Physics-Grounded Video Generation



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



Presented By
김소형

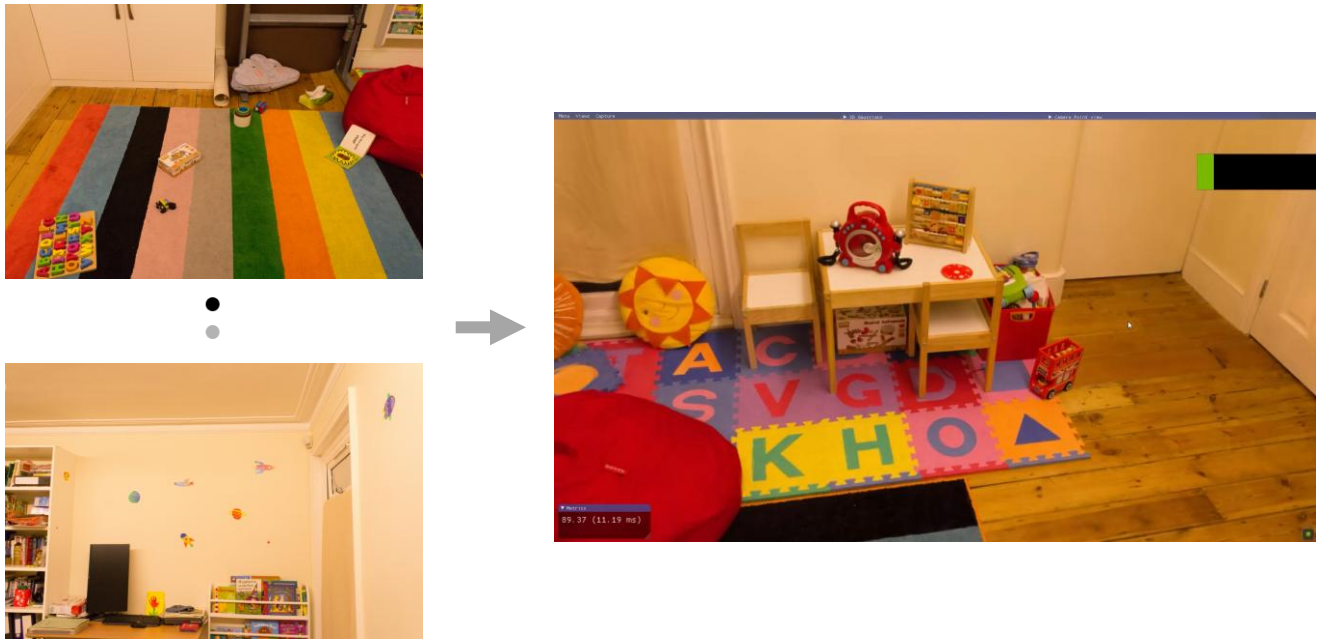
CVSP Lab
Computer Vision & Signal Processing

Outline

- Background
- PhysGen3D: Crafting a Miniature Interactive World from a Single Image [CVPR 2025]
- PhysCtrl: Generative Physics for Controllable andPhysics-Grounded Video Generation [NeurIPS 2025]

Background

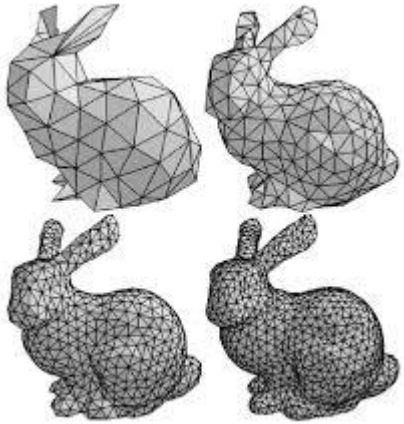
- 3D의 필요성
  - 인공지능이 현실 세계를 이해하기 위해서는 3D data가 필요함



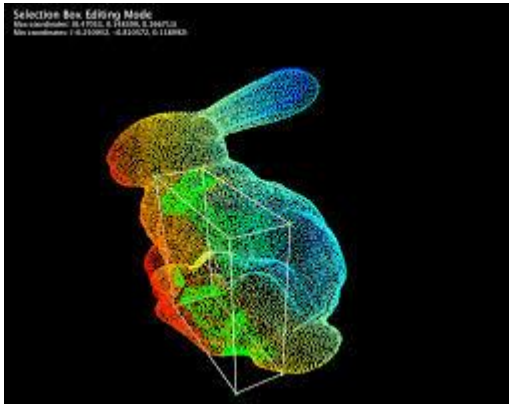
< Input and Output video of 3D Gaussian Splatting >

Background

- 3D Scene Representation
  - Mesh
  - Point Cloud
  - Gaussian Splatting(GS)



