planning becomes a shortest-path problem in the geometry induced by the scene, not in a fixed Euclidean space. A scalar potential $V(q)$ further turns the scene into a landscape: goals become valleys, while risky or inefficient regions become hills. This is

better matched to video motion planning than flat Euclidean penalties because goal reaching, safety, efficiency, and smoothness are modeled within one scene geometry.

Formally, let $\gamma : [0, 1] \to \mathcal{M}$ be a continuous path on the configuration manifold $\mathcal{M}$, with $\gamma(0) = q_s$ and $\gamma(1) \in \mathcal{G}(g)$. Its Riemannian length is

$$L_G(\gamma) = \int_0^1 \sqrt{\dot{\gamma}(s)^\top G(\gamma(s))\dot{\gamma}(s)}\, ds. \tag{7}$$

Minimizing the squared length energy gives a geodesic-like path under this geometry. We add the potential and dynamics terms to obtain the planning objective:

$$\gamma^\star = \arg\min_\gamma \int_0^1 \left[\frac{1}{2}\dot{\gamma}(s)^\top G(\gamma(s))\dot{\gamma}(s) + V(\gamma(s))\right] ds + \lambda_{\text{dyn}} \int_0^1 \left\|\nabla_{\dot{\gamma}}\dot{\gamma}(s)\right\|^2 ds, \tag{8}$$

where G is the task-induced metric, V is the scalar potential, and $\nabla_{\dot{\gamma}}\dot{\gamma}$ is the covariant acceleration along the curve. Intuitively, the path is optimized as if it moves on a terrain induced by the current scene: it is pulled toward the goal, pushed away from unsafe regions, discouraged from wasting motion, and smoothed for stable video generation.

Constraint-induced geometry. All physical constraints are encoded into the same geometry. We use three planning fields: goal reaching, safety, and efficiency, plus a smoothness regularizer on the path. Let ϕ_{safe} denote the unified safety cost from obstacles, buffers, and semantic hazards; let ϕ_{eff} denote the efficiency cost from path length, detours, and positive lift; and let a denote the gravity-up direction:

$$G(q) = \big(1 + \lambda_{\text{safe}}\phi_{\text{safe}}(q) + \lambda_{\text{eff}}\phi_{\text{eff}}(q)\big)I + \lambda_{\text{up}}\phi_{\text{eff}}(q)aa^\top, \tag{9}$$

$$V(q) = \lambda_{\text{goal}} d(q, \mathcal{G}(g))^2 + \lambda_{\text{safe}}\phi_{\text{safe}}(q) + \lambda_{\text{eff}}\phi_{\text{eff}}(q). \tag{10}$$

All λ terms are nonnegative weights. The goal term creates a low-potential basin at the target. The safety term stretches or raises unsafe regions, including geometric obstacles and semantic hazards such as a laptop near a cup of water. The efficiency term favors short, low-energy motion and discourages unnecessary upward movement or detours. Smoothness is handled by the covariant acceleration term in Eq. (8), which penalizes high acceleration, curvature, and abrupt turns. Together, these terms reshape the scene into a continuous planning manifold. Appendix C.2 gives implementation details.

Practical solver. The solver follows the geometry above in three steps. First, we convert the world state $\mathcal{W}$ into 3D cost fields for goal reaching, safety, and efficiency. Then, we search on the voxel graph to get a globally feasible seed path. Finally, we refine this seed as a continuous curve under Eq. (8), and resample it into the frame-level trajectory $\tau^\star$ used by the generator:

$$\mathcal{W} \to \text{3D cost fields} \to \text{global seed path} \to \gamma^\star \to \tau^\star. \tag{11}$$

Here $\tau^\star$ is the final optimized 3D trajectory in Euclidean coordinates.

For video dynamics editing, OptiWorld improves source 3D tracks into smoother, more efficient, and more physically reasonable motion while preserving the scene identity and task instruction. The strategy is hybrid: we combine an autonomous plan from Eq. (11) with source 3D tracks to form a new global seed path, and then refine this seed. This keeps the benefits of the planned trajectory, such as efficiency and smoothness, while preserving the real motion cues already present in the original video.

4.5 Generation: 3D-Aware Video Rendering

Finally, the generation module takes the first frame, a text prompt, and planned frame-level tracking controls as input, and renders a photorealistic video. OptiWorld is generator-agnostic: any mature 3D-trajectory-controlled image-to-video model can use the optimized motion guidance [37–44]. In this paper, we use Diffusion as Shader (DaS) [39] because it provides a convenient 3D tracking interface. For goal-conditioned image-to-video generation, the tracking controls come from the planned trajectory $\tau^\star$ of the selected object. For video dynamics editing, the input first frame is the first frame of the source video, and the tracking controls come from the optimized tracks.

5 Experiments

5.1 Implementation Details

Benchmark. We evaluate the effectiveness of OptiWorld on two video generation tasks: goal-conditioned image-to-video (I2V) generation and video dynamics editing. OptiBench-I2V contains 60 distinct environments (22 from DROID [88] and 38 recorded by us), evaluated over more than 300 combinations of objects, initial configurations, and goals. The environments span offices, kitchens, bathrooms, living rooms, bedrooms, industrial settings, and outdoors. Each case specifies an initial image, a selected object, a goal point, and an instruction. OptiBench-V2V contains 30 source videos (20 from DROID and 10 recorded by us) of robot manipulation and human hand-object motion. To assess robustness beyond the original benchmarks, we additionally collected 24 challenging I2V scenes and 12 V2V scenes, including low-light, reflective or textureless, cluttered, garden, and outdoor settings; each added I2V scene includes at least five target configurations. Across the original and additional data, evaluation covers 84 I2V environments with more than 420 task configurations and 42 V2V videos. Tables 2 and 3 report the results.

Metrics. We evaluate video quality and motion quality. For video quality, we use motion smoothness, background consistency, and temporal flickering from VBench [89]. For motion quality, we extract foreground 3D tracks from the generated videos and evaluate four aspects: goal reaching, safety, smoothness, and efficiency. For goal reaching, Goal err. measures the final 3D distance to the target in I2V, and Track dev. measures the average 3D deviation from the source motion after first-frame alignment in V2V. For safety, Viol.@succ. measures the fraction of videos that still enter unsafe regions. For smoothness, Accel. and Jerk measure second- and third-order changes of the 3D tracks. For efficiency, Path len. measures the total 3D travel distance, and Energy measures the accumulated motion cost along the track. Details are provided in the Appendix D.

Inference details. Our method follows a fully inference-time pipeline. In the understanding stage, we estimate 3D geometry using MoGe2 [45] to obtain metric depth and point maps. For prompt-based segmentation, we adopt SAM3 [46], which provides fine-grained masks for all scene pixels and extends the vocabulary to open-world settings. For VLM reasoning, we adopt Qwen2.5-VL-3B-Instruct [47]. For video dynamics editing requiring dense 3D tracks, we use SpatialTrackerV2 [85]. The planner is the core component, solving optimal control on a geometric manifold. The understanding and planning stages take approximately 2 minutes per scene on a single Nvidia A100-80GB GPU. For generation, we build upon the object manipulation pipeline of Diffusion as Shader (DaS) [39], enhance its tracking video rendering for real-world 3D motion, and use 30 denoising steps. A separate 30-scene comparison of VLM choices is summarized in Section 8.

5.2 Goal-Conditioned Image-to-Video Generation

Given a scene image, a goal point, and an instruction, the goal-conditioned I2V task is to move the selected object to the target while respecting physical constraints. We compare OptiWorld with state-of-the-art open-source I2V models HunyuanVideo-1.5 [90] and Wan2.2 [7], the world foundation model Cosmos-Predict2.5 (Image2World) [91], and the physics-aware video generator VLIPP [29].

