『2026 동계 세미나』

Controllable Video Generation with Physical Priors

Pusan National University
Dept. of Electronics Engineering
Computer Vision & Signal Processing Lab.

Presented By
문형주

Outline

- Motivation
- WonderPlay: Dynamic 3D Scene Generation from a Single Image and Actions [ICCV'25 Highlight]
- Force Prompting: Video Generation Models Can Learn and Generalize Physics-based Control Signals [NeurIPS'25]

Motivation

- ECCV2026
  - Topic : 3D motion generation
    - World model에서의 interaction을 통한 motion 생성을 목표로 함

∴ 기존 연구들이 *world model*에서의 *interaction*을 통한 *motion* 생성 어떤식으로 하고 있는지 동향 파악



WonderPlay: Dynamic 3D Scene Generation from a Single Image and Actions [ICCV'25 Highlight]

Introduction

- Goal
  - Goal
    - Single image 입력으로부터 다양한 physics materials를 포함하는 action-conditioned dynamic 3D scene 생성





Introduction

- Problem & Solution
  - Problem
    - 사용자 action 입력이 주어졌을 때 dynamic 3D scene을 계산하기 위해 physics simulation에 의존하는 경우가 많음
      - Scene에 포함된 모든 dynamic 유형에 대한 정확한 physcis solver가 필요함
        - ✓Solid-fluid two-way coupling과 같은 정밀한 physics solver 없음
      - 제한된 관측(single image)로부터 완전한 physics state를 복원해야 함
      - 기존 방법들 **강체**와 **단순 탄성체**에 해당하는 dynamic만으로 제함됨

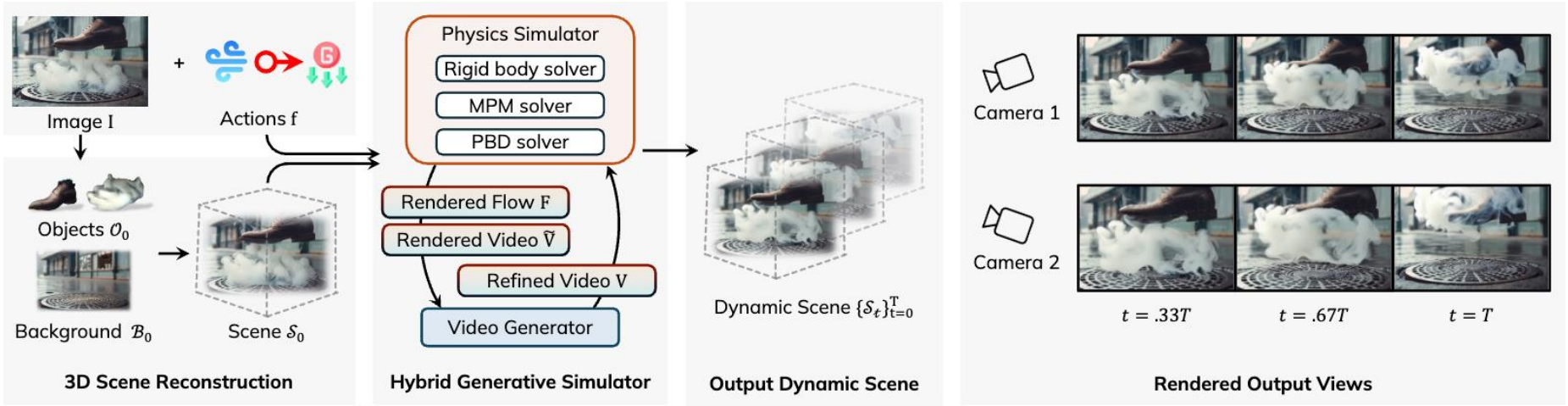
➡다양한 실제 물리 현상을 포함하는 extensive real-world 비디오로부터 학습된 Video Generation model이 제공하는 data prior 활용 필요성 제기

Introduction

- Problem & Solution
  - Problem
    - Video generation model은 정확한 3D action 입력을 받아 그에 따른 dynamic을 simulation하는 능력을 갖추고 있지 않음
  - Solution
    - Hybrid generative simulator
      - Physics solver와 video generation model을 결합
      - physics solver만으로 action을 반영한 dynamic 3D scene이 가지는 부정확하고 현실성이 떨어지는 문제를 video generation model의 data prior로 개선
    - Motion-appearance bimodal control
      - Conditioned video generation 단계에서 simulator 정보를 최적으로 활용하기 위한 전략

Method

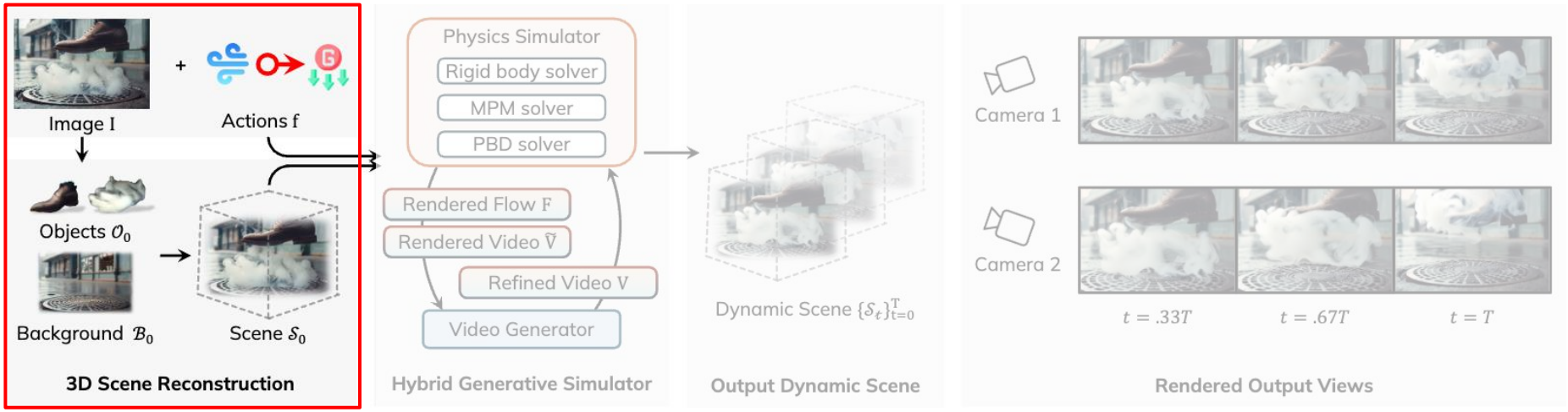
- Overview
  - Pipeline
    - Input : Single image I, Actions f
      - ✓3가지 action type : 중력 f_g, 3D force field $f_w(x, y, z, t)$, 3D point forces $f_p(t)$
    - Output : Initial scene에 action을 적용한 결과
      - dynamic 3D scene $\{S_t\}_{t=0}^{T}$
    - 강체, 탄성체, 천, 연기, 액체, 입자성 물질 및 이들 간의 상호작용을 포함하는 dynamics를 시뮬레이션