< Mesh>



< Point Cloud >

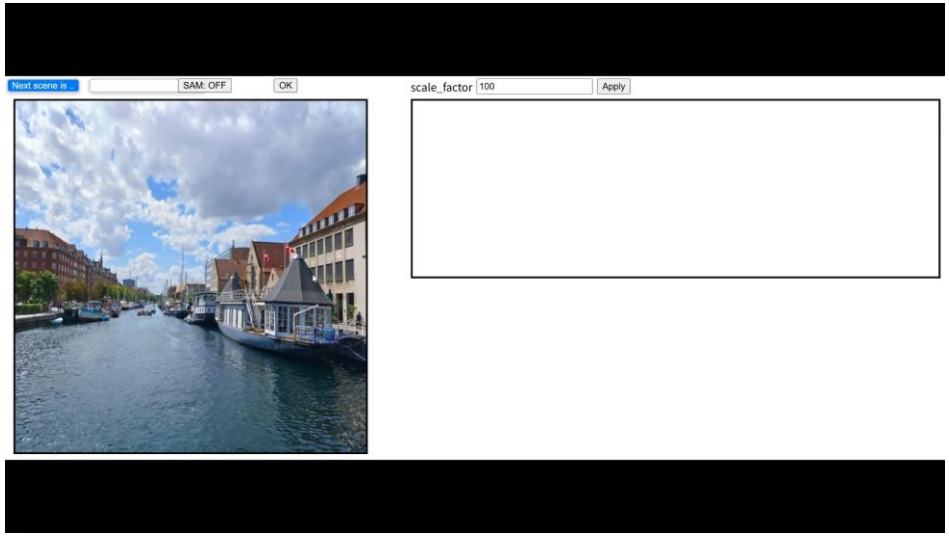


< 3D-GS>

CVSP Lab
Computer Vision & Signal Processing

Background

- 4D scene & Video
  - 4D scene
    - 자유시점
  - Video
    - 고정된 카메라 궤적



< 4D scene >

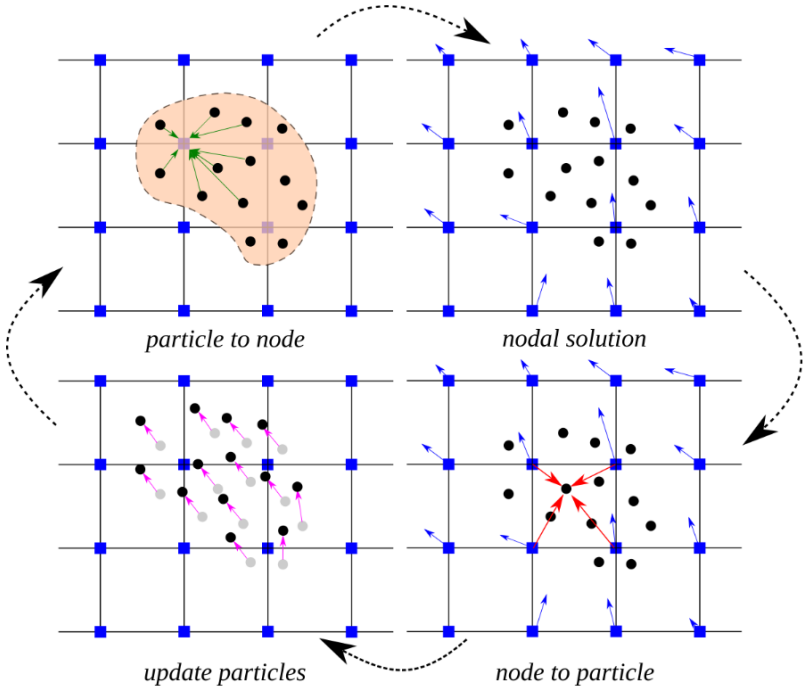


< Video >

CVSP Lab
Computer Vision & Signal Processing

Background

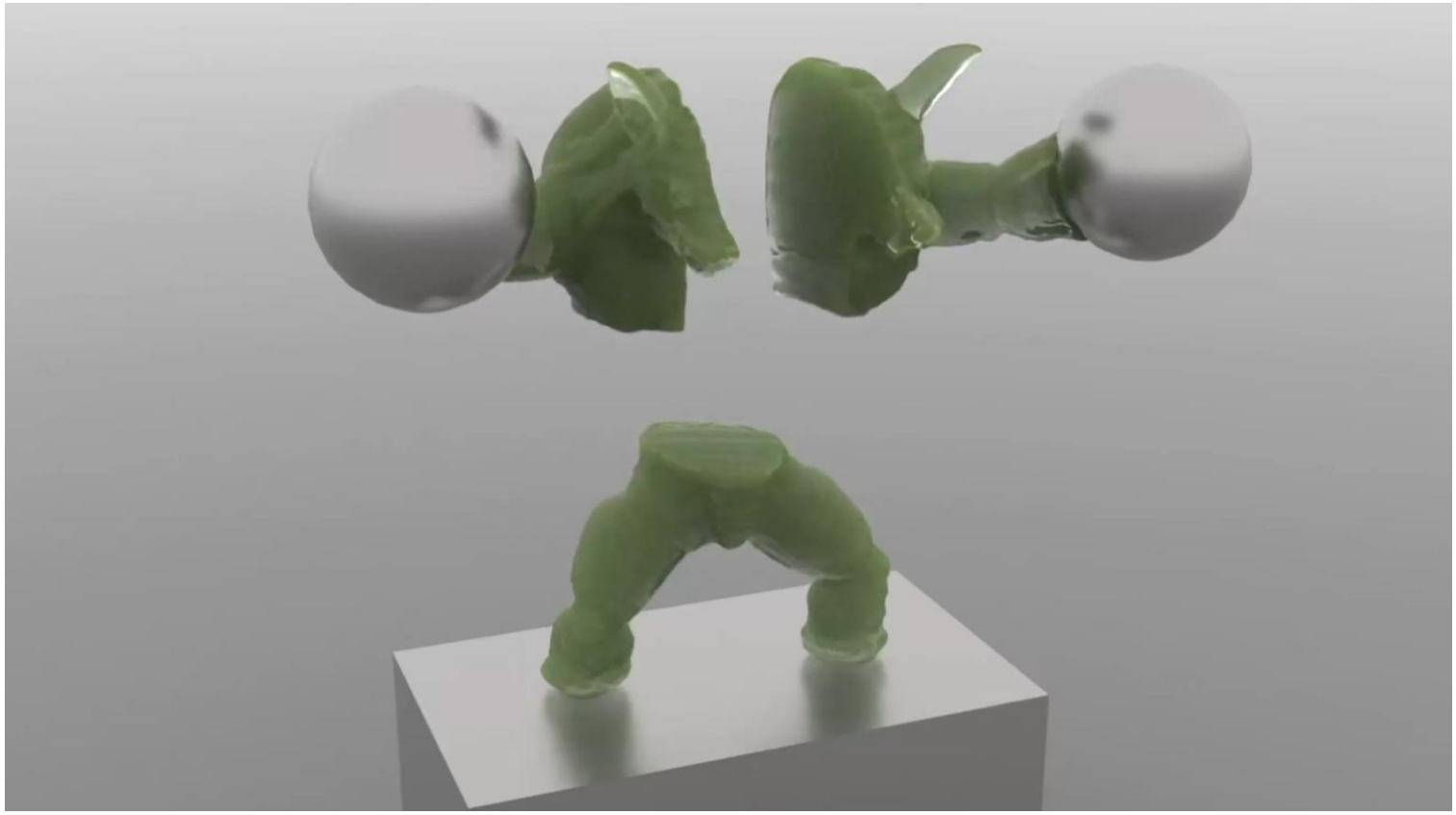
- MPM(Material Point Method)
  - particle to node : 각 입자의 물리적 속성을 grid의 인접 8개 node로 분배
  - nodal solution : 집계된 힘을 이용해 node의 가속도(=힘/질량), 속도를 업데이트
  - nodal to particle : node에서 계산된 가속도/속도를 입자로 전파. node와 입자 간의 거리로 weighted sum
  - update particles : 입자의 속도/가속도를 이용해 새로운 입자의 위치를 갱신



CVSP Lab
Computer Vision & Signal Processing

Background

- MPM(Material Point Method)
  - Visual result



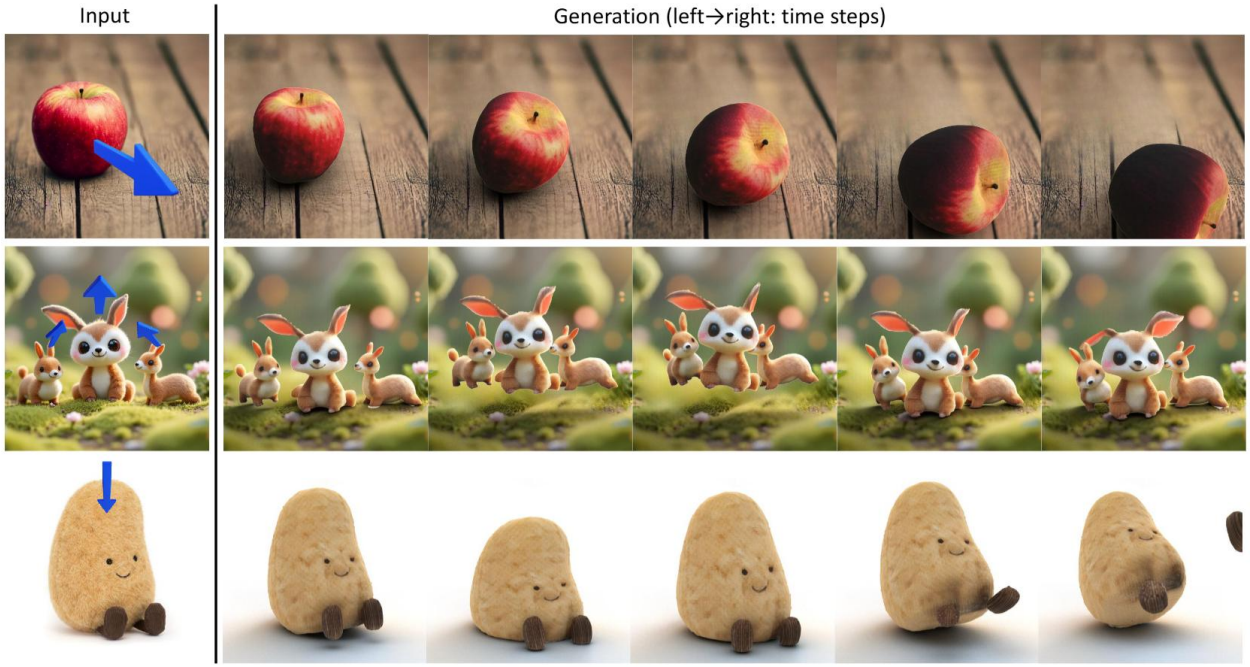
CVSP Lab
Computer Vision & Signal Processing

PhysGen3D: Crafting a Miniature Interactive World from a Single Image

[CVPR 2025]

PhysGen3D

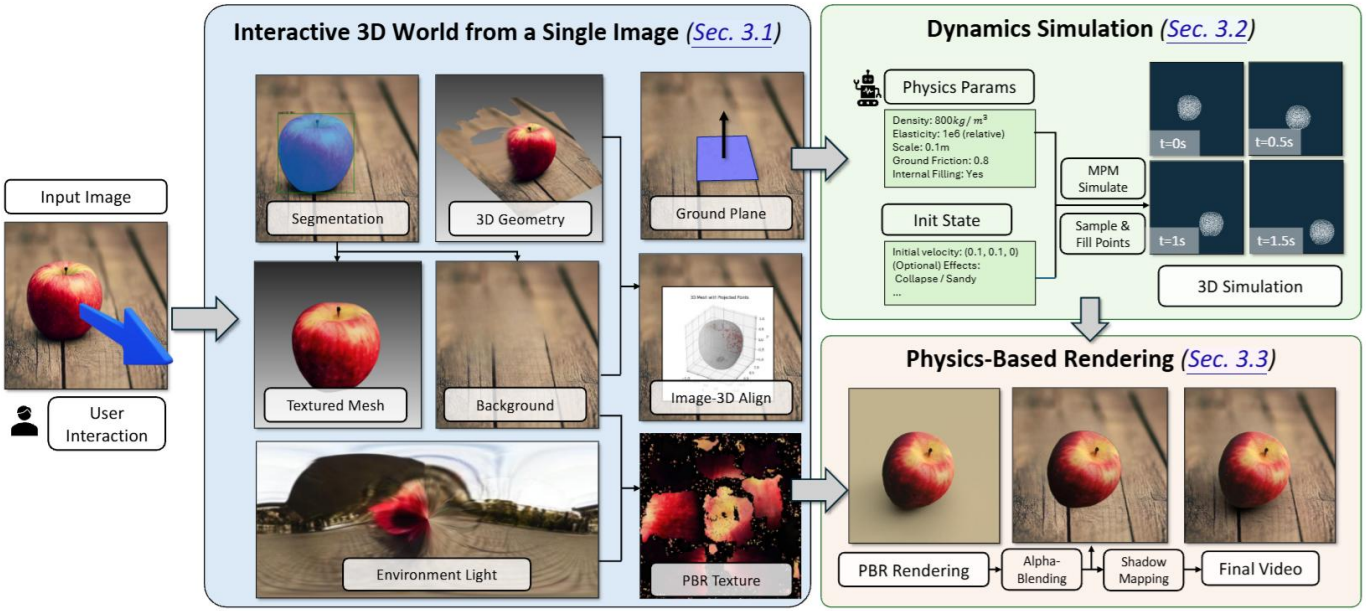
- Introduction
  - Contribution
    - Single image으로부터 interactive 3D scene 생성
    - Image-based geometric과 physics-based simulation의 강점을 결합



< PhysGen3D 적용 예시 >

PhysGen3D

- Method
  - Overview
    - 입력이미지를 가지고 3D Scene 표현
    - Dynamic Simulation에서 객체의 움직임을 생성
    - 물리 기반 렌더링을 통해 최종 비디오 출력



