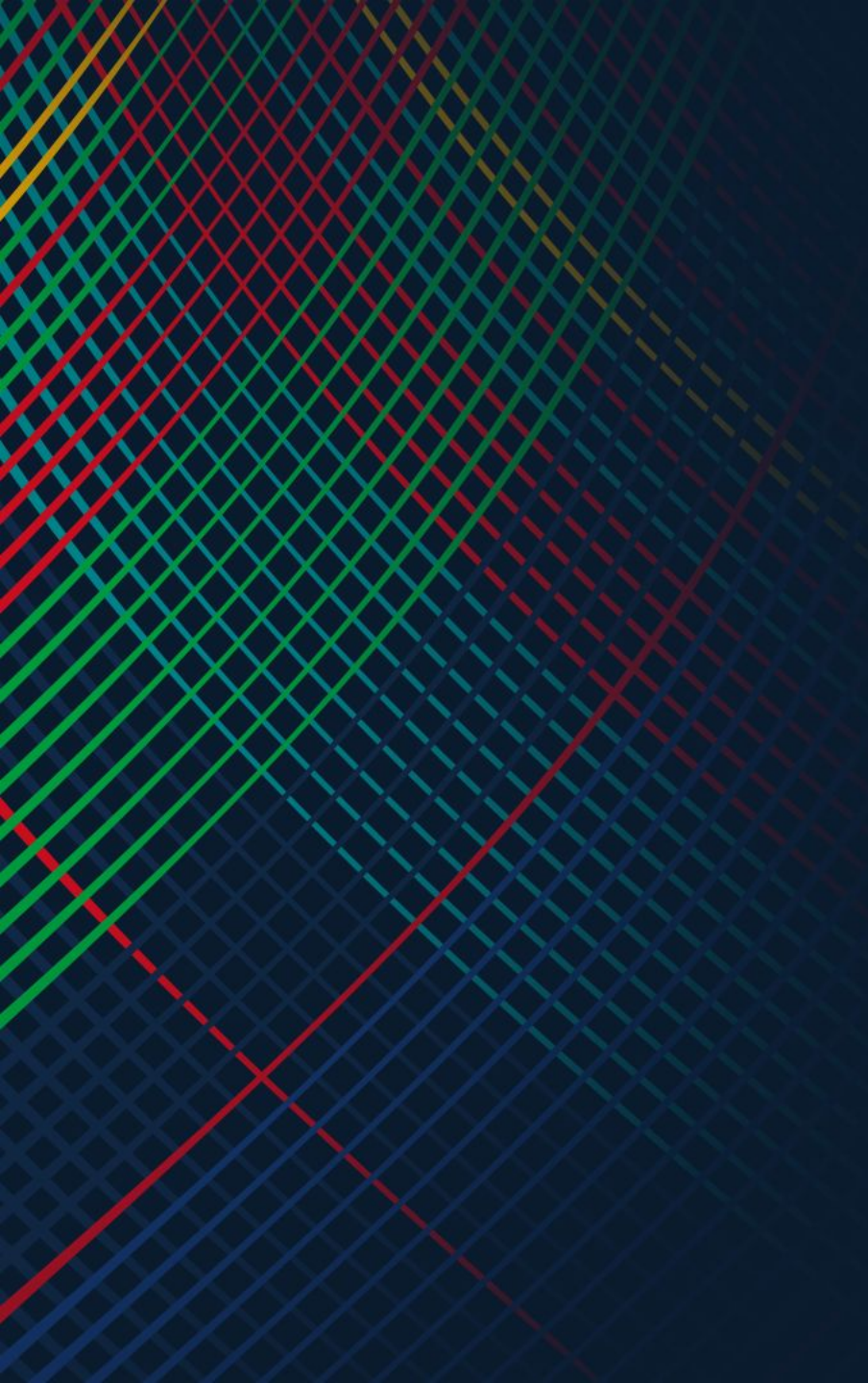
Research Talk

Adithya Narayan

Human Sensing Lab,
Carnegie Mellon University



About me

My Journey

2020	2021	2024	2025	2025
Research intern @ IIT-Bombay	**Bachelor's Thesis @ NUS**	**Applied researcher @ Klothed**	**Graduate research @ CMU**	**Research Intern @ HeyGen**
Monocular, visual odometry.	Fetal Medicine	Virtual try-ons	3D-VQA, adversarial scene exploration	Video Diffusion Models

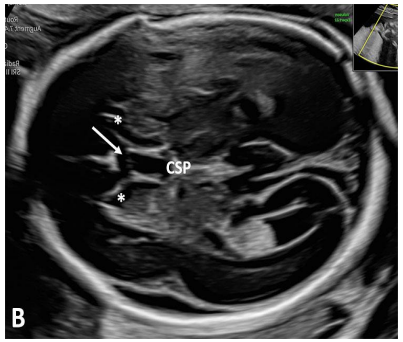
My Research

Video Diffusion Models

How can we effectively model character-centric camera motion generation?

3D Understanding

Like humans, can we improve the 3D understanding in 2D-VLMs using video diffusion prios?