< Overview of WonderPlay >

Method

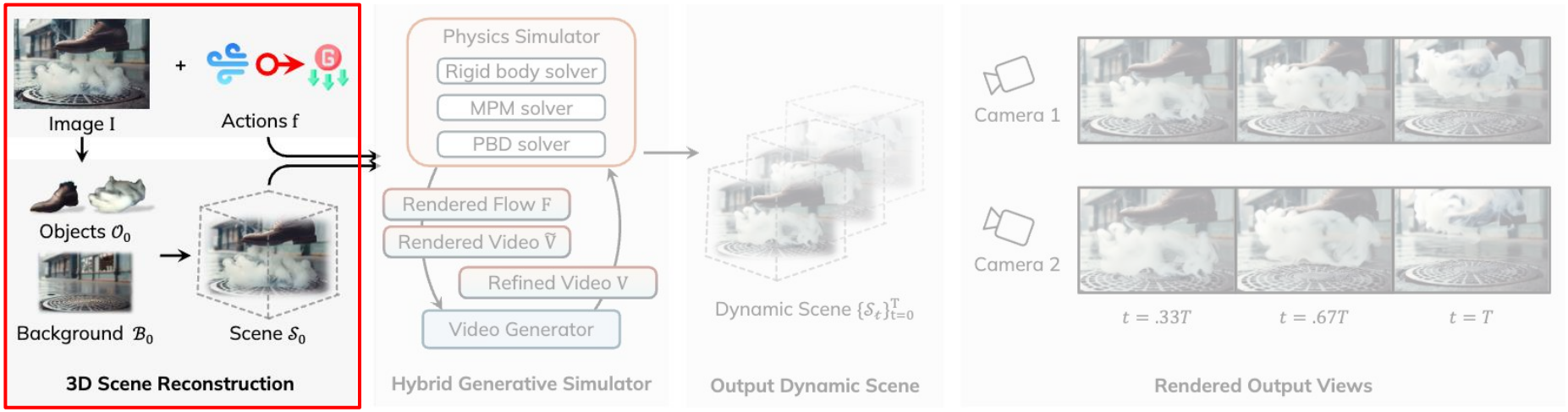
- 3D Scene Reconstruction
  - Initial 3D scene
    - Single image 입력으로부터 initial 3D scene S_0 reconstruction
      - Background와 object로 구성
        - ✓$S_0 = B_0 \cup O_0$



< Overview of WonderPlay >

Method

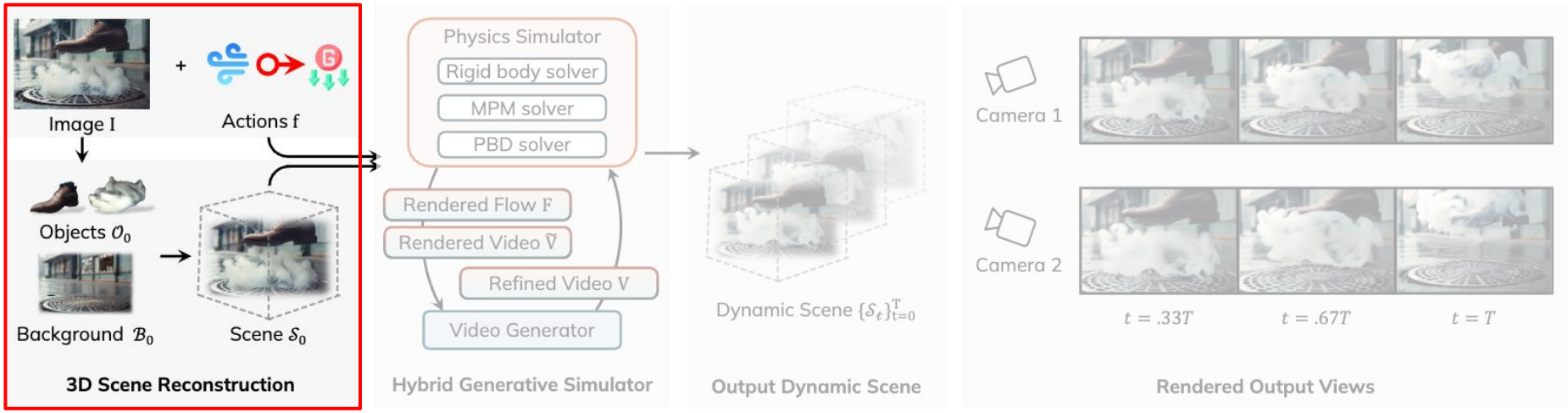
- 3D Scene Reconstruction
  - Background
    - $B_t = \{p^B, q^B, s^B, o^B, c_t^B\}$
      - N_B개의 Gaussian surface로 구성
      - p : 위치, q : 방향 quaternion, s : 크기, o : opacity, c_t : RGB 색상



< Overview of WonderPlay >

Method

- 3D Scene Reconstruction
  - Object
    - Physics solver에서 simulation되는 dynamic 대상
      - 호환 가능한 simulation object를 표현하기 위해 Gaussian surface에 connectivity 추가
      - Topological Gaussian surfaces로 확장