In Figure 3, HunyuanVideo-1.5, Wan2.2, and Cosmos-Predict2.5 can generate plausible frames, but they lack explicit scene-aware planning and often fail to reach the goal accurately or safely. VLIPP introduces physics-aware guidance, but its coarse motion prior can still produce hazardous artifacts around outlet and chair. OptiWorld first plans a 3D path under these constraints and then uses it as tracking guidance, leading to more direct goal reaching, fewer safety violations, and smoother trajectories. Table 2 reports the baseline results and OptiWorld performance on the larger benchmark. On this expanded evaluation, OptiWorld retains low goal error and achieves the lowest safety-violation rate while maintaining strong smoothness, efficiency, and video-quality scores.

5.3 Video Dynamics Editing

In video dynamics editing, we compare OptiWorld with the source video and Wan2.1 First-Last-Frame-to-Video (FLF2V) [7], which is a natural baseline for preserving the start and end states.

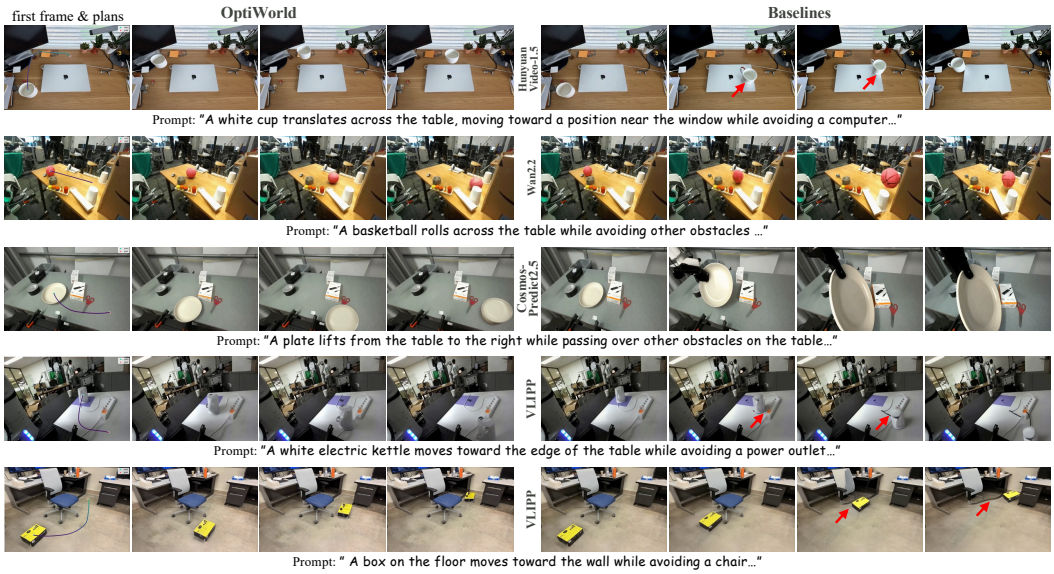


Figure 3: **Visual comparisons on goal-conditioned I2V.** OptiWorld generates videos with better goal reaching, safety, smoothness, and efficiency.

Table 2: **Quantitative comparisons on goal-conditioned I2V.** Best results are highlighted in red, and second-best results in blue.

Method	Motion quality					Video quality (VBench)		
	Goal err. ↓	Viol.@succ. ↓	Accel. ↓	Jerk ↓	Energy ↓	Mot. smooth ↑	BG cons. ↑	Flicker ↑
HunyuanVideo-1.5 [90]	0.571	0.539	0.0126	0.0342	0.0603	0.995	0.957	0.993
Wan2.2 [7]	0.567	0.535	0.0180	0.0320	0.0988	0.993	0.955	0.990
Cosmos-Predict2.5 [91]	0.759	0.301	0.0164	0.0194	0.0635	0.992	0.934	0.984
VLIPP [29]	0.365	0.547	0.0133	0.0185	0.0492	0.990	0.961	0.991
OptiWorld	0.335	0.240	0.0079	0.0129	0.0253	0.997	0.979	0.996

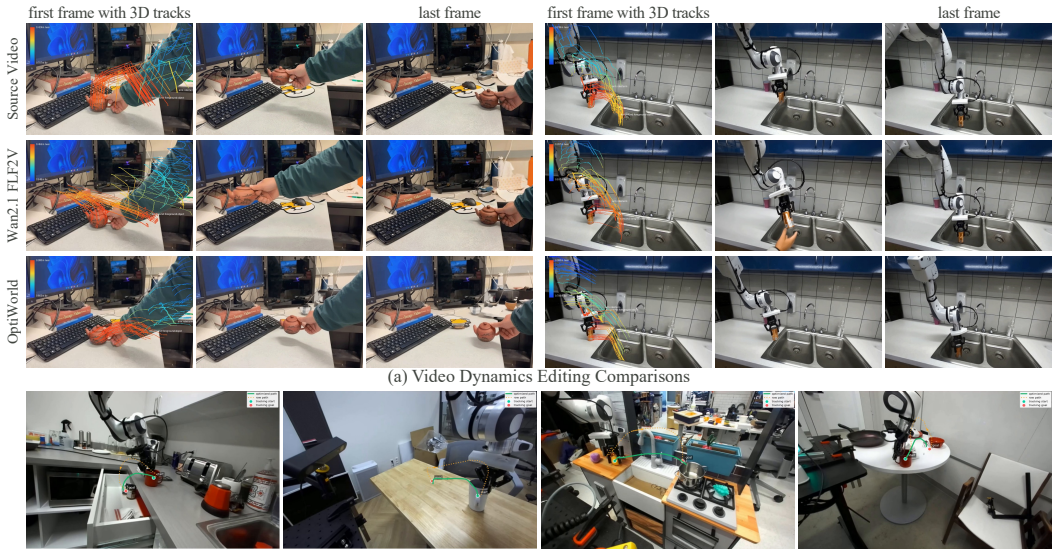
In Figure 4, Wan2.1 FLF2V preserves the coarse start and end frames, but the intermediate motion is not optimized and can remain long, jerky, or physically inefficient. OptiWorld instead extracts and refines the 3D tracks before generation, improving the object dynamics while preserving the scene and task intent. Table 3 reports the baseline results and OptiWorld performance on the larger benchmark, where the optimized tracks remain smoother and more efficient while preserving video quality.

5.4 Counterfactual Physics Generation

As shown in Figure 1(b), OptiWorld can also generate counterfactual motions by changing the planner before rendering. We use two settings. **(1) Change the goal.** For the same scene, a different goal produces a different trajectory while the understanding and generation modules stay fixed. **(2) Change the constraints.** For the same start and goal, setting the safety weight to zero can produce a dangerous path through a risky region. These counterfactual videos provide controlled labels such as goal, safety violation, energy, and smoothness for future robot-policy training or evaluation. More examples appear in Appendix H.

5.5 Ablation Study

We ablate OptiWorld by pipeline stage. In understanding, w/o VLM removes active VLM reasoning and keeps only geometry and segmentation, so task-dependent object roles are no longer inferred. In planning, w/o Man. replaces the constraint-induced manifold with a flatter Euclidean planning objective; w/o Safe, w/o Eff., and w/o Smooth respectively remove the safety, efficiency, and



(a) Video Dynamics Editing Comparisons

(b) Trajectory Comparison Between Before (dashed yellow) and After (solid green) Optimization. Please Zoom In.

Figure 4: **Visual comparisons on video dynamics editing.** OptiWorld improves the source trajectory, producing smoother and more efficient motion.