< PhysGen3D의 전체 프레임워크 >

CVSP Lab
Computer Vision & Signal Processing

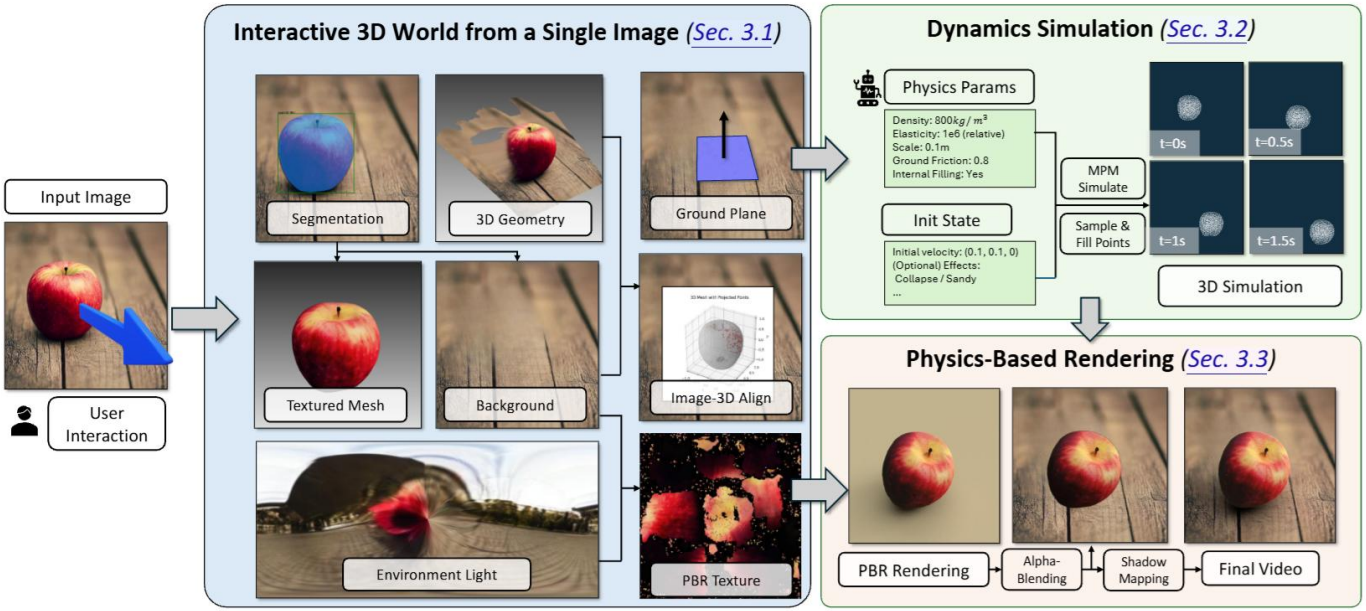
PhysGen3D

[1] Ren, Tianhe, et al. "Grounded sam: Assembling open-world models for diverse visual tasks." arXiv preprint arXiv:2401.14159 (2024).

- Method
  - Interactive 3D World from a Single Image – **Segmentation**
    - GPT-4o로 foreground object의 카테고리를 식별 후 프롬프트 생성
    - Grounded-SAM[1]: object o^i를 segment함

$$\{o^i \in R^{W \times H \times 3}\}_i^N$$

✓ N: object 개수



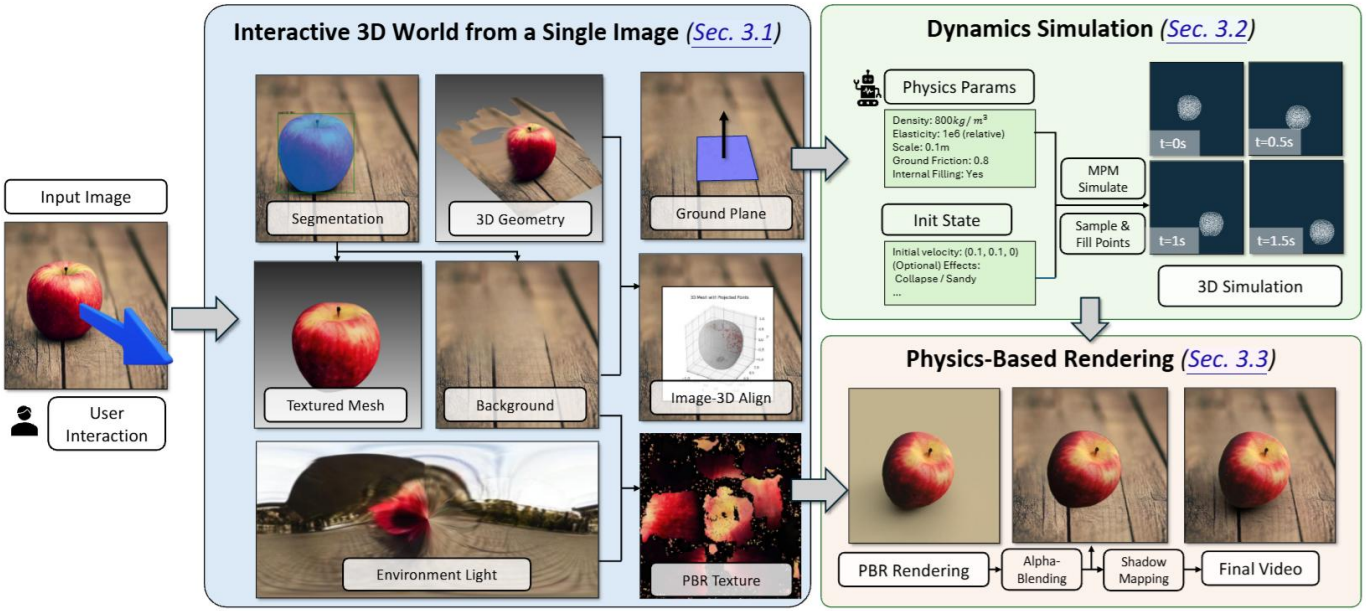
< PhysGen3D의 전체 프레임워크 >

CVSP Lab Computer Vision & Signal Processing

PhysGen3D

[2] Xu, Jiale, et al. "Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models." arXiv preprint arXiv:2404.07191 (2024).

- Method
  - Interactive 3D World from a Single Image – **Mesh Generation**
    - InstantMesh[2]: segment된 object에 대하여 multi-view 이미지 합성
    - 이 이미지들로 object 3D mesh $\mathcal{O}$ 생성



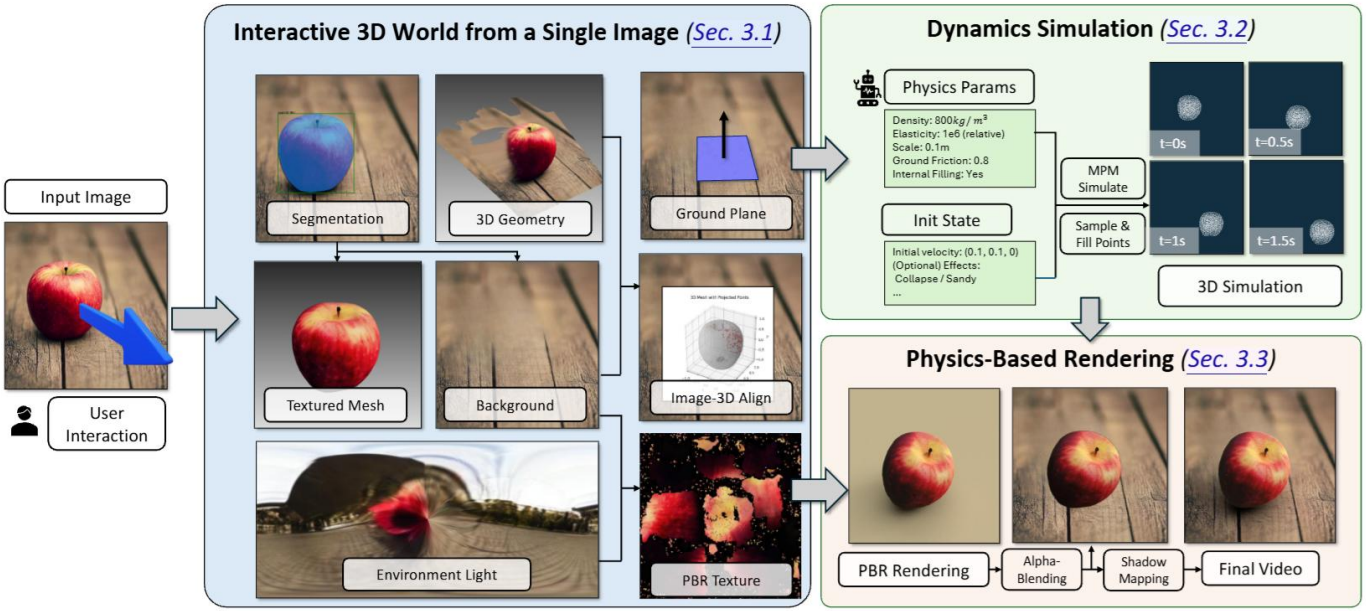
< PhysGen3D의 전체 프레임워크 >

CVSP Lab
Computer Vision & Signal Processing

PhysGen3D

[3] Wang, Shuzhe, et al. "Dust3r: Geometric 3d vision made easy." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
[4] Cao, Xu, et al. "Bilateral normal integration." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.
[5] Suvorov, Roman, et al. "Resolution-robust large mask inpainting with fourier convolutions." Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2022.

- Method
  - Interactive 3D World from a Single Image – **Background Handling**
    - Simulation
      - Dust3r [3]: 이미지 depth추정 및 3D point cloud $\mathcal{P}$ 생성
      - Bilateral Normal Integration [4]: 3D point cloud Smooth surface $\mathcal{S}$ 생성
    - Rendering
      - LaMA inpainting model [5]: object로 인해 가려진 부분을 채워넣음