< Overview of WonderPlay >

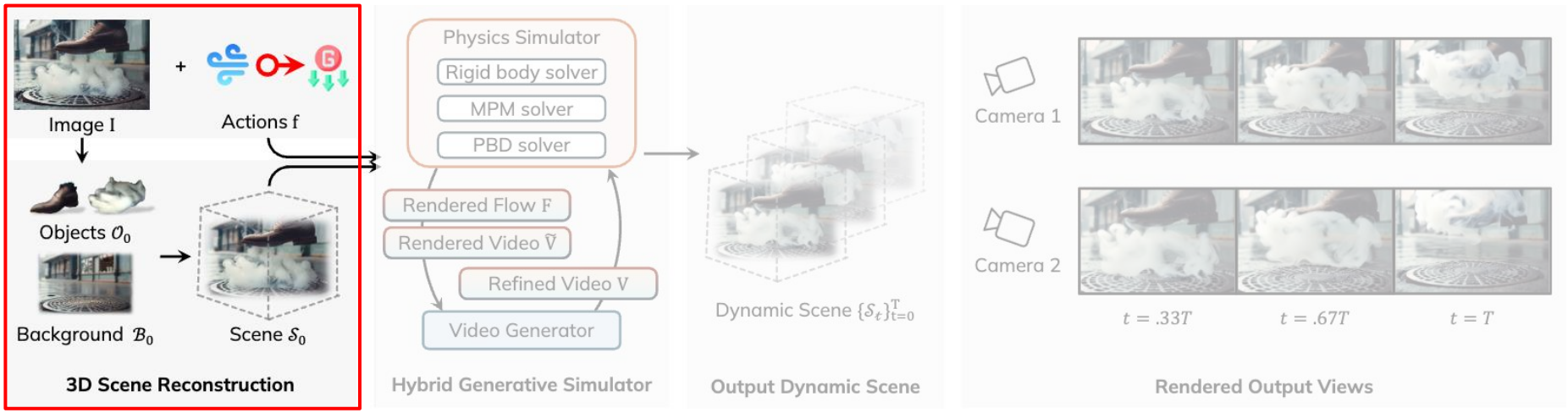
Method

- 3D Scene Reconstruction
  - Object
    - Topological Gaussian surfaces

$O_t = \{E, v_t, p_t^o, q_t^o, s^o, o^o, c_t^o\}$

✓ $E \in \{0,1\}^{N_o \times N_o}$: edge 행렬, surface 간 위상적 연결성

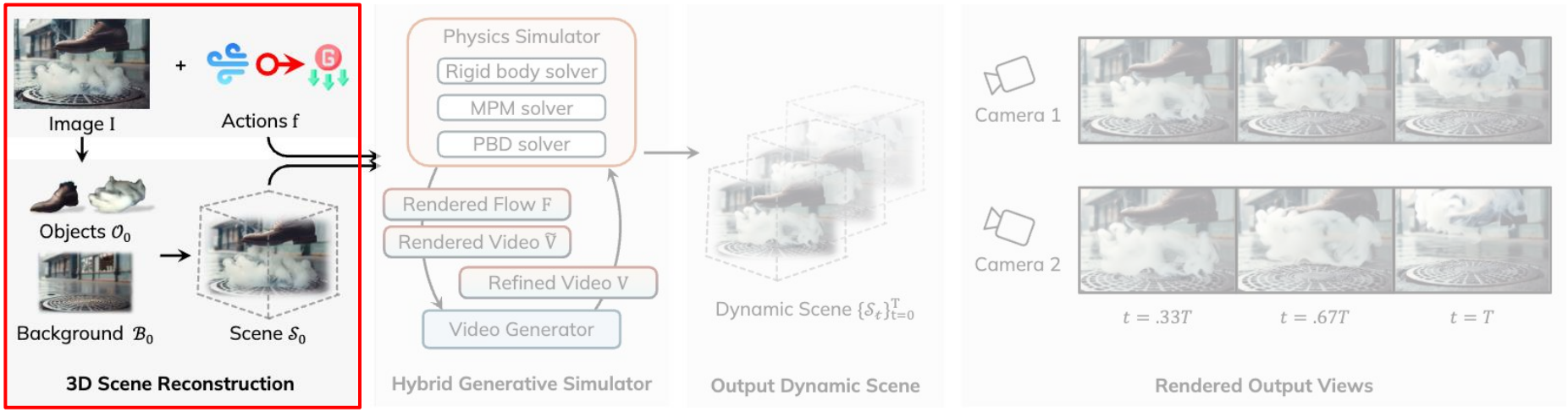
✓ $v_t \in \mathbb{R}^{3N_o}$: 속도



< Overview of WonderPlay >

Method

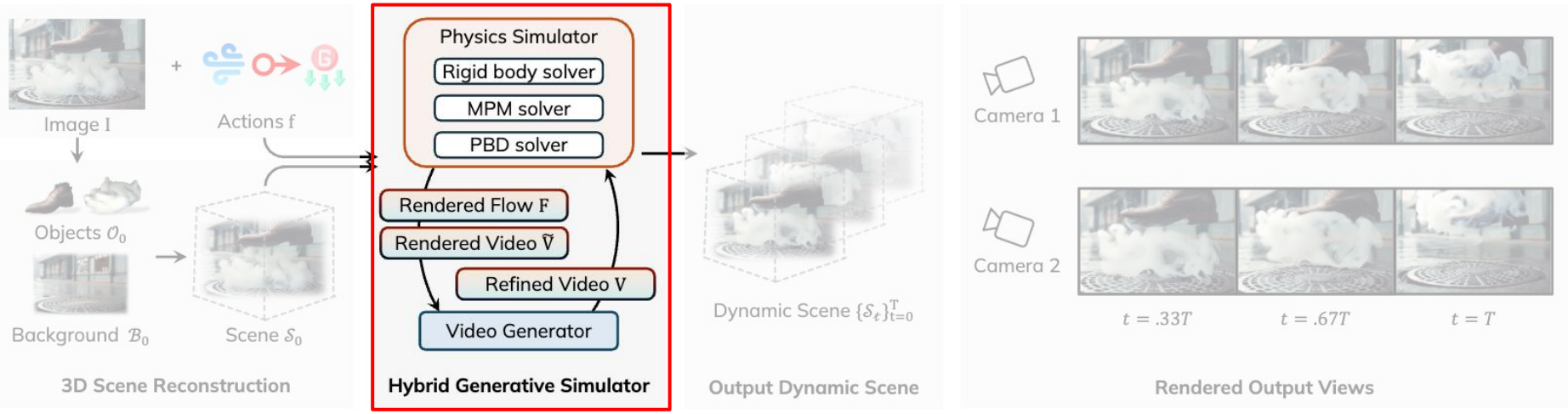
- 3D Scene Reconstruction
  - Object
    - Materials
      - Geometry & appearance 표현 외에도 object는 material properties m을 가짐
        - ✓VLM이 object type과 physics parameter를 추정



< Overview of WonderPlay >

Method

- Hybrid Generative Simulator
  - Main idea
    - Real-world 물리현상이 포함된 video로 학습된 video generation model로부터 dynamic 지식을 추출
      1. Physics solver를 사용해 coarse dynamic scene $\{\tilde{S}_t\}_{t=1}^{T}$ 추정
      2. Coarse dynamic scene을 condition으로 사용하여 video generator를 통해 현실적인 dynamics를 갖는 비디오 V를 합성
      3. V 를 감독신호로 사용하여 coarse dynamic scene을 갱신하여 최종 dynamic scene $\{S_t\}_{t=1}^{T}$ 얻음



< Overview of WonderPlay >

Method

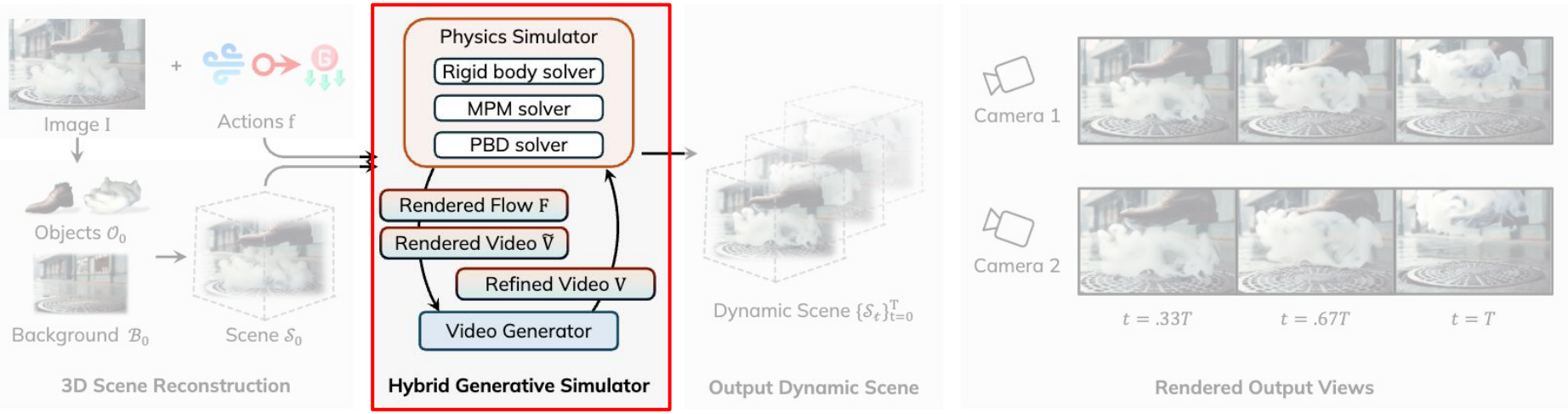
- Hybrid Generative Simulator
  - Physics solver
    - 각 simulation time step에서 다음 time step에서의 object dynamic 속성 계산

$$v_{t+1}, p^o_{t+1}, q^o_{t+1} = \text{solver}\left(\tilde{S}_t, f_g, f_w(t), f_p(t)\right)$$

$$\tilde{S}_0 = S_0$$

✓Next step : $\tilde{S}_{t+1} = B_0 \cup \{E, v_{t+1}, p^o_{t+1}, q^o_{t+1}, s^o, o^o, c^o_t\}$