Table 3: **Quantitative comparisons on video dynamics editing.**

Method	Motion quality					Video quality (VBench)		
	Track dev. ↓	Accel. ↓	Jerk ↓	Path len. ↓	Energy ↓	Mot. smooth ↑	BG cons. ↑	Flicker ↑
Source video	–	0.0129	0.0258	0.645	0.0301	0.992	0.966	0.987
Wan2.1 FLF2V [7]	0.135	0.0184	0.0296	0.786	0.0502	0.992	0.954	0.993
OptiWorld	0.120	0.0129	0.0209	0.533	0.0253	0.995	0.962	0.991

smoothness terms. For generation, 2D guidance replaces DaS with projected optical-flow guidance and Go-with-the-Flow. Table 4 shows that removing VLM reasoning or smoothness worsens motion regularity; removing the manifold or safety increases Viol.@succ.; removing efficiency increases Energy; and projected 2D guidance degrades video metrics.

6 Conclusion

We introduced optimal control into video generation. Through active scene understanding and manifold-based planning, OptiWorld can generate video worlds with better dynamics: motions are more controllable, smoother, safer, and more efficient. More broadly, this suggests that video world models should not only imagine plausible futures, but also choose futures that respect the structure and consequences of the physical world.

7 Limitations

OptiWorld improves motion planning before video generation, but final video quality remains limited by the renderer. DaS can introduce artifacts, object deformation, or imperfect texture preservation, especially for large motions or thin, reflective, or heavily occluded objects. The system is designed as an offline inference pipeline; faster geometry estimation, semantic reasoning, planning, and rendering are needed for real-time use.

Table 4: **Quantitative comparisons on ablation study.** Planning diagnostics are computed on optimized trajectories before rendering, while video diagnostics are computed on generated videos. Dashes indicate metrics that are not the primary diagnostic for that ablation.

Stage	Variant	Planning diagnostics				Video diagnostics		
		Viol.@succ. ↓	Accel. ↓ (10^{-4})	Jerk ↓ (10^{-4})	Energy ↓ (10^{-4})	Goal err. ↓	Viol.@succ. ↓	Energy ↓
Understanding	w/o VLM	–	3.10	1.25	0.35	–	–	–
Planning	w/o Man.	0.349	–	–	–	–	–	–
	w/o Safe	0.589	–	–	–	–	–	–
	w/o Eff.	–	–	–	0.27	–	–	–
	w/o Smooth	–	3.20	1.26	0.37	–	–	–
Generation	2D guidance	–	–	–	–	0.361	0.523	0.0518
OptiWorld		0.304	2.40	0.86	0.22	0.310	0.260	0.0126

8 Questions and Answers

Why use a Riemannian formulation? A conventional Euclidean planner combines distance, collision, safety, height, energy, and smoothness through weighted penalties. These terms have different units and gradient scales, so weights are difficult to tune across scenes, and improving one objective may degrade another. Our metric encodes where and in which directions motion is costly in the local geometry. The "w/o Man." ablation replaces the constraint-induced manifold with a flatter Euclidean objective while retaining the same scene understanding and tuned weights; it shows a benefit in our evaluated settings, especially for collision avoidance. A broader comparison with carefully tuned constrained planners remains useful future work.

Can the method handle articulated or deformable objects? The current representation and experiments focus on rigid-object motion. Articulated tasks such as opening a drawer could be supported by adding articulation variables and constraints. Cloth folding and liquid pouring require deformable-state perception and dynamics that are beyond the current method.

How robust is planning to depth errors? Depth errors can distort the inferred 3D geometry and planning cost fields. Our additional scenes include low-light, reflective or textureless surfaces, clutter, gardens, and outdoor settings, but we have not conducted a dedicated uncertainty-aware robustness study. Propagating depth uncertainty into planning remains an important direction.

How does voxel resolution affect safety and smoothness? Voxelization is used only for the initial A* search that produces a feasible seed. The final trajectory is refined continuously on the constraint-induced manifold, so voxel resolution mainly affects initialization quality and computation. Continuous refinement does not compensate for inaccurate or overly coarse scene geometry; safety still depends on the quality of the estimated obstacles and cost fields.

Why use DaS, and could a stronger renderer help? We use DaS because it provides a practical 3D-trajectory interface for testing the planning contribution. OptiWorld is renderer-agnostic in principle, and a stronger trajectory-conditioned renderer could improve visual fidelity. Adapting Wan2.2 with a learned 3D-conditioning interface would require training a new renderer and is outside this inference-time planning study. Our 2D-guidance ablation shows that alternative trajectory conditioning is possible, although it reduces trajectory-following accuracy and video quality in our implementation.

What downstream utility is established? The intended applications include robotics simulation, controllable content creation, and action-data generation. We evaluate trajectory and video metrics, but do not measure downstream robot-policy performance; that utility remains to be established.

Acknowledgments

This work is supported, in part, by the United States National Science Foundation under the grants 2133032, 2431505, and a research award from Samsung Research America. We thank Yichen Sheng for helpful discussion.

References

[1] David Ha and Jürgen Schmidhuber, "World models," *arXiv preprint arXiv:1803.10122*, 2018.

[2] Xun Huang, "Towards video world models," 2025.

[3] OpanAI, "Video generation models as world simulators," `https://openai.com/index/video-generation-models-as-world-simulators/`.

[4] Google, "Veo 3.1: Our state-of-the-art video generation model," `https://aistudio.google.com/models/veo-3/`, 2025.

[5] Team Seedance, "Seedance 2.0: Advancing video generation for world complexity," *arXiv preprint arXiv:2604.14148*, 2026.

[6] Team HunyuanWorld, "Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels," *arXiv preprint arXiv:2507.21809*, 2025.

[7] Team Wan, "Wan: Open and advanced large-scale video generative models," *arXiv preprint arXiv:2503.20314*, 2025.

[8] Philip J. Ball, Jakob Bauer, Frank Belletti, Bethanie Brownfield, Ariel Ephrat, Shlomi Fruchter, Agrim Gupta, Kristian Holsheimer, Aleksander Holynski, Jiri Hron, Christos Kaplanis, Marjorie Limont, Matt McGill, Yanko Oliveira, Jack Parker-Holder, Frank Perbet, Guy Scully, Jeremy Shar, Stephen Spencer, Omer Tov, Ruben Villegas, Emma Wang, Jessica Yung, Cip Baetu, Jordi Berbel, David Bridson, Jake Bruce, Gavin Buttimore, Sarah Chakera, Bilva Chandra, Paul Collins, Alex Cullum, Bogdan Damoc, Vibha Dasagi, Maxime Gazeau, Charles Gbadamosi, Woohyun Han, Ed Hirst, Ashyana Kachra, Lucie Kerley, Kristian Kjems, Eva Knoepfel, Vika Koriakin, Jessica Lo, Cong Lu, Zeb Mehring, Alex Moufarek, Henna Nandwani, Valeria Oliveira, Fabio Pardo, Jane Park, Andrew Pierson, Ben Poole, Helen Ran, Tim Salimans, Manuel Sanchez, Igor Saprykin, Amy Shen, Sailesh Sidhwani, Duncan Smith, Joe Stanton, Hamish Tomlinson, Dimple Vijaykumar, Luyu Wang, Piers Wingfield, Nat Wong, Keyang Xu, Christopher Yew, Nick Young, Vadim Zubov, Douglas Eck, Dumitru Erhan, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Raia Hadsell, Aäron van den Oord, Inbar Mosseri, Adrian Bolton, Satinder Singh, and Tim Rocktäschel, "Genie 3: A new frontier for world models," 2025.

[9] Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al., "Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory," *arXiv preprint arXiv:2604.08995*, 2026.

[10] Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, Zian Wang, Nicholas Sharp, Zan Gojcic, Sanja Fidler, Jiahui Huang, Huan Ling, Jun Gao, and Xuanchi Ren, "Lyra 2.0: Explorable generative 3d worlds," *arXiv preprint arXiv:2604.13036*, 2026.

[11] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi Fan, and Joel Jang, "World action models are zero-shot policies," *arXiv preprint arXiv:2602.15922*, 2026.

[12] NVIDIA, "World simulation with video foundation models for physical ai," *arXiv preprint arXiv:2511.00062*, 2025.