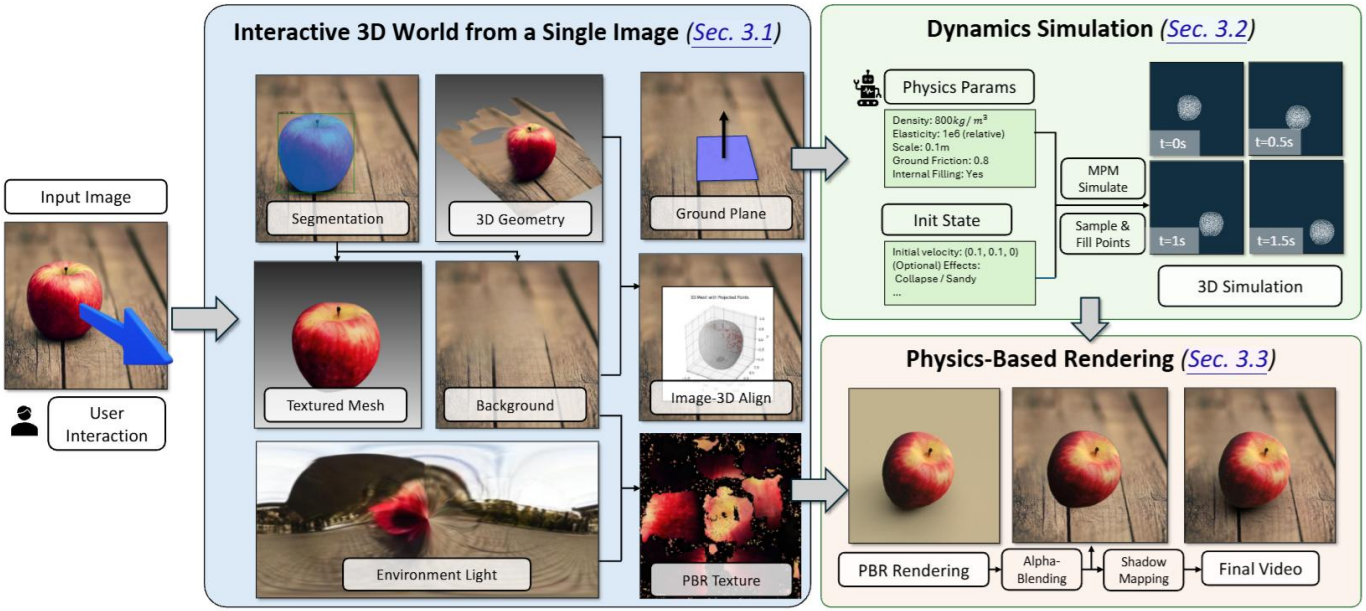


< PhysGen3D의 전체 프레임워크 >

CVSP Lab
Computer Vision & Signal Processing

[3] Wang, Shuzhe, et al. "Dust3r: Geometric 3d vision made easy." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
[4] Cao, Xu, et al. "Bilateral normal integration." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.
[5] Suvorov, Roman, et al. "Resolution-robust large mask inpainting with fourier convolutions." Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2022.

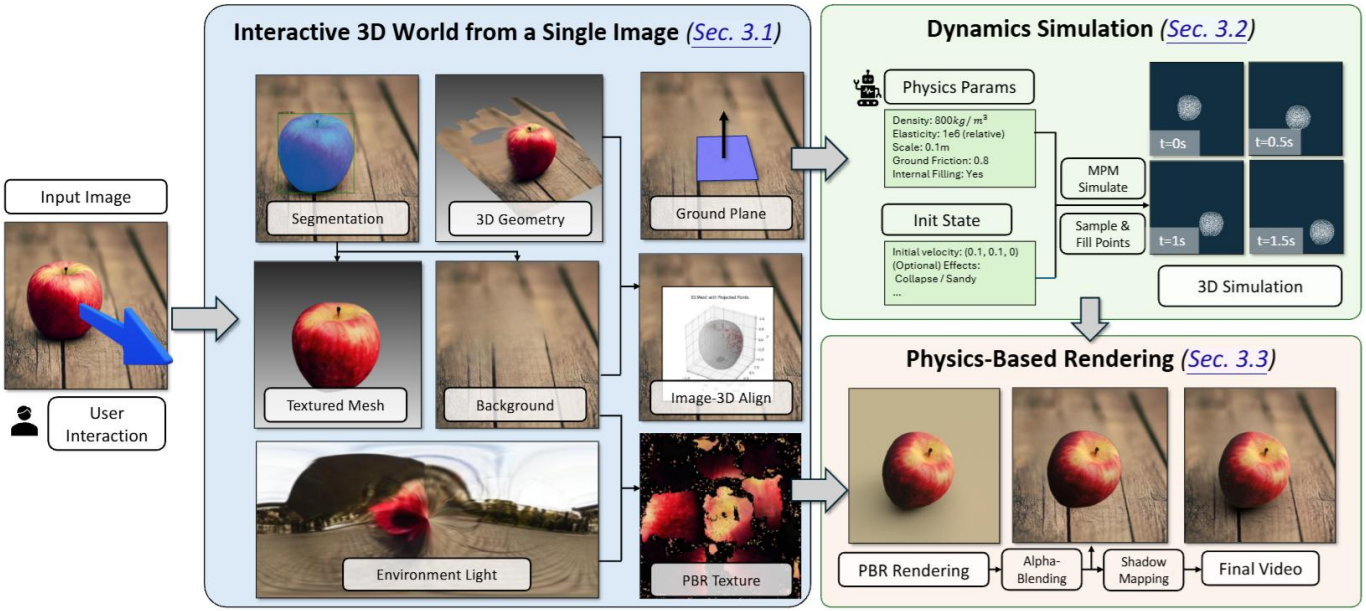
- Method
  - Interactive 3D World from a Single Image – **Object Pose and Scale Estimation**
    - 3D mesh $\mathcal{O}$는 object-centric coordinate system임
      - camera coordinate에서의 위치와 스케일을 모름
    - 3D Scene $\mathcal{P}$ 위에 3D mesh $\mathcal{O}$를 올바른 스케일과 6DoF pose로 배치해야 함
    - 이를 위해 Multi-stage coarse-to-fine alignment 전략 사용



< PhysGen3D의 전체 프레임워크 >

[6] Sarlin, Paul-Edouard, et al. "Superglue: Learning feature matching with graph neural networks." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
[7] Fischler, Martin A., and Robert C. Bolles. "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography." *Communications of the ACM* 24.6 (1981): 381-395.

- Method
  - Interactive 3D World from a Single Image – **Object Pose and Scale Estimation**
    - Coarse alignment stage
      - SuperGlue [6]: Feature level에서 original object image의 카메라 뷰를 알아냄
      - PnP algorithm [7]: 3D mesh $\mathcal{O}$의 6DoF pose 추정
      - $\sum_i^N = \left\| P_i^N - P_i'^N \right\|_2$: Scale ambiguity 완화
        - ✓ P_i^N: 3D feature point in camera coordinate, P'^N_i: 3D feature point in object coordinate



< PhysGen3D의 전체 프레임워크 >

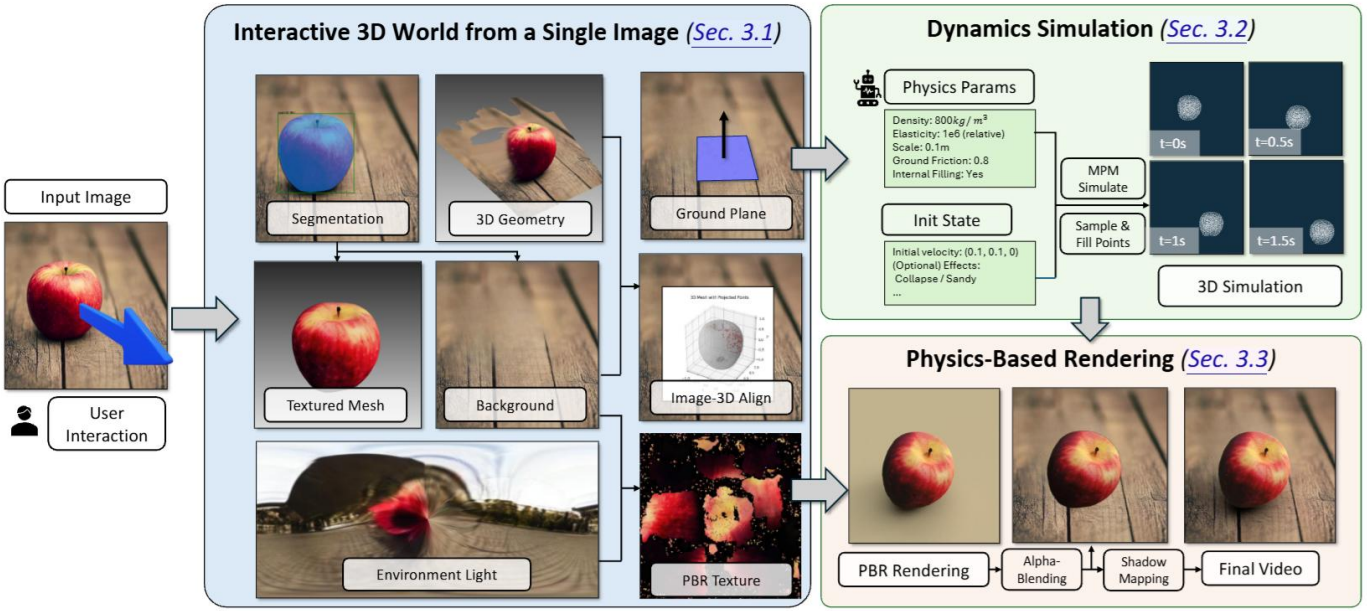
Chen, Boyuan, et al. "Physgen3d: Crafting a miniature interactive world from a single image." *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025.

- Method
  - Interactive 3D World from a Single Image – **Object Pose and Scale Estimation**
    - Fine alignment stage

$\mathcal{L}_{dice} = 1 - \frac{2\times|M_A \cap M_B|}{|M_A|+|M_B|}$: 얼마나 겹치는지를 의미

$\mathcal{L}_{depth} = \frac{mse(M_B * z_A - M_B * z_B)}{|M_B|}$: object 영역에서 depth 차이를 줄이기 위함