✓비동적 속성은 S_0와 동일하게 유지



< Overview of WonderPlay >

Method

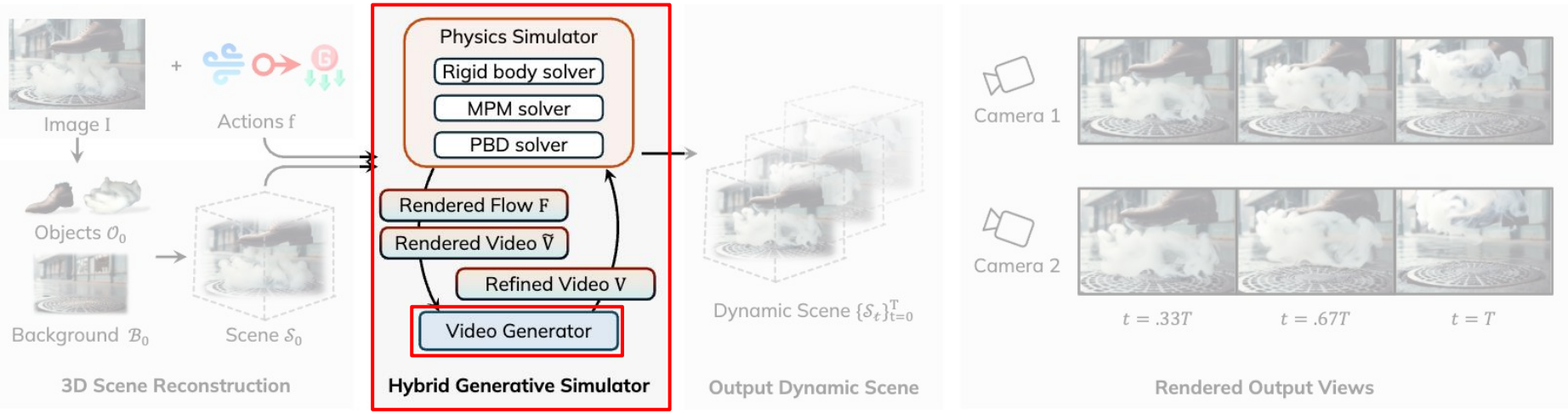
- Hybrid Generative Simulator
  - Conditioning Video Generator
    - Coarse dynamic scene $\{S_t\}_{t=0}^{T}$을 조건으로 하여 더 세밀한 움직임을 포함하면서도 action 결과를 따르는 video $V \in \mathbb{R}^{(T+1)\times H\times W\times 3}$를 합성

➡ 이를 위해 bimodal control 방식을 도입 {motion(flow), appearance(RGB)}

$V = g(\mathrm{F}, \tilde{V}, \mathrm{I})$

✓ g : video generator, $\mathrm{F} \in \mathbb{R}^{T\times H\times W\times 2}$: 속도 $\{v_t\}_{t=1}^{T}$로부터 렌더링된 pixel flow

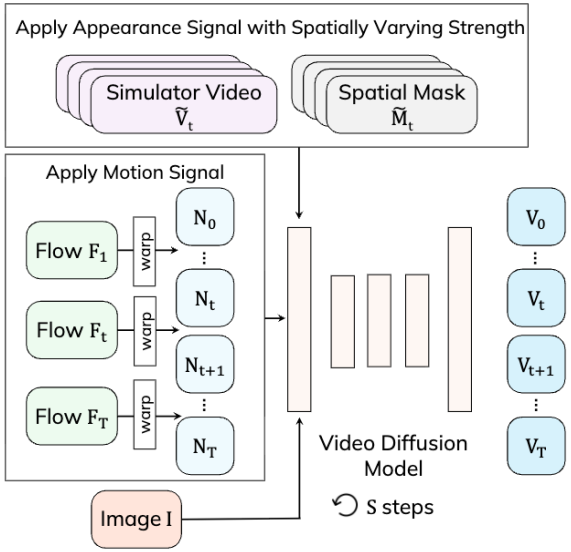
✓ $\tilde{V} \in \mathbb{R}^{(T+1)\times H\times W\times 3}$: coarse scene $\{\tilde{S}_t\}_{t=0}^{T}$로부터 렌더링된 video, I : input image



< Overview of WonderPlay >

Method

- Hybrid Generative Simulator
  - Conditioning Video Generator
    - Bimodal control – Motion control
      - Pretrained motion controlled i2v diffusion model Go with the Flow를 generator g로 사용
      - Motion control은 noise warping에 기반
        - ✓Random Gaussian noise 대신 warping 기반 구조화된 noise $\mathrm{N(F)} \in \mathbb{R}^{(T+1)\times H\times W\times 3}$ 사용

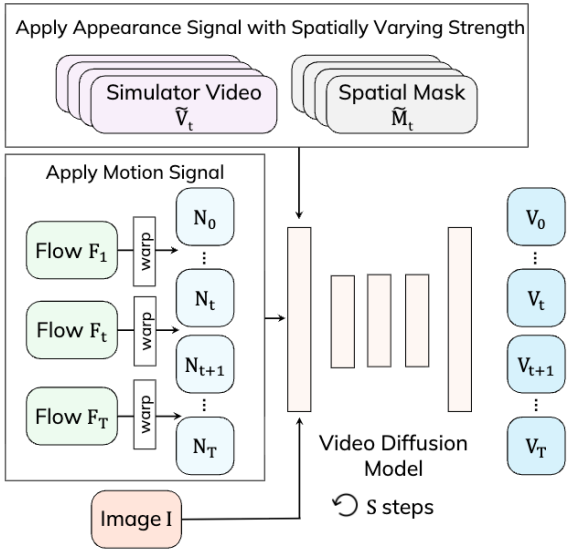


< Overview of WonderPlay >

1. Random Gaussian $\mathrm{N}_0 \in \mathbb{R}^{H\times W\times 3}$을 샘플링
2. $\mathrm{N}_{t+1} = warp(\mathrm{N}_t, \mathrm{F}_{t+1})$을 반복해서 생성
3. Visual quality 향상을 위해 일부 random noise와 결합 (감쇠 계수 γ에 의해 제어됨

Method

- Hybrid Generative Simulator
  - Conditioning Video Generator
    - Bimodal control – RGB control
      - RGB 프레임 $\tilde{V}$를 통한 추가 control을 위해 SDEdit 사용
        - ✓Diffusion 기반 generation 과정 N(F)를 점진적으로 복원
          - $V_{s=1} = Denoise(V_s, s)$
            - $s = S, S-1, \ldots, 1$: Diffusion timestep, $V_S = \mathrm{N(F)}$: initial noise



< Overview of WonderPlay >

처음 몇 step을 건너뛰고 $s_1 < S$에서

$V_{s_1} = \alpha_{s_1}\tilde{V} + \sqrt{1-\alpha_{s_1}^2}\,\mathrm{N(F)}$로 시작함으로써 제어됨

α_i : diffusion coefficient

Experiments

- Baselines
  - Physics-based method
    - PhysGen
      - 2D 강체
    - PhysGaussian
      - MPM 프레임워크를 사용해 장면을 탄성 객체로 모델링
  - Conditioned video generation method
    - CogVideoX-I2V
      - 텍스트 프롬프트 기반
    - Tora
      - 드래그 기반 conditioned

Experiments

- Metrics
  - 입력 시점에서 렌더링된 비디오를 사용해 quntitaive result 평가
    - Vbench에서 제안한 maging, aesthetic, motion quality, consistency 지표 사용
    - GPT-4o 기반 physical realism 지표 사용
  - 실제 사진 7개, 사실적인 합성 이미지 8개로 총 15개 예제 구성