[13] Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, and Jinwei Gu, "Cosmos policy: Fine-tuning video models for visuomotor control and planning," in *International Conference on Learning Representations (ICLR)*, 2026.

[14] Gemini Robotics Team, "Evaluating gemini robotics policies in a veo world simulator," *arXiv preprint arXiv:2512.10675*, 2025.

[15] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter, "Diffusion models are real-time game engines," in *International Conference on Learning Representations*, 2025.

[16] Junshu Tang, Jiacheng Liu, Jiaqi Li, Longhuang Wu, Haoyu Yang, Penghao Zhao, Siruis Gong, Xiang Yuan, Shuai Shao, and Qinglin Lu, "Hunyuan-gamecraft-2: Instruction-following interactive game world model," *arXiv preprint arXiv:2511.23429*, 2025.

[17] Georgy Savva, Oscar Michel, Daohan Lu, Suppakit Waiwitlikhit, Timothy Meehan, Dhairya Mishra, Srivats Poddar, Jack Lu, and Saining Xie, "Solaris: Building a multiplayer video world model in minecraft," *arXiv preprint arXiv:2602.22208*, 2026.

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

[19] Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos, "Video models are zero-shot learners and reasoners," *arXiv preprint arXiv:2509.20328*, 2025.

[20] 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," in *International Conference on Machine Learning*, 2025.

[21] Daochang Liu, Junyu Zhang, Anh-Dung Dinh, Eunbyung Park, Shichao Zhang, and Chang Xu, "Generative physical ai in vision: A survey," 2025.

[22] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini Jaini, and Robert Geirhos, "Do generative video models understand physical principles?," 2025.

[23] Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang, "VideoPhy-2: A challenging action-centric physical commonsense evaluation in video generation," *arXiv preprint arXiv:2503.06800*, 2025.

[24] Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Cheng Yu, Dianqi Li, Yu Qiao, and Ping Luo, "Towards world simulator: Crafting physical commonsense-based benchmark for video generation," in *International Conference on Machine Learning*, 2025.

[25] Chenyu Zhang, Daniil Cherniavskii, Andrii Zadaianchuk, Antonios Tragoudaras, Antonios Vozikis, Thijmen Nijdam, Derck W. E. Prinzhorn, Mark Bodracska, Nicu Sebe, and Efstratios Gavves, "Morpheus: Benchmarking physical reasoning of video generative models with real physical experiments," *arXiv preprint arXiv:2504.02918*, 2025.

[26] 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," in *International Conference on Machine Learning*, 2025.

[27] Jiahan 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," *arXiv preprint arXiv:2510.18135*, 2025.

[28] Jiaxi Lv, Yi Huang Huang, Mingfu Yan, Jiancheng Huang, Jianzhuang Liu, Yifan Liu Liu, Yafei Wen, Xiaoxin Chen, and Shifeng Chen, "Gpt4motion: Scripting physical motions in text-to-video generation via blender-oriented gpt planning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[29] Xindi Yang, Baolu Li, Yiming Zhang, Zhenfei Yin, Lei Bai, Liqian Ma, Zhiyong Wang, Jianfei Cai, Tien-Tsin Wong, Huchuan Lu, and Xu Jia, "Vlipp: Towards physically plausible video generation with vision and language informed physical prior," *arXiv preprint arXiv:2503.23368*, 2025.

[30] Karran Pandey, Matheus Gadelha, Yannick Hold-Geoffroy, Karan Singh, Niloy J. Mitra, and Paul Guerrero, "Motion modes: What could happen next?," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[31] Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao, "PhyT2V: Llm-guided iterative self-refinement for physics-grounded text-to-video generation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[32] Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Zhanjie Zhang, Jiasong Feng, Shanyuan Liu, Yuhang Ma, Bo Cheng, Dawei Leng, Yuhui Yin, and Xiaodan Liang, "WISA: World simulator assistant for physics-aware text-to-video generation," *arXiv preprint arXiv:2503.08153*, 2025.

[33] Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin, "Videojam: Joint appearance-motion representations for enhanced motion generation in video models," *arXiv preprint arXiv:2502.02492*, 2025.

[34] Yu Yuan, Xijun Wang, Tharindu Wickremasinghe, Zeeshan Nadir, Bole Ma, and Stanley H. Chan, "NewtonGen: Physics-consistent and controllable text-to-video generation via neural newtonian dynamics," *arXiv preprint arXiv: 2509.21309*, 2025.

[35] Yu Yuan, Tharindu Wickremasinghe, Zeeshan Nadir, Xijun Wang, Yiheng Chi, and Stanley H. Chan, "SeeU: Seeing the unseen world via 4d dynamics-aware generation," *arXiv preprint arXiv: 2512.03350*, 2025.

[36] Robert F. Stengel, *Optimal Control and Estimation*, Dover Publications, New York, USA, 1994.

[37] Hanlin Wang, Hao Ouyang, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Qifeng Chen, Yujun Shen, and Limin Wang, "Levitor: 3d trajectory oriented image-to-video synthesis," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[38] Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, Michael Ryoo, Paul Debevec, and Ning Yu, "Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[39] Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, Wenping Wang, and Yuan Liu, "Diffusion as shader: 3d-aware video diffusion for versatile video generation control," *arXiv preprint arXiv:2501.03847*, 2025.

[40] Angtian Wang, Haibin Huang, Zhiyuan Fang, Yiding Yang, and Chongyang Ma, "ATI: Any trajectory instruction for controllable video generation," *arXiv preprint arXiv:2505.22944*, 2025.

[41] Yingjie Chen, Yifang Men, Yuan Yao, Miaomiao Cui, and Liefeng Bo, "Perception-as-control: Fine-grained controllable image animation with 3d-aware motion representation," *arXiv preprint arXiv:2501.05020*, 2025.

[42] Jinbo Xing, Long Mai, Cusuh Ham, Jiahui Huang, Aniruddha Mahapatra, Chi-Wing Fu, Tien-Tsin Wong, and Feng Liu, "Motioncanvas: Cinematic shot design with controllable image-to-video generation," *arXiv preprint arXiv:2502.04299*, 2025.

[43] Yao-Chih Lee, Zhoutong Zhang, Jiahui Huang, Jui-Hsien Wang, Joon-Young Lee, Jia-Bin Huang, Eli Shechtman, and Zhengqi Li, "Generative video motion editing with 3d point tracks," *arXiv preprint arXiv:2512.02015*, 2025.

[44] Ryan Burgert, Charles Herrmann, Forrester Cole, Michael S Ryoo, Neal Wadhwa, Andrey Voynov, and Nataniel Ruiz, "Motionv2v: Editing motion in a video," *arXiv preprint arXiv:2511.20640*, 2025.

[45] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang, "Moge-2: Accurate monocular geometry with metric scale and sharp details," *arXiv preprint arXiv:2507.02546*, 2025.

[46] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shubham Debnath, Ronghang Hu, Dídac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Anqi Huang, Jiawen Lei, Tengyu Ma, Bowen Guo, Aditya Kalla, Matthew Marks, John Greer, Muxin Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Eleni Mavroudi, Kuan Xu, Tsung-Han Wu, Yitian Zhou, Lily Momeni, Rishav Hazra, Sifei Ding, Sagar Vaze, Florian Porcher, Fan Li, Songtao Li, Aishwarya Kamath, H. K. Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer, "Sam 3: Segment anything with concepts," *arXiv preprint arXiv:2511.16719*, 2025.

[47] Qwen Team, "Qwen2.5-vl," 2025.

[48] Yuze Wu, Mo Zhu, Xingxing Li, Yuheng Du, Yuxin Fan, Wenjun Li, Zhichao Han, Xin Zhou, and Fei Gao, "VLA-AN: An efficient and onboard vision-language-action framework for aerial navigation in complex environments," *arXiv preprint arXiv:2512.15258*, 2025.