✓ M_A: 렌더링 된 마스크, M_B: SAM의 예측마스크, z_A: 렌더링 된 depth, z_B: Dust3r의 예측 depth

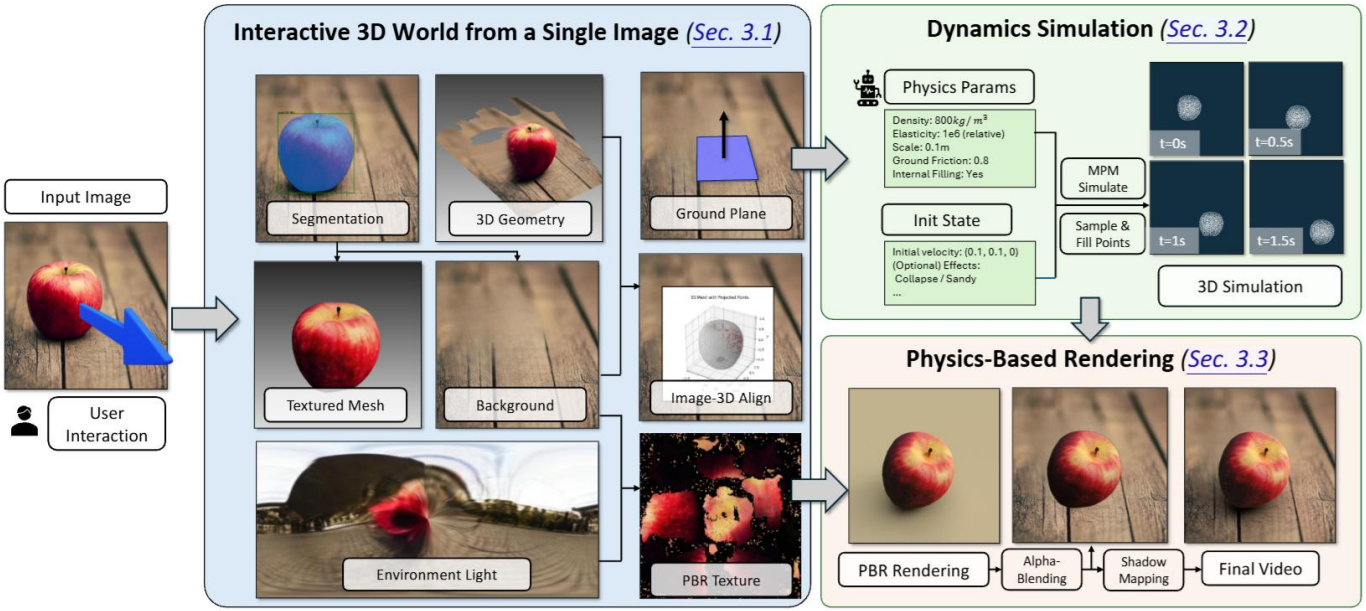


< PhysGen3D의 전체 프레임워크 >

PhysGen3D

[8] Jakob, Wenzel, et al. "Dr. jit: A just-in-time compiler for differentiable rendering." ACM Transactions on Graphics (TOG) 41.4 (2022): 1-19.
[9] Phongthawee, Pakkapon, et al. "Diffusionlight: Light probes for free by painting a chrome ball." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024.

- Method
  - Interactive 3D World from a Single Image – **Appearance Optimization**
    - Mitsuba3 [8]: 렌더링 된 mesh에서 물질 성질 추정
    - DiffusionLight [9]: 입력이미지로부터 Lighting 추정
    - 안 보이는 쪽의 texture를 단순화하기 위해 uniform roughness & metallic value 가정



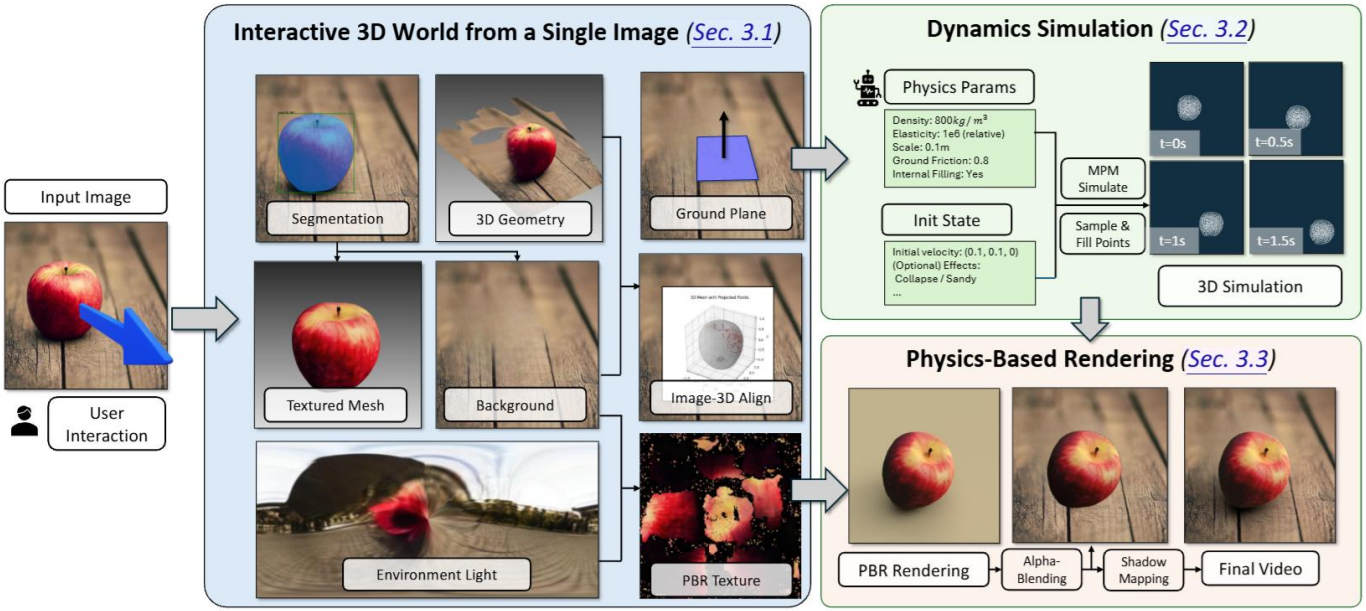
< PhysGen3D의 전체 프레임워크 >

CVSP Lab
Computer Vision & Signal Processing

PhysGen3D

[3] Wang, Shuzhe, et al. "Dust3r: Geometric 3d vision made easy." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

- Method
  - Interactive 3D World from a Single Image – **Physics Reasoning**
    - GPT4o로 physics parameter(탄성력, 밀도, 마찰계수 등) 추정
    - Dust3r [3]로 추정한 depth의 부정확성을 scaling factor k로 보정



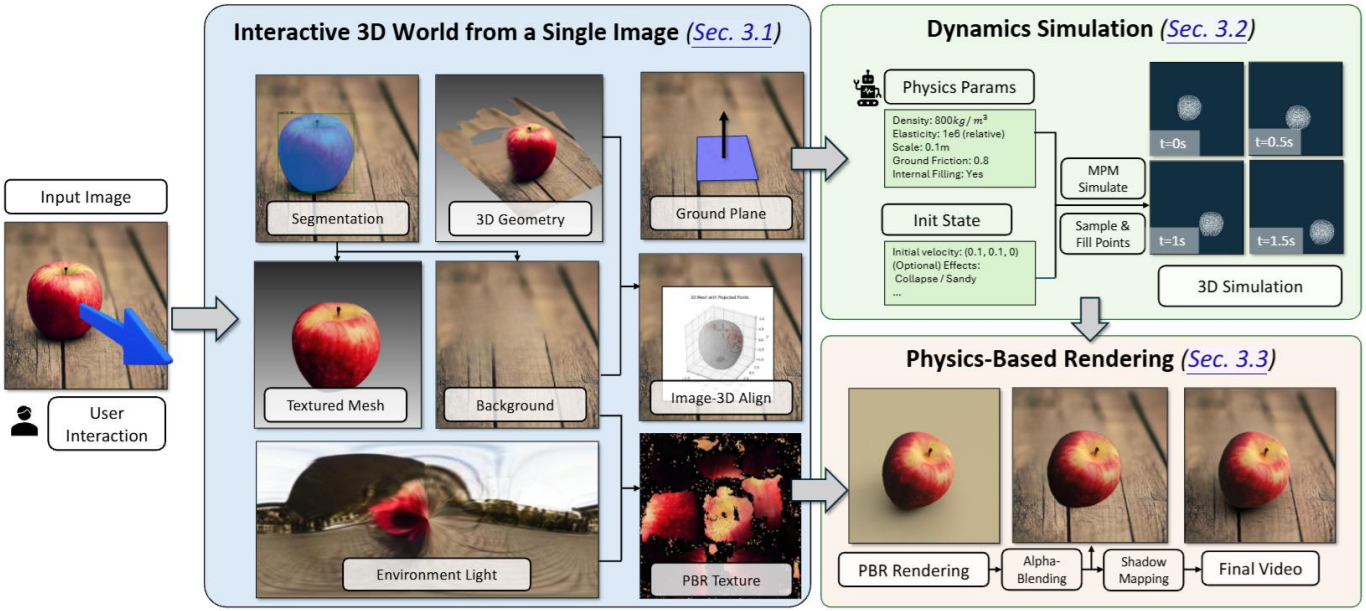
< PhysGen3D의 전체 프레임워크 >

CVSP Lab
Computer Vision & Signal Processing

PhysGen3D

[10] Hu, Yuanming, et al. "Taichi: a language for high-performance computation on spatially sparse data structures." ACM Transactions on Graphics (TOG) 38.6 (2019): 1-16.

- Method
  - Dynamics Simulation
    - Internal filling: 3D mesh 내부를 particle로 채움
    - Taichi Elements [10]: 앞서 얻은 physics parameter와 scaling factor k로 trajectory 생성