Experiments

- Quantitative Comparison to baselines
  - Human study 2AFC results

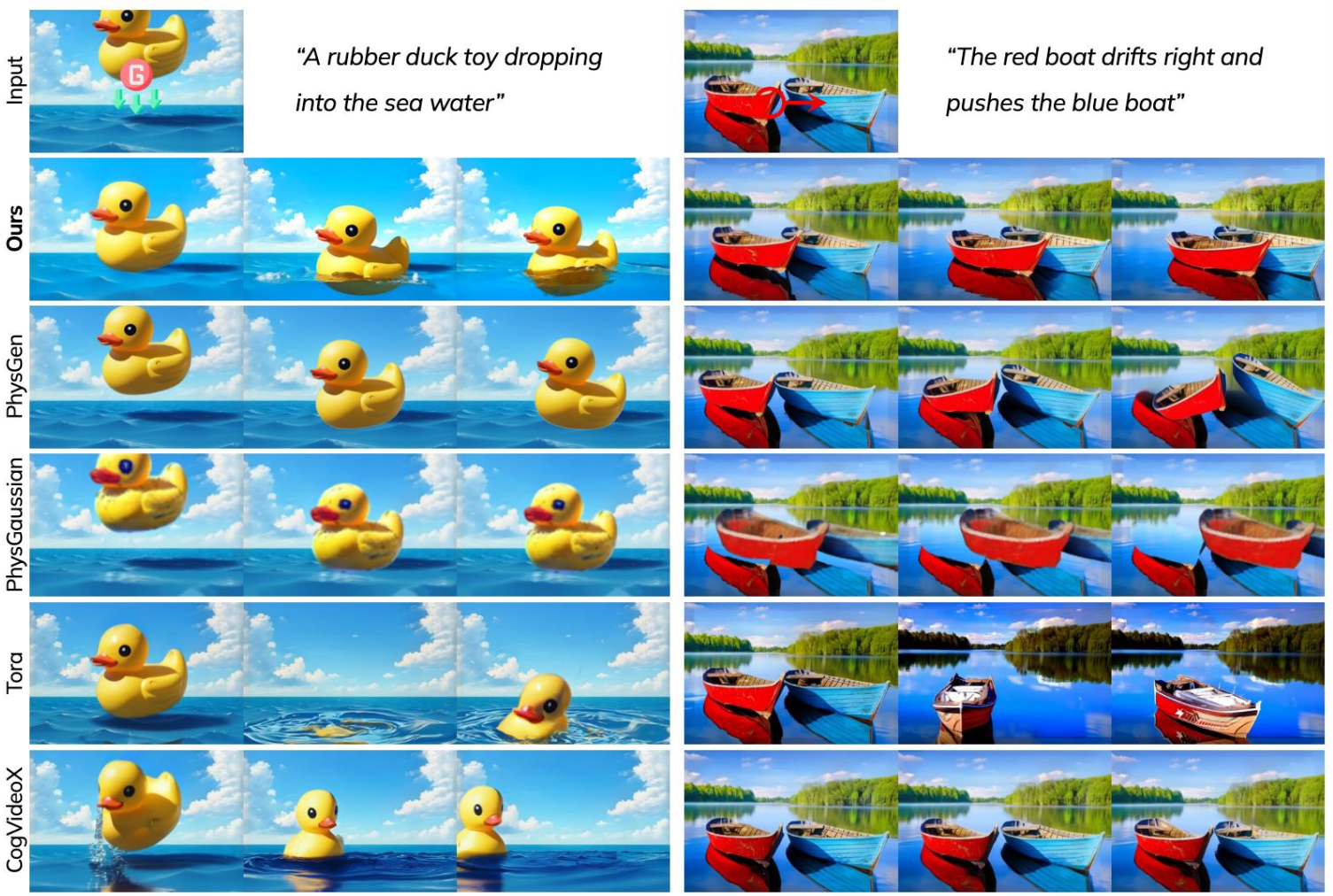
	Physics Plausibility	Motion Fidelity	Visual Quality
Over PhysGen [43]	78.0%	78.0%	80.1%
Over PhysGaussian [67]	80.2%	81.2%	85.2%
Over Tora [74]	77.0%	72.0%	71.0%
Over CogVideoX-I2V [70]	80.2%	73.0%	74.6%

  - Vbench results

Methods	Imaging (↑)	Aesthetic (↑)	Motion (↑)	Consistency (↑)	PhysReal (↑)
PhysGen	<u>0.692</u>	0.593	0.992	0.212	0.545
PhysGaussian	0.492	0.564	<u>0.994</u>	0.206	0.350
CogVideoX	0.686	0.574	0.993	**0.219**	0.670
Tora	0.644	**0.620**	0.992	0.210	0.530
Ours	**0.695**	<u>0.610</u>	**0.995**	<u>0.217</u>	**0.700**

Experiments

- Qualitative Comparison to baselines



Experiments

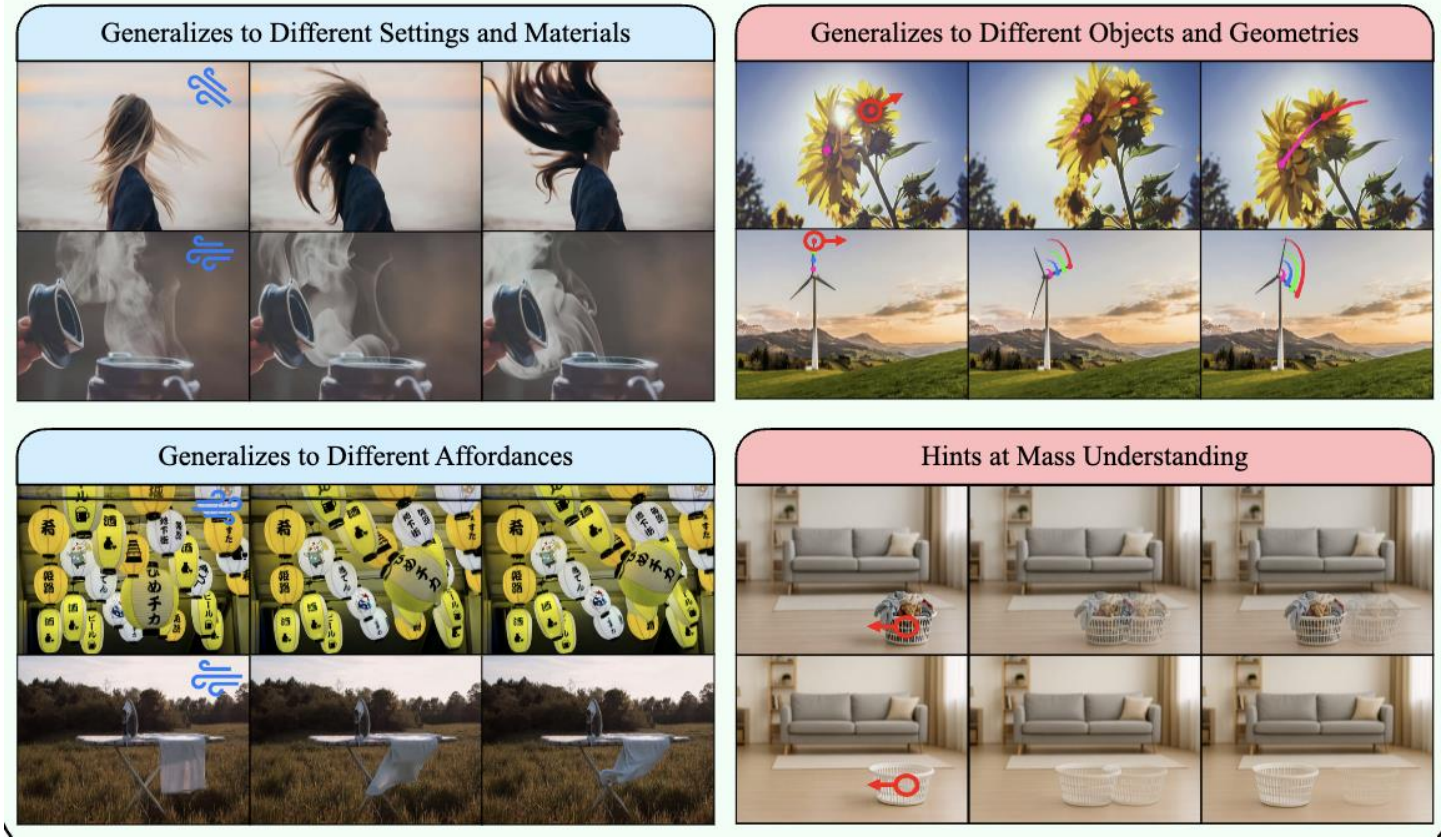
- Ablation
  - Ablation studies about bimodal control