[49] Zhengtong Xu, Zichen Miao, Qiang Qiu, Zhe Zhang, and Yu She, "Diffog: Differentiable policy trajectory optimization with generalizability," *IEEE Transactions on Robotics*, 2026.

[50] Sertac Karaman and Emilio Frazzoli, "Sampling-based algorithms for optimal motion planning," *The International Journal of Robotics Research (IJRR)*, 2011.

[51] Peter E. Hart, Nils J. Nilsson, and Bertram Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE Transactions on Systems Science and Cybernetics*, 1968.

[52] J. Rawlings, D.Q. Mayne, and Moritz Diehl, *Model Predictive Control: Theory, Computation, and Design*, 2017.

[53] Francesco Bullo and Andrew D Lewis, *Geometric control of mechanical systems: modeling, analysis, and design for simple mechanical control systems*, 2005.

[54] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, 2020, p. 6840–6851.

[55] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2021.

[56] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv: 2001.08361*, 2020.

[57] William Peebles and Saining Xie, "Scalable diffusion models with transformers," in *International Conference on Computer Vision*, 2023.

[58] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al., "Cogvideox: Text-to-video diffusion models with an expert transformer," in *International Conference on Learning Representations*, 2025.

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

[60] Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu, "WorldScore: A unified evaluation benchmark for world generation," *arXiv preprint arXiv:2504.00983*, 2025.

[61] Jianhao Yuan, Fabio Pizzati, Francesco Pinto, Lars Kunze, Ivan Laptev, Paul Newman, Philip Torr, and Daniele De Martini, "Likephys: Evaluating intuitive physics understanding in video diffusion models via likelihood preference," in *International Conference on Learning Representations*, 2026.

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

[63] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat, "CamCo: Camera-controllable 3d-consistent image-to-video generation," *arXiv preprint arXiv:2406.02509*, 2024.

[64] Chen Hou, Guoqiang Wei, Yan Zeng, and Zhibo Chen, "Training-free camera control for video generation," *arXiv preprint arXiv:2406.10126*, 2024.

[65] Dejia Xu, Yifan Jiang, Chen Huang, Liangchen Song, Thorsten Gernoth, Liangliang Cao, Zhangyang Wang, and Hao Tang, "Cavia: Camera-controllable multi-view video diffusion with view-integrated attention," *arXiv preprint arXiv:2410.10774*, 2024.

[66] 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*, 2024, pp. 1–11.

[67] Robin Courant, Nicolas Dufour, Xi Wang, Marc Christie, and Vicky Kalogeiton, "E.T. the exceptional trajectories: Text-to-camera-trajectory generation with character awareness," *arXiv preprint arXiv:2407.01516*, 2024.

[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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[69] Zhenghao Zhang, Junchao Liao, Xiangyu Meng, Long Qin, and Weizhi Wang, "Tora2: Motion and appearance customized diffusion transformer for multi-entity video generation," in *ACM International Conference on Multimedia*, 2025.

[70] 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," *arXiv preprint arXiv:2412.02700*, 2024.

[71] 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.

[72] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang, "Physgaussian: Physics-integrated 3d gaussians for generative dynamics," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[73] Xiyang Tan, Ying Jiang, Xuan Li, Zeshun Zong, Tianyi Xie, Yin Yang, and Chenfanfu Jiang, "Physmotion: Physics-grounded dynamics from a single image," *arXiv preprint arXiv:2411.17189*, 2024.

[74] 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*, 2024.

[75] Hao-Yu Hsu, Zhi-Hao Lin, Albert Zhai, Hongchi Xia, and Shenlong Wang, "Autovfx: Physically realistic video editing from natural language instructions," *arXiv preprint arXiv:2411.02394*, 2024.

[76] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz, "Physdiff: Physics-guided human motion diffusion model," in *International Conference on Computer Vision*, 2023.

[77] Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang, "Physgen: Rigid-body physics-grounded image-to-video generation," in *European Conference on Computer Vision*, 2024.

[78] Luca Savant Aira, Antonio Montanaro, Emanuele Aiello, Diego Valsesia, and Enrico Magli, "Motioncraft: Physics-based zero-shot video generation," in *Advances in Neural Information Processing Systems*, 2024.

[79] Boyuan Chen, Hanxiao Jiang, Shaowei Liu, Saurabh Gupta, Yunzhu Li, Hao Zhao, and Shenlong Wang, "Physgen3d: Crafting a miniature interactive world from a single image," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[80] Tianyi Xie, Yiwei Zhao, Ying Jiang, and Chenfanfu Jiang, "Physanimator: Physics-guided generative cartoon animation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[81] Zhengqi Li, Richard Tucker, Noah Snavely, and Aleksander Holynski, "Generative image dynamics," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[82] Yu Yuan, Xijun Wang, Yichen Sheng, Prateek Chennuri, Xingguang Zhang, and Stanley Chan, "Generative photography: Scene-consistent camera control for realistic text-to-image synthesis," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[83] Jianhao Yuan, Xiaofeng Zhang, Felix Friedrich, Nicolas Beltran-Velez, Melissa Hall, Reyhane Askari-Hemmat, Xiaochuang Han, Nicolas Ballas, Michal Drozdzal, and Adriana Romero-Soriano, "Inference-time physics alignment of video generative models with latent world models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

[84] Openai, "GPT-4o," `https://openai.com/index/hello-gpt-4o/`.

[85] Yuxi Xiao, Jianyuan Wang, Nan Xue, Nikita Karaev, Iurii Makarov, Bingyi Kang, Xin Zhu, Hujun Bao, Yujun Shen, and Xiaowei Zhou, "Spatialtrackerv2: 3d point tracking made easy," in *International Conference on Computer Vision*, 2025.

[86] Nathan D. Ratliff, Jan Issac, Daniel Kappler, Stan Birchfield, and Dieter Fox, "Riemannian motion policies," *arXiv preprint arXiv:1801.02854*, 2018.

[87] Ching-An Cheng, Mustafa Mukadam, Jan Issac, Stan Birchfield, Dieter Fox, Byron Boots, and Nathan Ratliff, "Rmpflow: A geometric framework for generation of multitask motion policies," *IEEE Transactions on Automation Science and Engineering*, 2021.

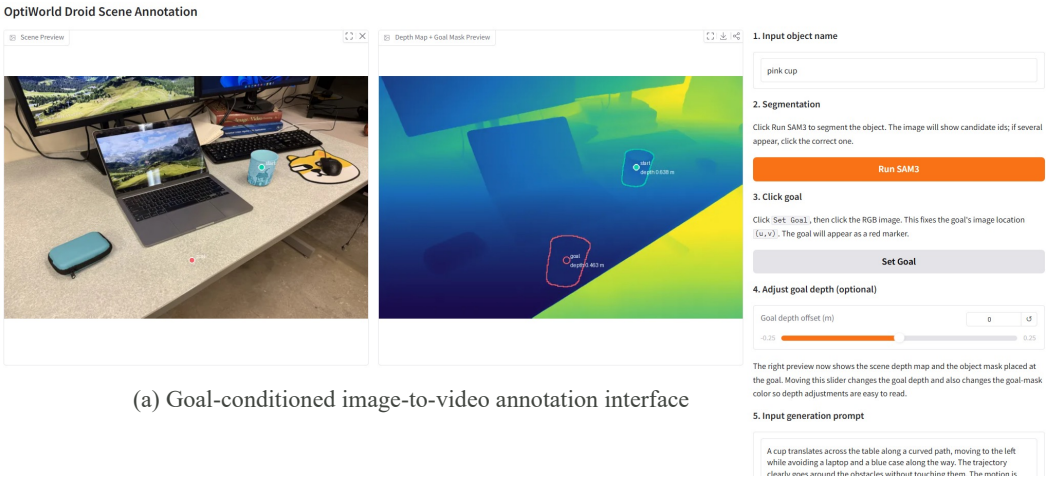
[88] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, Vitor Guizilini, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Muhammad Zubair Irshad, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy

Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn, "Droid: A large-scale in-the-wild robot manipulation dataset," *arXiv preprint arXiv:2403.12945*, 2024.