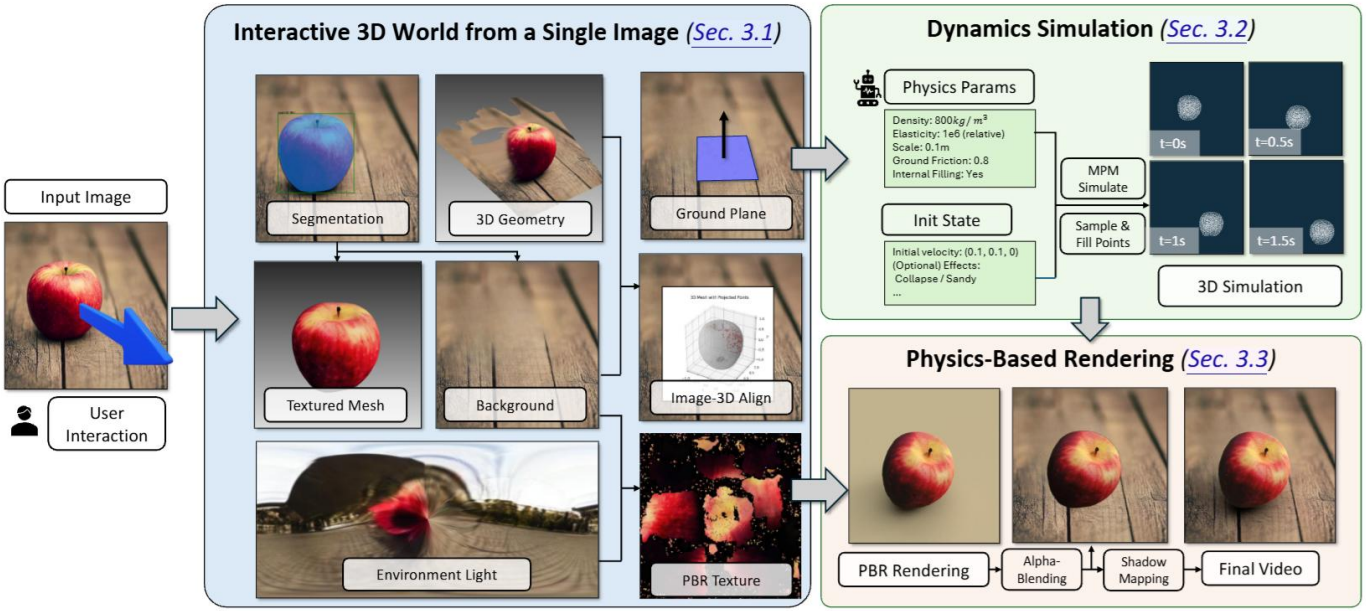
< PhysGen3D의 전체 프레임워크 >

CVSP Lab
Computer Vision & Signal Processing

PhysGen3D

[8] Jakob, Wenzel, et al. "Dr. jit: A just-in-time compiler for differentiable rendering." ACM Transactions on Graphics (TOG) 41.4 (2022): 1-19.

- Method
  - Physics-Based Rendering
    - Mitsuba3 [8]: 물질 성질과 lighting정보를 가지고 렌더링
    - 렌더링된 object와 shadow를 inpainted background 위에 합성하여 최종 영상 생성



< PhysGen3D의 전체 프레임워크 >

CVSP Lab
Computer Vision & Signal Processing

PhysGen3D

- Experiment
  - Quantitative Evaluation
    - Human Evaluation & GPT-4o Evaluation
      - PhysReal: 물리 법칙 준수 여부, 탄성(Elasticity) 및 마찰(Friction) 등 실제 물리적 특성 재현력 측정
      - PhotoReal: 전반적인 시각 품질, 아티팩트(왜곡) 유무, 빛/그림자/질감/재료의 정확한 묘사 수준
      - Align: 생성된 비디오 내용이 의도된 텍스트 프롬프트의 의미와 얼마나 잘 일치하는가
    - Vbench Score
      - Motion: 프레임 간 전환이 자연스럽고 끊김이 없는지 측정
      - Imaging: 개별 프레임의 선명도와 화질 측정

Methods	PhysReal	PhotoReal	Align
Kling 1.0 [3]	2.811	3.566	2.467
Runway Gen-3 [2]	2.283	**3.582**	1.886
Pika 1.5 [1]	2.412	3.314	2.016
Ours	**3.707**	3.411	**3.866**

< Human Evaluation Results >

Methods	Motion↑	Imaging↑	PhysReal↑	PhotoReal↑	Align↑
Kling 1.0	**0.996**	0.671	0.563	0.874	0.596
Runway Gen-3	0.991	**0.723**	0.141	**0.896**	0.144
Pika 1.5	0.994	0.671	0.544	0.863	0.563
MOFA-Video	0.994	0.634	0.384	0.764	0.304
DragAnything	0.985	0.428	0.645	0.756	0.380
Ours	0.995	0.666	0.752	0.867	**0.796**
Ours+VEnhancer	0.994	0.677	**0.766**	0.880	0.745

< VBench Scores and GPT-4o Evaluation Results>

CVSP Lab
Computer Vision & Signal Processing

PhysGen3D

Chen, Boyuan, et al. "Physgen3d: Crafting a miniature interactive world from a single image." *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025.

- Experiment
  - Qualitative Evaluation





< Video Results >

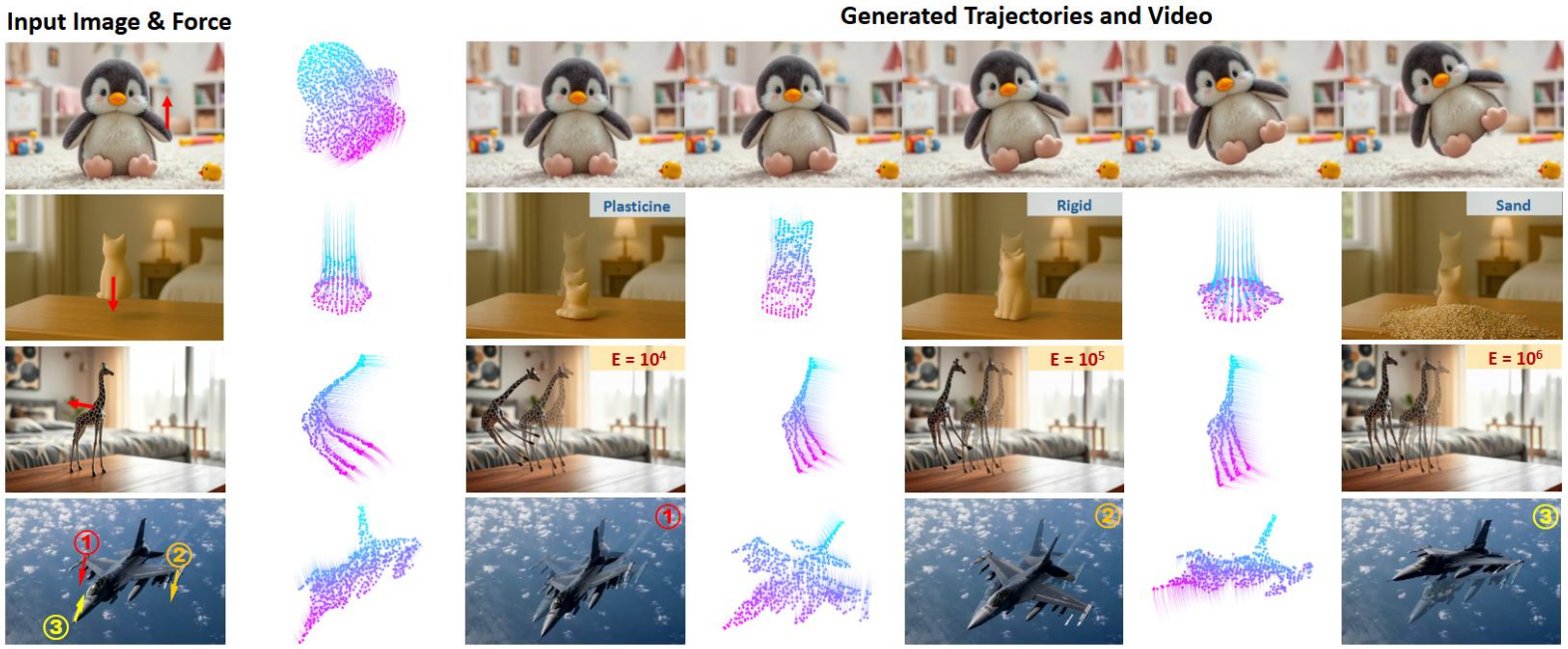
PhysCtrl: Generative Physics for Controllable andPhysics-Grounded Video Generation

[NeurIPS 2025]

PhysCtrl

Wang, Chen, et al. "Physctrl: Generative physics for controllable and physics-grounded video generation." *arXiv preprint arXiv:2509.20358* (2025).

- Introduction
  - Contribution
    - 디퓨전 모델 기반 point trajectory 생성 모델 프레임워크를 제안
    - 사용자가 임의의로 external force와 physics parameter를 부여하여 비디오를 생성할 수 있음

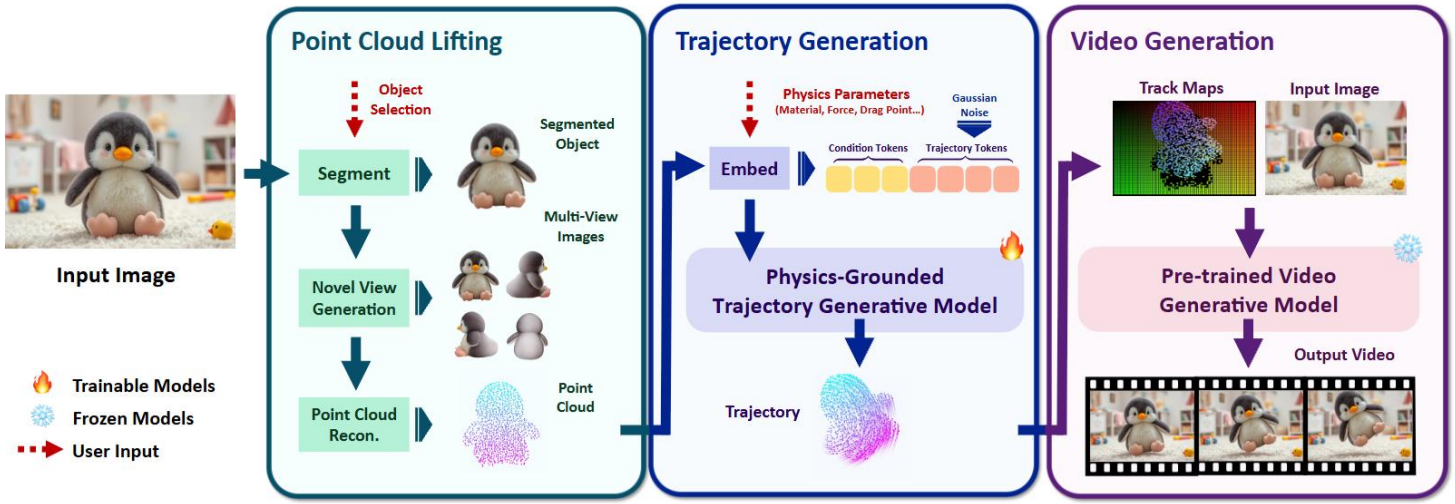


< PhyCtrl 적용 예시 >

CVSP Lab
Computer Vision & Signal Processing

PhysCtrl

- Method
  - Overview
    - 입력이미지로부터 3D point cloud reconstruction
    - Point cloud의 위치 정보 및 physics parameter를 바탕으로 Trajectory model 학습
    - Trajectory model의 Output track maps와 입력이미지로 비디오 생성



< PhyCtrl 전체 프레임워크 >

PhysCtrl

Wang, Chen, et al. "Physctrl: Generative physics for controllable and physics-grounded video generation." *arXiv preprint arXiv:2509.20358* (2025).