Force Prompting: Video Generation Models Can Learn and Generalize Physics-based Control Signals[ACM SIGGRAPH 24]

Introduction

- Goal
  - Goal
    - 사용자가 물리적 힘을 통해 이미지와 interaction할 수 있도록 하는 것

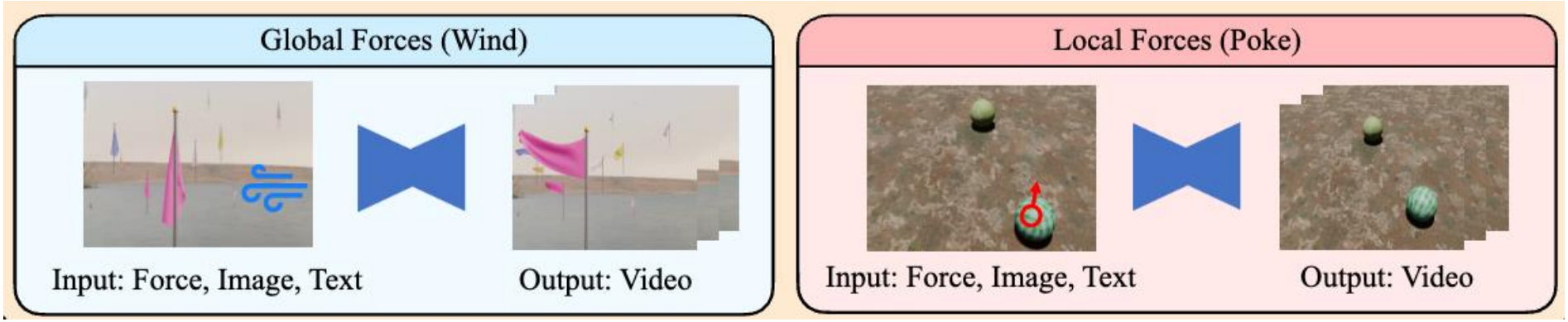


Introduction

- Problem & Solution
  - Problem
    - Text, trajectory 기반 control은 왜 움직이는지 설명하지 못함
    - Physics simulator 기반 method는 3D asset, 계산 비용이 큼
    - 실제 비디오에서 force label 수집이 거의 불가능
  - Solution
    - Force Prompting
      - 힘의 크기, 방향, 위치를 명시적 control 신호로 정의
      - Global force(바람), local force(찌르기) 두 가지 설정
      - ControlNet으로 force 정보를 video model에 주입
    - Outcome
      - 소량의 합성 데이터만으로도 강한 일반화 성능
      - 질량, material type에 따른 직관적 물리 반응 생성
      - 물리 엔진 없이도 action-conditioned world model 구현

Method

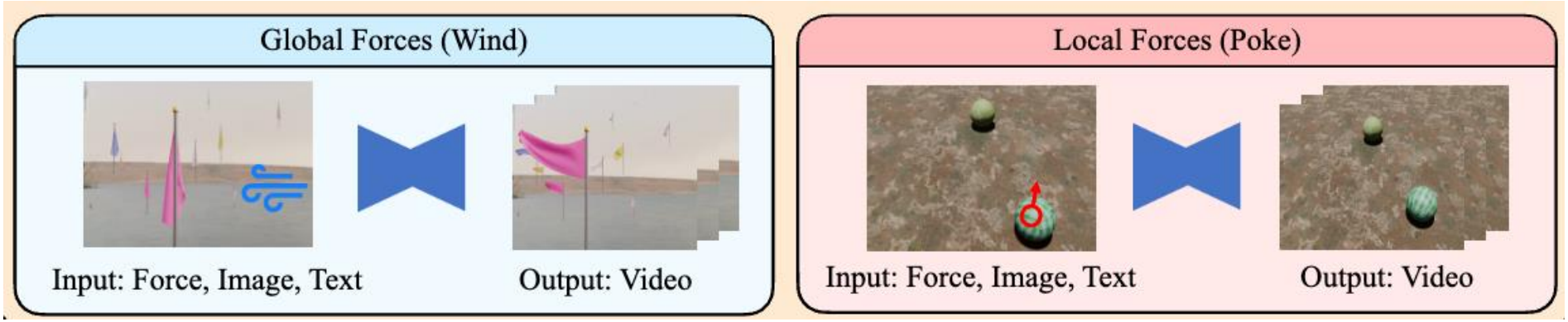
- Overview
  - Pipeline
    - Input : Image + text + Force Prompt
      - Force Prompt는 ControlNet으로 인코딩하여 주입
    - Output : force가 적용된 비디오
    - Synthetic Force-Video data
      - Physics simulator로 force-video 쌍 생성
      - 제한된 객체만 사용하여 생성



< Motion-I2V Framework>

Method

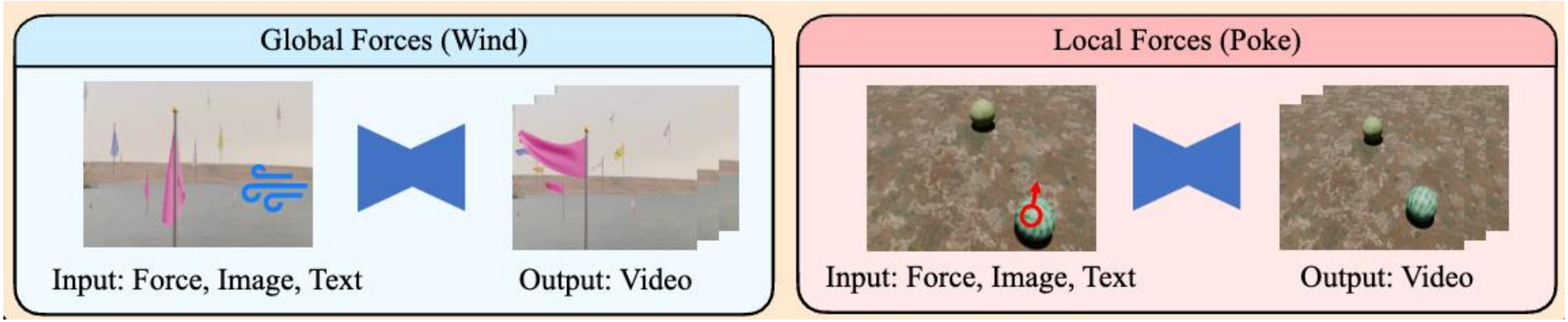
- Synthetic training data
  - Global forces + Local forces\
    - Global forces dataset
      - Physics simulator를 사용하여 바람에 흔들리는 깃발의 비디오를 생성
    - Local forces dataset
      - 두 부분으로 구성
        - ✓Physics simulator를 사용해 공이 바닥 위를 굴러가는 비디오를 생성
        - ✓3D Gaussian과 physics simulator를 결합한 모델을 사용하여 식물이 찔리는 비디오를 생성