[89] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu, "VBench: Comprehensive benchmark suite for video generative models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[90] Tencent Hunyuan Foundation Model Team, "Hunyuanvideo 1.5 technical report," *arXiv preprint arXiv:2511.18870*, 2025.

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



(a) Goal-conditioned image-to-video annotation interface

Figure 5: **Data annotation interface.**

A Benchmark Details

Table 5 summarizes the original benchmark data; the additional challenging sets contain 24 I2V scenes and 12 V2V scenes and are evaluated separately. DROID scenes provide real robot manipulation layouts. Our collected data add daily indoor scenes and hand-object motion examples that contain semantic hazards such as laptops, outlets, chairs, fragile objects, and cluttered tables.

Table 5: **Benchmark data sources.**

Task	DROID	Collected by us	Total
Goal-conditioned image-to-video generation	22	38	60
Video dynamics editing	20	10	30

Annotation. For goal-conditioned image-to-video generation, each example is annotated with a selected movable object, a goal point, and a language instruction. The selected-object mask is extracted by SAM3 [46]. The goal point is given by the user's click.

B Additional VLM Comparison

As a supplementary component analysis, we compared the original Qwen2.5-VL-3B-Instruct with Qwen3.5-4B on 30 I2V scenes. The main benchmark results above use Qwen2.5-VL-3B-Instruct; the comparison below shows that Qwen3.5-4B yields more semantic regions per scene and improves planning success and trajectory smoothness in this test.

Table 6: **VLM comparison on 30 I2V scenes.**

Model	Planning success ↑	Semantic regions/scene ↑	Acceleration ↓ (10^{-4})	Jerk ↓ (10^{-4})
Qwen2.5-VL-3B-Instruct	90.0%	3.40	2.526	0.823
Qwen3.5-4B	93.3%	5.77	2.255	0.750

C OptiWorld Pipeline Details

C.1 Understanding

The understanding stage turns an image or a video into a compact 3D semantic world state.

Understanding for goal-conditioned image-to-video generation. For each input image, we first resize it to a fixed working resolution. MoGe2 [45] estimates metric depth, a dense 3D point map, and camera intrinsics. The annotation gives the selected object and the goal point. SAM3 [46] refines the selected object mask and the task-relevant scene masks. Qwen2.5-VL [47] reads the image and instruction, then assigns semantic roles such as manipulable object, support surface, obstacle, safety buffer, safety hazard, and target. We lift every valid mask pixel to 3D using the point map. The output contains the object centroid, goal position, semantic point cloud, signed-distance fields, hazard fields, and task weights.

Video dynamics editing. Video dynamics editing uses the same first-frame geometry and semantic reasoning, but adds source motion. We run SpatialTrackerV2 [85] on the source video to obtain dense 2D and 3D tracks, depth, point maps, intrinsics, and camera information. The foreground and arm masks select the tracks that can be edited. Background tracks are kept as anchors. The output world state therefore contains both the static scene constraints and the raw source 3D motion that should be preserved in intent but improved in dynamics.

C.2 Planning

The planner optimizes object motion in the 3D constraint-induced manifold defined in the main text. Let $q_k \in \mathbb{R}^3$ be the object state at planning step k, $\mathcal{G}(g)$ be the 3D goal set, $d_{\mathrm{obs}}(q)$ be the signed distance to geometric obstacles, $d_{\mathrm{haz}}(q)$ be the distance to semantic hazards, and a be the gravity-up direction.

Constraint fields. We use the same goal, safety, and efficiency fields as Eq. (9) and Eq. (10). The safety field combines geometric obstacles and semantic hazards:

$$\phi_{\mathrm{safe}}(q) = \phi_{\mathrm{obs}}(q) + \phi_{\mathrm{haz}}(q), \tag{12}$$

$$\phi_{\mathrm{obs}}(q) = \exp\left(-\frac{d_{\mathrm{obs}}(q) - m_{\mathrm{obs}}}{\sigma_{\mathrm{obs}}}\right), \tag{13}$$

$$\phi_{\mathrm{haz}}(q) = r(q)\exp\left(-\frac{d_{\mathrm{haz}}(q) - m_{\mathrm{haz}}}{\sigma_{\mathrm{haz}}}\right). \tag{14}$$

Here m_{obs} and m_{haz} are safety margins, σ_{obs} and σ_{haz} control the spatial falloff, and $r(q)$ is the VLM-estimated semantic risk weight. The efficiency field is

$$\phi_{\mathrm{eff}}(q) = \max\big((q - q_s)^\top a, 0\big)^2, \tag{15}$$

which discourages unnecessary positive lift from the start state q_s. Path length and detours are handled by the metric length term during graph search and continuous refinement.

These fields define the local metric

$$G(q) = \big(1 + \lambda_{\mathrm{safe}}\phi_{\mathrm{safe}}(q) + \lambda_{\mathrm{eff}}\phi_{\mathrm{eff}}(q)\big)I + \lambda_{\mathrm{up}}\phi_{\mathrm{eff}}(q)aa^\top, \tag{16}$$

and the scalar potential

$$V(q) = \lambda_{\mathrm{goal}}d(q, \mathcal{G}(g))^2 + \lambda_{\mathrm{safe}}\phi_{\mathrm{safe}}(q) + \lambda_{\mathrm{eff}}\phi_{\mathrm{eff}}(q). \tag{17}$$

Large safety or efficiency costs stretch the metric and raise the potential. A path through a hazard or obstacle therefore becomes expensive even if it is short in Euclidean distance, while the goal term keeps the endpoint close to $\mathcal{G}(g)$.

Discrete seed. We voxelize the 3D world and run A* [51] graph search to find a globally connected seed path. For neighboring voxels q_i and q_j, with midpoint q_{ij} and $\Delta q = q_j - q_i$, the edge cost is

$$c_{ij} = \sqrt{\Delta q^\top G(q_{ij})\Delta q} + \|\Delta q\|_2 V(q_{ij}). \tag{18}$$

This seed already avoids most unsafe shortcuts.

Continuous refinement. We resample the seed into a curve and optimize a small set of control points with field, smoothness, and anchor terms:

$$E(\tau) = E_G + E_V + \lambda_{\text{dyn}} E_{\text{dyn}} + \lambda_{\text{anchor}} E_{\text{anchor}}. \tag{19}$$

Here E_G samples the Riemannian metric in Eq. (16), and E_V samples the potential in Eq. (17). Thus goal reaching, safety, and efficiency keep the same symbols and weights as the main objective. The dynamics term penalizes acceleration and sharp turns:

$$E_{\text{dyn}} = \sum_k \|q_{k+1} - 2q_k + q_{k-1}\|_2^2. \tag{20}$$

For goal-conditioned image-to-video generation, the anchor term keeps the refined path near the graph seed. For video dynamics editing, it anchors the refined motion to the source tracks so that the task intent is preserved. We use a gradient-based solver on the control points and then resample the optimized curve into frame-level controls for generation.

C.3 Generation

We use Diffusion as Shader (DaS) [39] as the tracking-conditioned renderer. The original DaS interface expects a tracking video as condition input. We modify the bridge so that this control video is generated from metric 3D points rather than only 2D image displacements.

For goal-conditioned image-to-video generation, we load the MoGe2 [45] point map, camera intrinsics, selected object mask, and planned 3D path. Points inside the selected object mask receive the planned rigid translation. Support-surface and background points remain static. We then project all 3D points back to pixels in each frame and render a dense tracking video. This makes the control signal camera-aware and depth-aware.