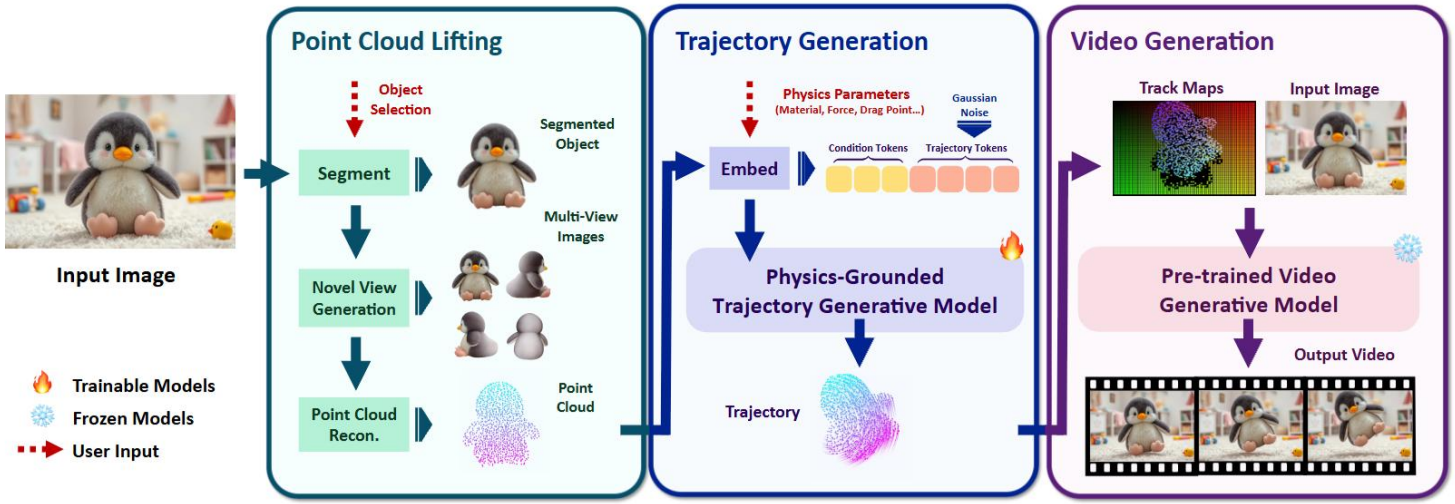
- Method
  - Problem Setting
    - 주어진 object를 N개의 3D point로 표현

$$\boldsymbol{P}_0 = \{x_i^0 \in R^3\}_{i=1}^{N}$$

✓ x_i^0: 3D point 개별 입자

✓ i : 3D point index

✓ $\boldsymbol{P}_0$: t = 0에서의 초기 Point cloud



< PhyCtrl 전체 프레임워크 >

CVSP Lab Computer Vision & Signal Processing

PhysCtrl

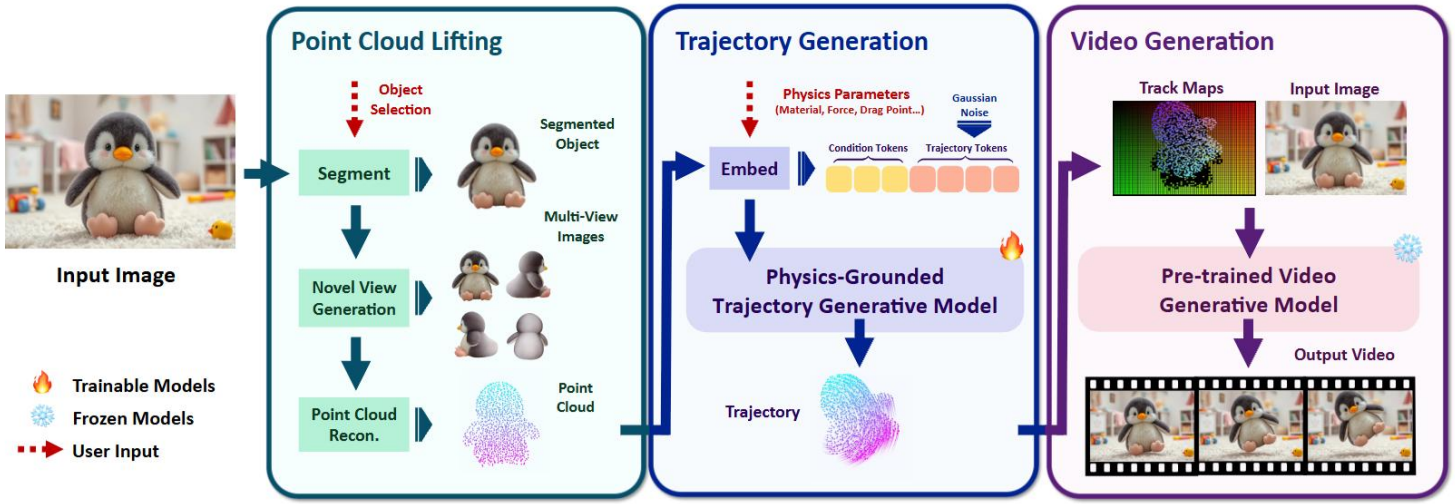
Wang, Chen, et al. "Physctrl: Generative physics for controllable and physics-grounded video generation." *arXiv preprint arXiv:2509.20358* (2025).

- Method
  - Problem Setting
    - PhysCtrl은 physics condition **c** 에서 $\mathcal{P}$를 잘 예측하는 것을 목표로 함

$$\mathcal{P} = \{\boldsymbol{P}^f\}_{f=1}^{F} = \left\{\left\{x_p^f\right\}_{p=1}^{N}\right\}_{f=1}^{F}$$

$$\mathrm{c} = \{\, \boldsymbol{P}_0, \mathbf{f}, \mathbf{D}, \{E, \nu\}, h, [mat] \,\}$$

f : 외력
D : input drag
E : Young's modulus (탄성계수)
ν : Poisson's ratio
h : height
$[mat]$: material type

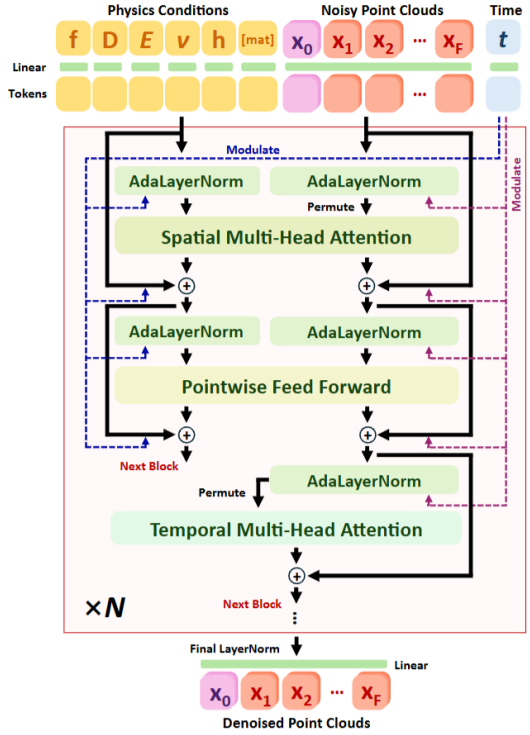


< PhysCtrl 전체 프레임워크 >

CVSP Lab
Computer Vision & Signal Processing

Wang, Chen, et al. "Physctrl: Generative physics for controllable and physics-grounded video generation." *arXiv preprint arXiv:2509.20358* (2025).

- Method
  - Physics-grounded Trajectory Generative model
    - Noisy point cloud를 MLP 를 통해 point token으로 임베딩
    - Physics condition **c**는 별도의 MLP로 physics token으로 임베딩



< Trajectory generation architecture >

CVSP Lab
Computer Vision & Signal Processing

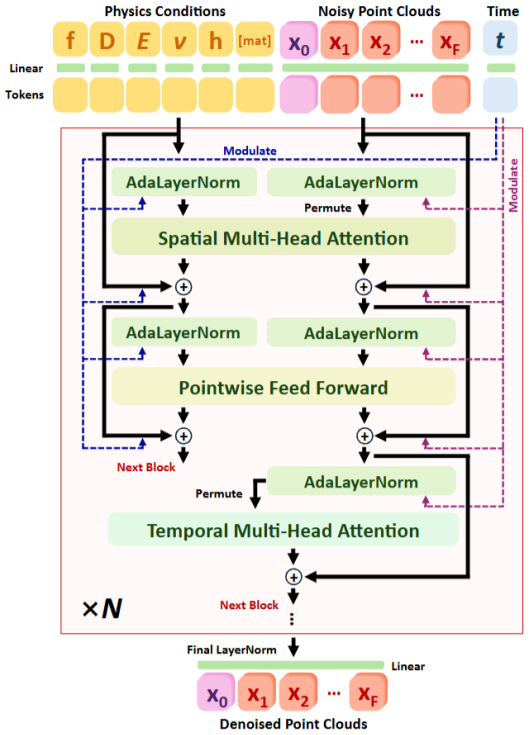
PhysCtrl

- Method
  - Physics-grounded Trajectory Generative model
    - Spatial attention과 Temporal attention으로 분리
      - Naïve 4D attention보다 계산 비용을 줄임

✓ $\hat{\boldsymbol{P}}^f = SelfAttn\left(AdaLN\left(\left[\boldsymbol{P}^f; cond\right]\right)\right)$