< Motion-I2V Framework>

Method

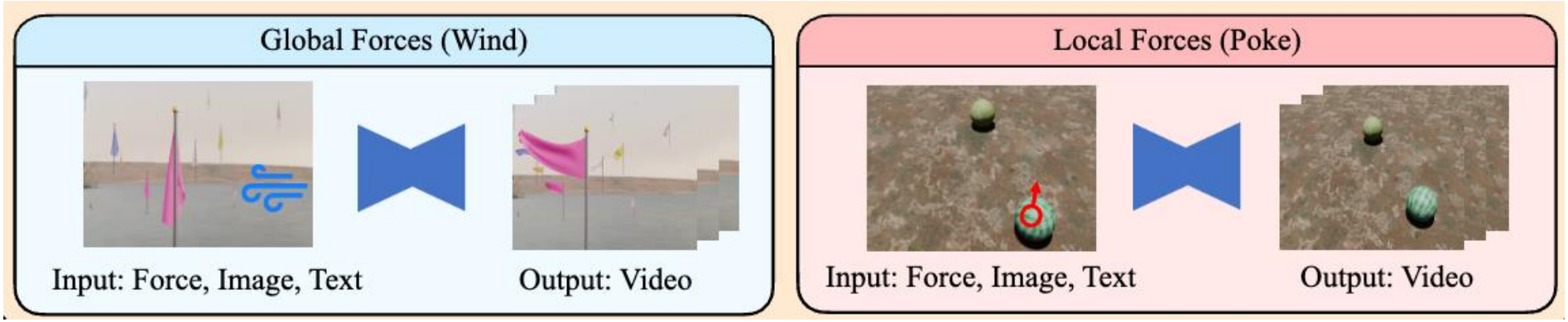
- Synthetic training data
  - Global forces dataset
    - Blender를 사용해 다양한 바람 condititon에 반응하며 움직이는 깃발 데이터셋을 구축
      - 데이터 다양성을 확보하기 위해, 각 비디오마다 여러 파라미터 무작위로 설정
      - 깃발 개수, 색상, 위치, 카메라 배치, HDRI, 바람 방향, 바람세기 등이 포함됨
    - 총 15000개 비디오로 구성



< Motion-I2V Framework>

Method

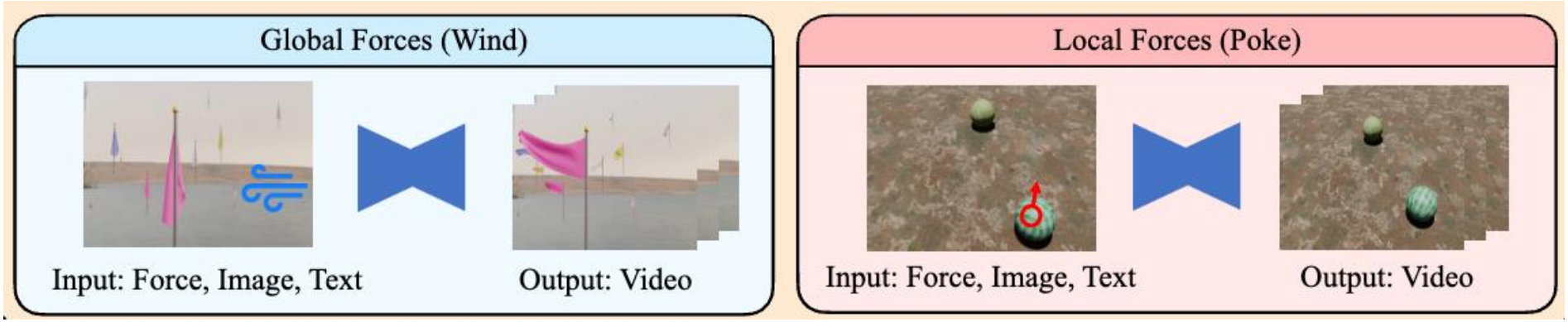
- Synthetic training data
  - Local forces dataset – 2개로 구성
    - 첫 번째 시나리오
      - Blender를 사용해 파라미터들을 무작위로 설정하여 하나의 공만의 보이지 않는 point-wise force에 의해 굴러가는 비디오를 생성
      - 공 개수, texture, 색상, 위치, 카메라 위치, 지면 texture, 목표 공 선택, 힘 각도, 힘 크기 포함
        - ✓질량 기반 dynamic을 모델이 학습하도록 하기 위해, 볼링공의 질량을 축구공의 4배로 설정
      - 12000개로 구성



< Motion-I2V Framework>

Method

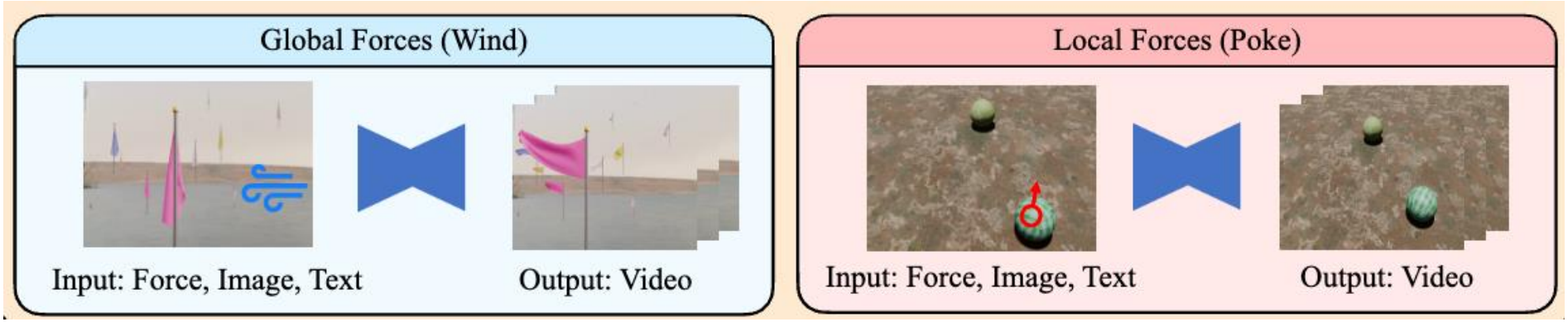
- Synthetic training data
  - Local forces dataset – 2개로 구성
    - 두 번째 시나리오
      - Generation-simulator hybrid 모델인 PhysDreamer를 사용하여 보이지 않는 힘에 의해 찔릴 때 식물이 앞뒤로 흔들리는 비디오를 포함하는 데이터셋 생성
      - 카메라 위치, 접촉지점, 힘 각도, 힘 크기를 무작위로 설정하여 생성
      - Object의 종류에 따라 복잡한 진동 dynamic도 모델에 가르치기 위해, 두 시나리오를 혼합한 데이터셋 사용



< Motion-I2V Framework>

Method

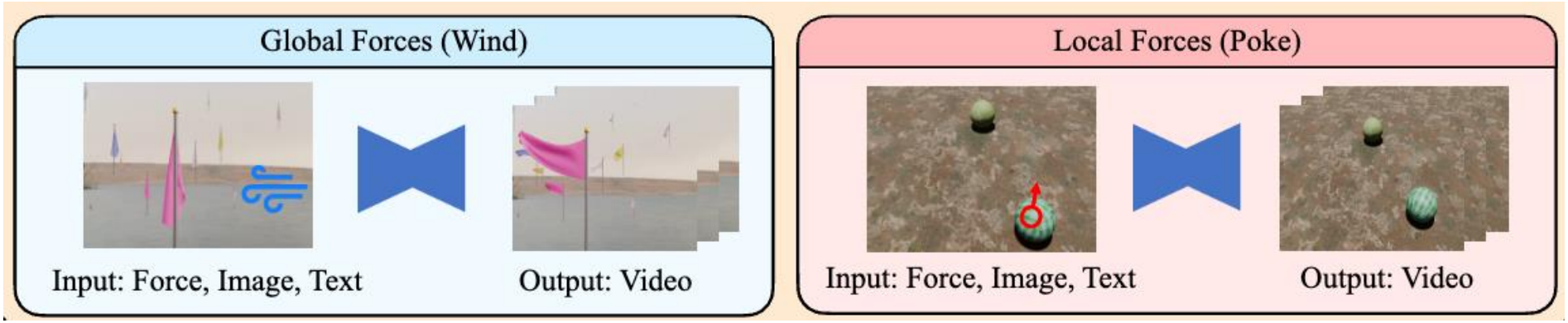
- Synthetic training data
  - Both dataset
    - 카메라 파라미터를 사용해 3D 공간의 forces를 2D pixel 평면으로 투영
    - 물리 세계 좌표계에서의 힘 벡터와 object 위치를 화면 좌표계로 매핑
      - 이미지 프레임 내에서 힘을 모델링할 수 있도록 함
    - GPT-4o API를 사용해 상세한 text prompt 생성
      - 각 HDRI 배경과 지면 texture에 대해 고유한 설명을 생성