For video dynamics editing, foreground and arm tracks are initialized from SpatialTrackerV2. The optimized 3D tracks replace the raw foreground motion. Background tracks keep their source motion or remain static, depending on the camera motion. We render the optimized tracks into the same DaS tracking-video format. These changes let DaS follow the planned 3D dynamics while preserving scene appearance.

D Metric Details

We compute motion metrics from foreground 3D tracks. Let p_t be the representative object track at frame t, g be the 3D goal, and S_{unsafe} be the unsafe region.

Goal reaching. For goal-conditioned image-to-video generation, goal error is the final distance to the target:

$$\text{GoalError} = \|p_T - g\|_2. \tag{21}$$

For video dynamics editing, track deviation compares the generated track $\hat{p}_t$ with the source track p_t after first-frame alignment:

$$\text{TrackDeviation} = \frac{1}{T} \sum_{t=1}^{T} \|(\hat{p}_t - \hat{p}_1) - (p_t - p_1)\|_2. \tag{22}$$

Safety. Violation among successful cases is measured only on goal-conditioned image-to-video generations that reach the goal. It is the fraction of these generations whose foreground track enters the unsafe set:

$$\text{ViolationAmongSuccess} = \frac{\sum_i \mathbb{1}[\text{success}_i]\mathbb{1}[\exists t,\ p_{i,t} \in S_{\text{unsafe}}]}{\sum_i \mathbb{1}[\text{success}_i]}. \tag{23}$$

Smoothness. Acceleration and jerk measure second- and third-order temporal changes:

$$\text{Acceleration} = \frac{1}{T-2} \sum_{t=2}^{T-1} \|p_{t+1} - 2p_t + p_{t-1}\|_2, \tag{24}$$

$$\text{Jerk} = \frac{1}{T-3} \sum_{t=2}^{T-2} \|p_{t+2} - 3p_{t+1} + 3p_t - p_{t-1}\|_2. \tag{25}$$

Efficiency. Path length is the total traveled distance:

$$\text{PathLength} = \sum_{t=1}^{T-1} \|p_{t+1} - p_t\|_2. \tag{26}$$

Energy is the accumulated motion cost:

$$\text{Energy} = \sum_{t=1}^{T-1} \|p_{t+1} - p_t\|_2^2. \tag{27}$$

For video quality, we report VBench motion smoothness, background consistency, and temporal flickering using the official VBench implementation.

E More Results for Goal-Conditioned Image-to-Video Generation

Table 7 reports motion-quality results with standard deviations. Figure 6 provides additional visual examples for goal-conditioned image-to-video generation.

Table 7: **Additional quantitative results on goal-conditioned image-to-video generation.** Values are mean ± standard deviation.

Method	Goal error ↓	Violation among successful cases ↓	Acceleration ↓	Jerk ↓	Energy ↓
HunyuanVideo-1.5	0.548 ± 0.465	0.5232 ± 0.3114	0.0118 ± 0.0168	0.0197 ± 0.0301	0.0576 ± 0.1550
Wan2.2	0.549 ± 0.371	0.5230 ± 0.2607	0.0172 ± 0.0235	0.0304 ± 0.0437	0.0860 ± 0.2172
Cosmos-Predict2.5	0.733 ± 0.375	0.2857 ± 0.2440	0.0104 ± 0.0134	0.0186 ± 0.0256	0.0398 ± 0.1409
VLIPP	0.331 ± 0.316	0.5110 ± 0.2694	0.0099 ± 0.0190	0.0176 ± 0.0366	0.0419 ± 0.2171
OptiWorld	0.310 ± 0.317	0.2599 ± 0.2641	0.0070 ± 0.0031	0.0116 ± 0.0053	0.0126 ± 0.0100



Figure 6: **Additional goal-conditioned image-to-video results.**

F More Results for Video Dynamics Editing

Table 8 reports video dynamics editing results with standard deviations. Figure 7 provides additional visual examples for video dynamics editing.

Table 8: **Additional quantitative results on video dynamics editing.** Values are mean ± standard deviation.

Method	Track deviation ↓	Acceleration ↓	Jerk ↓	Path length ↓	Energy ↓
Source video	–	0.0124 ± 0.0128	0.0205 ± 0.0229	0.647 ± 0.342	0.0269 ± 0.0398
Wan2.1 first-last-frame-to-video	0.123 ± 0.070	0.0157 ± 0.0125	0.0266 ± 0.0235	0.733 ± 0.343	0.0387 ± 0.0431
OptiWorld	0.113 ± 0.072	0.0107 ± 0.0121	0.0181 ± 0.0213	0.524 ± 0.346	0.0236 ± 0.0489

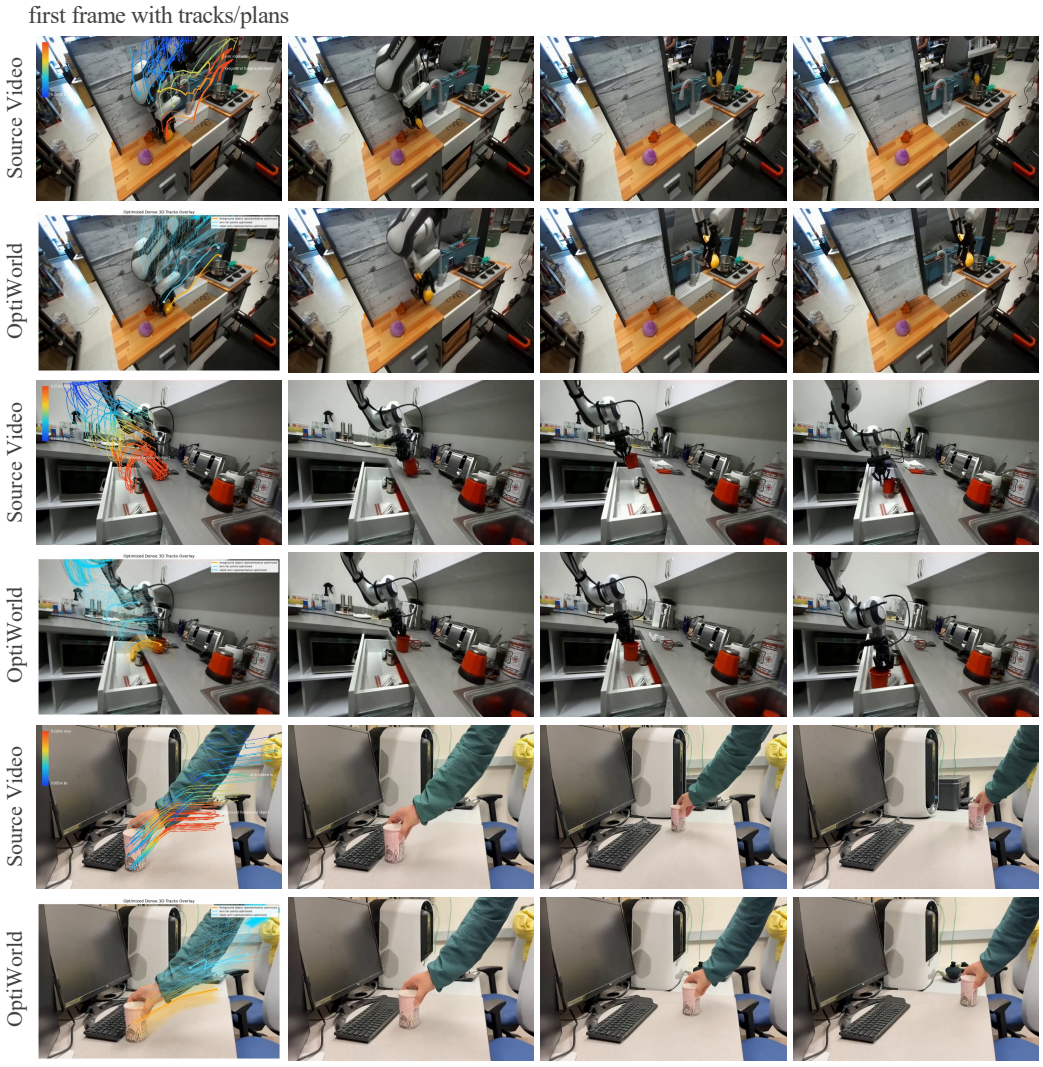


Figure 7: **Additional video dynamics editing results.** OptiWorld refines the source 3D motion before rendering, producing smoother and shorter object trajectories while keeping the first-frame scene content stable. Note that OptiWorld only relies on the **first frame**, prompt, and the optimized tracks for video dynamics editing.