✓ $\hat{\boldsymbol{T}}_p = SelfAttn(AdaLN([\boldsymbol{T}_p]))$

- $cond = MLP_{phys}([\mathbf{f}; \mathbf{D}, ; \{E, \nu\}, h, [mat]\,])$

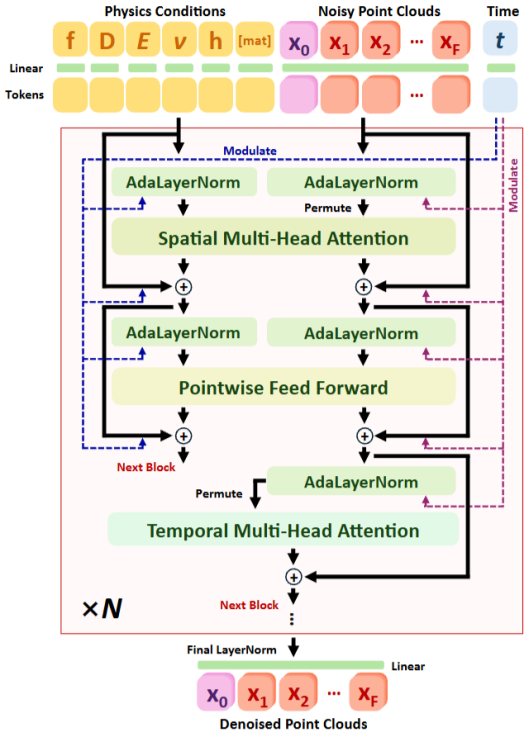


< Trajectory generation architecture >

CVSP Lab
Computer Vision & Signal Processing

Wang, Chen, et al. "Physctrl: Generative physics for controllable and physics-grounded video generation." *arXiv preprint arXiv:2509.20358* (2025).

- Method
  - Training Losses
    - $\mathcal{L} = \mathcal{L}_{diff} + \lambda_{vel}\mathcal{L}_{vel} + \lambda_{phys}\mathcal{L}_{phys} + \lambda_{floor}\mathcal{L}_{floor}$ 를 최소화
      - $\mathcal{L}_{diff}$: point cloud trajectory를 복원
      - $\mathcal{L}_{vel}$: 물리적으로 자연스러운 움직임 유도
      - $\mathcal{L}_{phys}$: 물리 법칙에 가까운 움직임을 만들기 위함
      - $\mathcal{L}_{floor}$: 바닥 아래로 penetration 방지

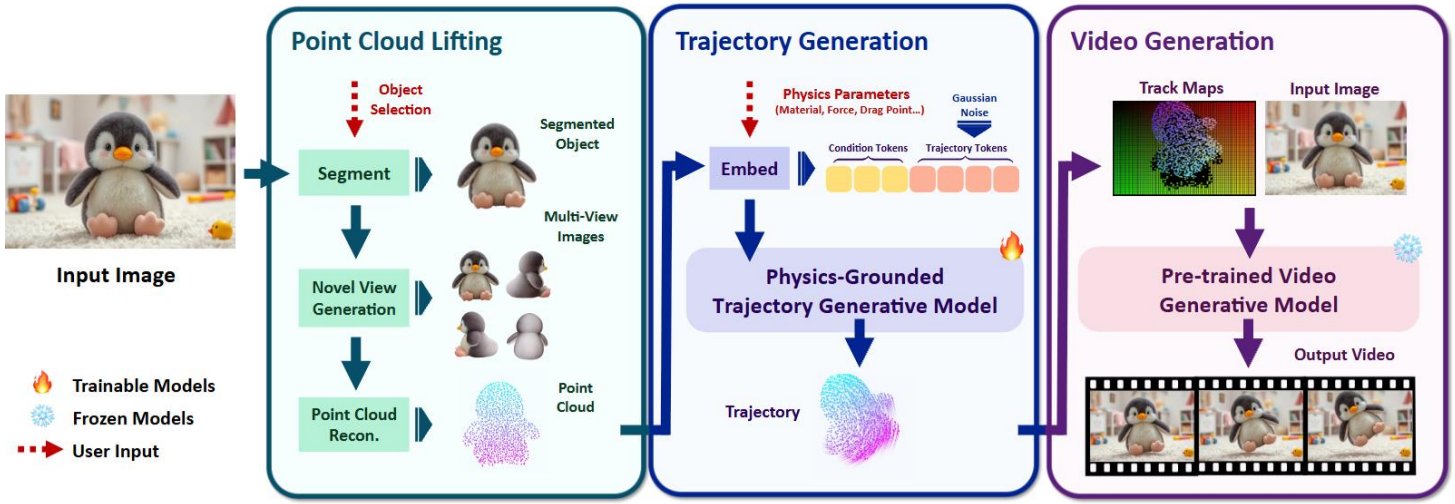


< Trajectory generation architecture >

PhysCtrl

[11] Wang, Jianyuan, et al. "Vggt: Visual geometry grounded transformer." *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025.

- Method
  - Physics-grounded Image-to-Video Generation
    - 입력이미지로부터 object segmentation과 novel view 이미지 생성
    - Novel view 이미지와 segmented 이미지로 3D asset 생성 후 point cloud 추출
    - 3D scene reconstruction model인 VGGT [11]로 3D scene 복원
    - VGGT [11]의 coordinate system과 obeject point의 align을 맞춤

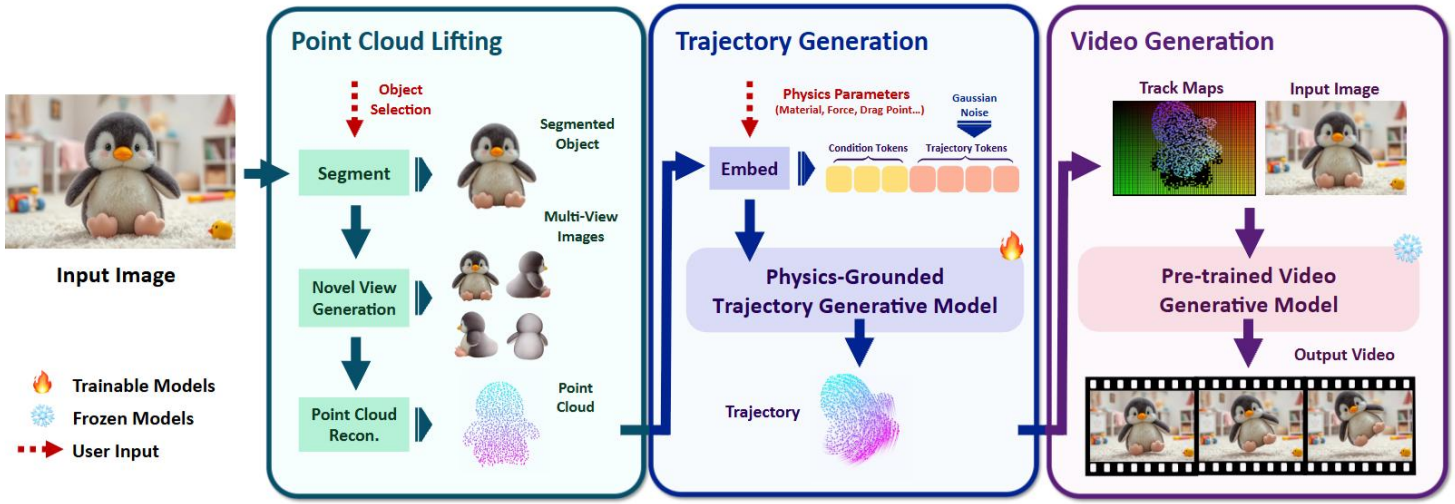


< PhyCtrl 전체 프레임워크 >

CVSP Lab
Computer Vision & Signal Processing

PhysCtrl

- Method
  - Physics-grounded Image-to-Video Generation
    - Trajectory Generative Model로 object point의 trajectory를 생성
    - 생성된 3D point trajectory를 image space로 projection시킴
    - Projection된 track maps을 비디오 생성 모델에 넣어 Output video 생성



< PhyCtrl 전체 프레임워크 >

PhysCtrl

- Experiments
  - Quantitative Evaluation
    - Video evaluation
      - SA: 텍스트 프롬프트와 비디오의 내용/동작이 얼마나 일치하는가
      - PC: 가해진 힘의 방향과 위치에 따라 물체의 움직임이 직관적이고 물리적으로 타당하게 구현되었는가
      - VQ: 비디오의 전반적인 시각적 품질 및 프레임 간의 시간적 일관성(Temporal quality) 평가
    - Trajectory generation comparison
      - vIoU: 예측된 포인트 클라우드와 정답(Ground Truth) 사이의 중첩 정도를 측정
      - CD: 두 포인트 클라우드 간의 쌍별 근접 이웃 거리의 평균을 측정. (전반적인 형태의 유사도 평가)
      - Corr: 대응되는 두 포인트 클라우드 사이의 유클리드 거리(Euclidean distance)를 측정

Table 1: Results of video evaluation.

	SA↑	PC↑	VQ↑
DragAnything [89]	2.9	2.8	2.8
ObjCtrl [87]	1.5	1.3	1.4
Wan2.1 [82]	3.8	3.7	3.6
CogVideoX [94]	3.2	3.2	3.1
Ours	**4.5**	**4.5**	**4.3**

Table 2: Quantitative comparison on trajectory generation.

Method	vIoU↑	CD↓	Corr↓
M2V [6]	24.92%	0.2160	0.1064
MDM [79]	53.78%	0.0159	0.0240
Ours	**77.59%**	**0.0028**	**0.0015**

CVSP Lab
Computer Vision & Signal Processing

PhysCtrl

- Experiments
  - Qualitative Evaluation



< Output video>

논문 선택 및 발표 이유

- ECCV 2026
  - 유체뿐만 아니라 다양한 물질 유형들에 motion을 부여하기 위해 Physics simulator에 관한 논문을 survey함

- 향후 연구 계획
  - Video generation 연구에 대한 기초를 쌓기 위함

Thank you

CVSP Lab
Computer Vision & Signal Processing