< Motion-I2V Framework>

Method

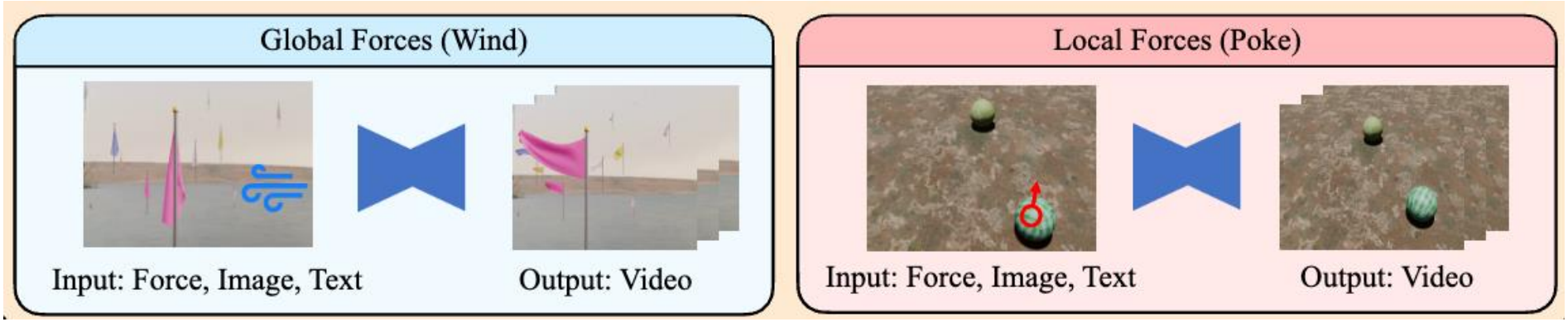
- Local and Global Force Prompts
  - Encoding strategy
    - Global forces는 global하게 적용되고, point forces는 local하게 적용되므로 두 가지 서로 다른 force encoding 전략 제안
    - Global forces
      - 바람에 대한 control 신호는 힘 $F \in [0,1]$와 각도 $\theta \in [0,360)$으로 파라미터화됨
      - 목표 : force prompt π에 대한 텐서 표현을 정의하는 것
        - ✓$\tilde{\pi} \in \mathbb{R}^{f \times c \times h \times w}$로 표기
          - f : 프레임 수, c : color 채널 수, h : 높이, w : 높이
          - $\tilde{\pi}$의 첫 번째 채널: $-1 + 2 \cdot F \in [-1, 1]$, 두 번째 채널: $cos\theta$, 세 번째 채널: $sin\theta$



< Motion-I2V Framework>

Method

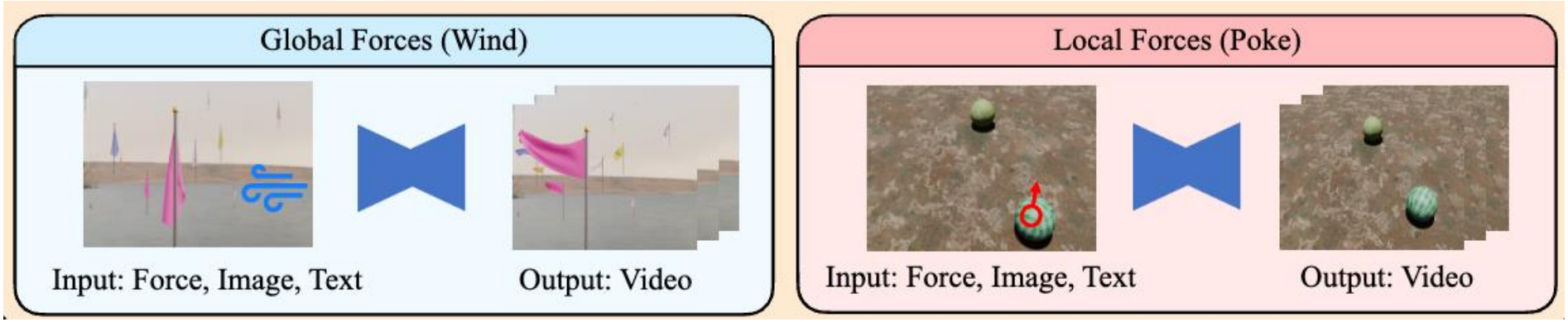
- Local and Global Force Prompts
  - Encoding strategy
    - Local force
      - 점 힘 control 신호는 local한 힘을 지정
      - 힘 크기 $F \in [0,1]$와 각도 $\theta \in [0,360)$에 더해 픽셀 좌표 $(x, y) \in \{0, \dots, w-1\} \times \{0, \dots, h-1\}$로 파라미터화됨



< Motion-I2V Framework>

Method

- Architecture and Training
  - Text와 초기 frame을 condition 입력으로 받는 video generation model인 CogVideoX-5B-I2V를 기반으로 Force prompting 모델 구축
  - Force prompt condition을 통합
    - Physics control prompt π를 입력으로 받는 ControlNet을 추가



< Motion-I2V Framework>

Experiment

- Dataset
  - Pexels에서 수집한 이미지를 사용해 두 force prompting 모델 모두를 평가하기 위한 benchmark datset 제안
    - Prolific을 활용해 2AFC Human study를 수행하여 세 가지 베이스라인과 비교

- Baseline models
  - Text 전용, zero-shot 방식
    - 기존 CogVideoX 모델을 사용하며 의도된 힘을 문자열로 기존 text prompt 끝에 추가
  - Text 전용, fine-tuned 방식
    - Force prompting 모델과 동일한 ControlNet 구조 사용하되, control 신호는 zero-tensor로 설정
    - 학습 시 text prompt 끝에 force 관련 문구 추가
  - Motion Prompting(Geng et al., 2024)
    - 시간적으로 희소한 트랙을 condition 신호로 받는 track-conditioned모델

Experiment

- Human study

Point Force Model	Linear Motion			Oscillatory Motion			Complex Motion		
	Force Adh.	*Real. Physics*	*Visual Qual.*	*Force Adh.*	*Real. Physics*	*Visual Qual.*	*Force Adh.*	*Real. Physics*	*Visual Qual.*
Text-only, zero-shot	72%	50%	48%	67%	48%	52%	73%	48%	49%
Text-only, fine-tuned	79%	53%	52%	62%	52%	58%	74%	55%	54%
Motion Prompting	91%	93%	100%	89%	76%	99%	86%	76%	98%

Global Force Model	Tethered Motion			Aerodynamic Motion			Fluid Dynamics		
	Force Adh.	*Real. Physics*	*Visual Qual.*	*Force Adh.*	*Real. Physics*	*Visual Qual.*	*Force Adh.*	*Real. Physics*	*Visual Qual.*
Text-only, zero-shot	91%	50%	54%	97%	48%	47%	84%	53%	47%
Text-only, fine-tuned	62%	48%	47%	57%	70%	50%	71%	58%	49%
Motion Prompting	93%	82%	100%	90%	75%	100%	90%	80%	95%

- Qualitative results
  - https://force-prompting.github.io/

감사합니다