G More Results for Ablation Study

Figure 8 provides additional visual examples for the ablation study. Removing VLM reasoning weakens task-dependent semantic constraints. Removing the manifold, safety, efficiency, or smoothness terms produces paths that are less safe, less direct, or less stable.

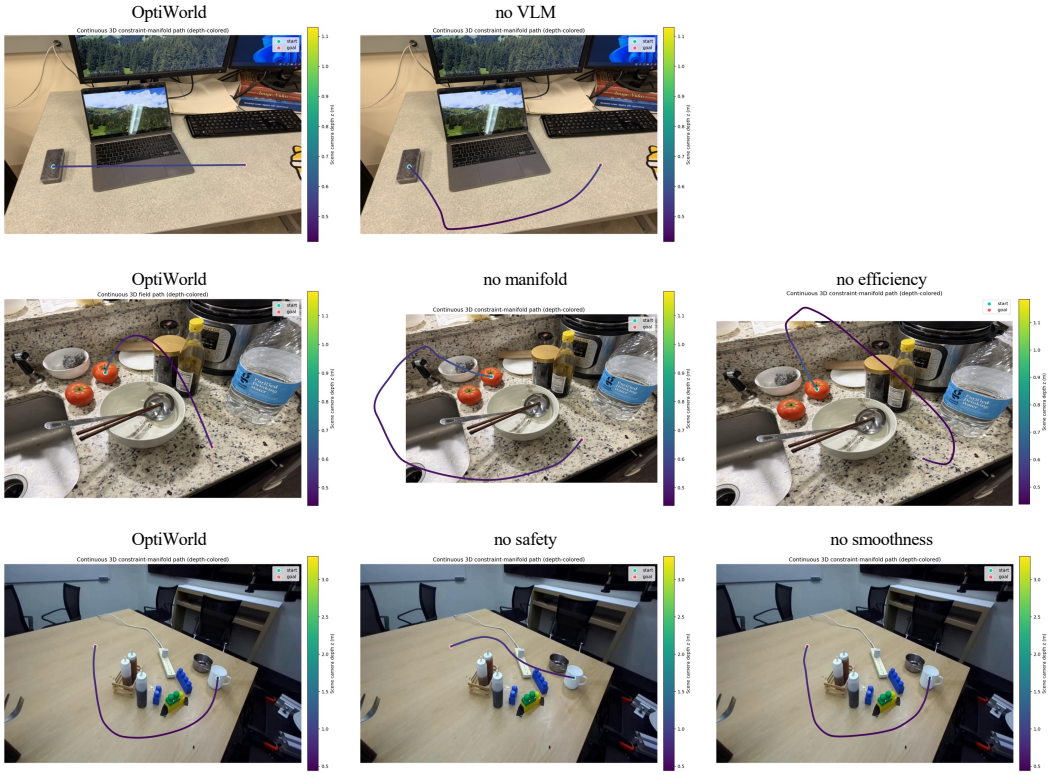


Figure 8: **Additional ablation study results.**

H More Results for Counterfactual Physics Generation

Figures 9 and 10 show additional counterfactual generations produced by changing only the planner. Changing the goal gives different optimized trajectories for the same initial scene. Changing the safety weight changes whether the path avoids or enters risky regions, while the understanding and rendering modules remain fixed.

I Broader Impacts

OptiWorld may support creative and educational applications. If used without clear AI-generated content markers, its outputs could contribute to misinformation.

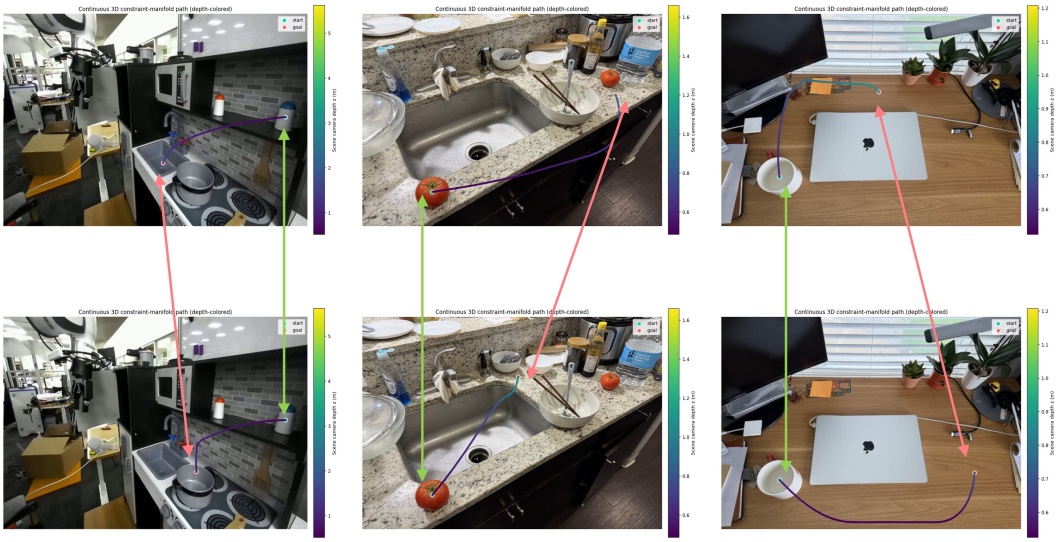


Figure 9: **Additional counterfactual results of goal control.** Different goal sets produce different optimized paths.

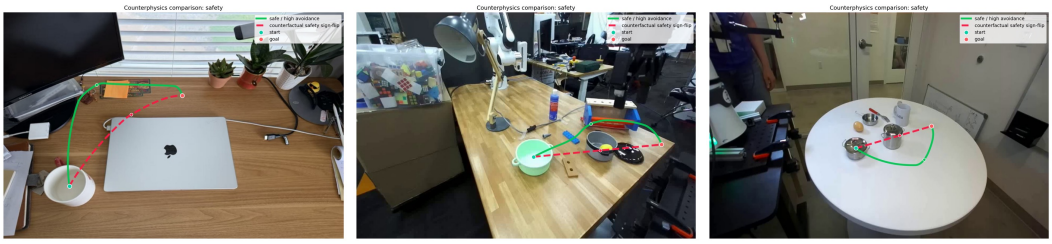


Figure 10: **Additional counterfactual results of safety control.** Lowering the safety constraint weight allows hazardous shortcuts, while the full planner avoids risky regions.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally

expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: For each claim made in the abstract and introduction, we provide reference or experiments.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in a separate section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide the full set of assumptions and complete (and correct) proof.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all the information needed to reproduce the main experimental results. We talk the data, model and training strategy we use.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will release all the coda and data we use in this paper.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the training and test details in the main paper and supplemental material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide detailed analysis in the Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the compute resources and time in the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We checked the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss the Broader impacts at the Appendix.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: We will make arrangements to host the data and code on a secure platform and to provide the necessary licenses.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: For all data and pre-trained models, we put them in safe platform and provide necessary licenses.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer:[Yes]

    Justification: New assets introduced in the paper are well documented.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with real human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with real human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper describe the usage of LLMs in the Method section.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.