Fetal Anomaly Detection

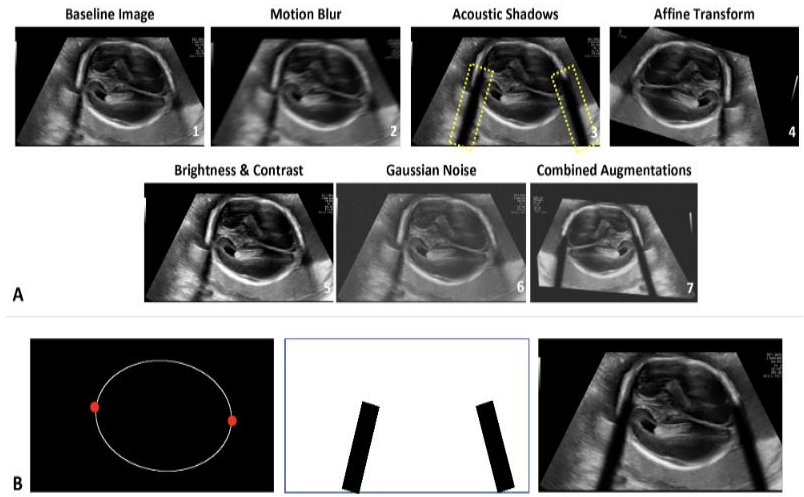
How can we effectively segment, measure and detect fetal brain anomalies during the second trimester of pregnancy?

Previous Research

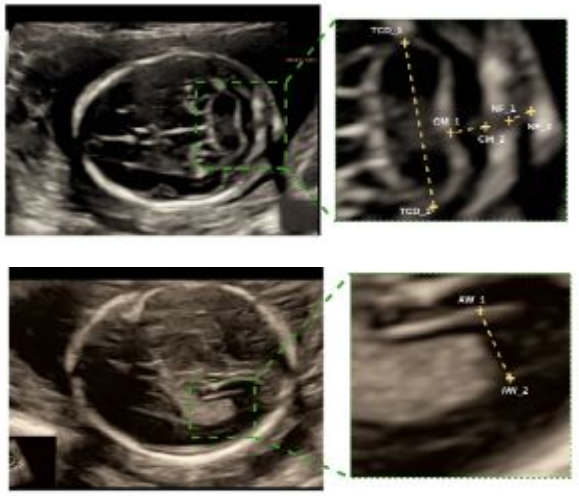
Fetal Anomaly Detection

Can we detect fetal anomalies during the second trimester?
Can we generalize across ultrasound machines?

Synthetic Data Gen [1]

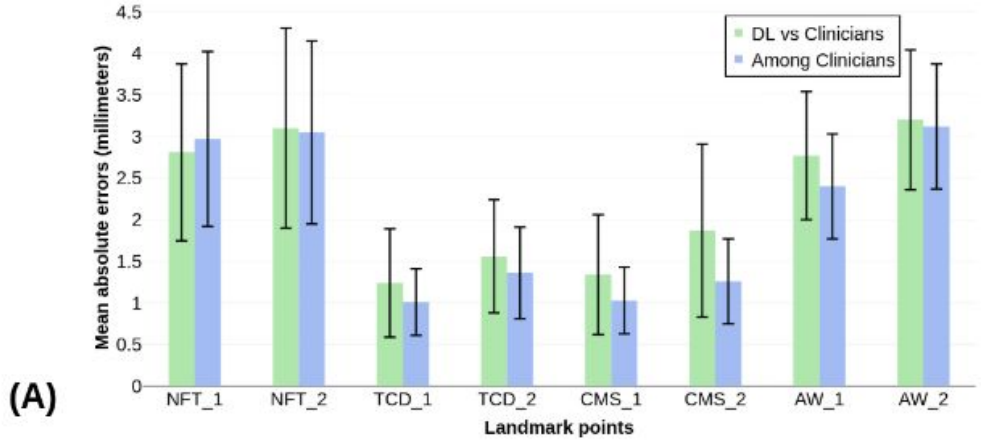


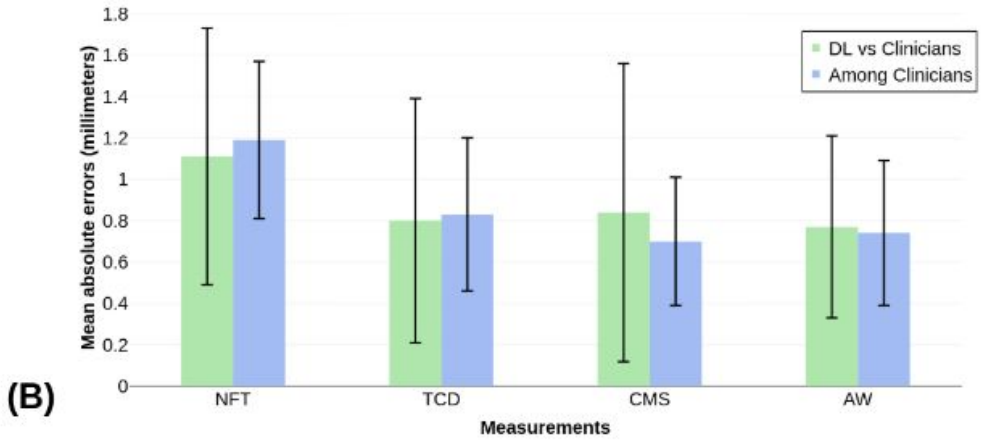
Anatomy measurement [2]



[1] Lad et. al. SPIE Medical Imaging 2022: Computer Aided Diagnosis (12033-75), 2022; [2] Shankar et. al. arXiv, 2022

Clinician Approved

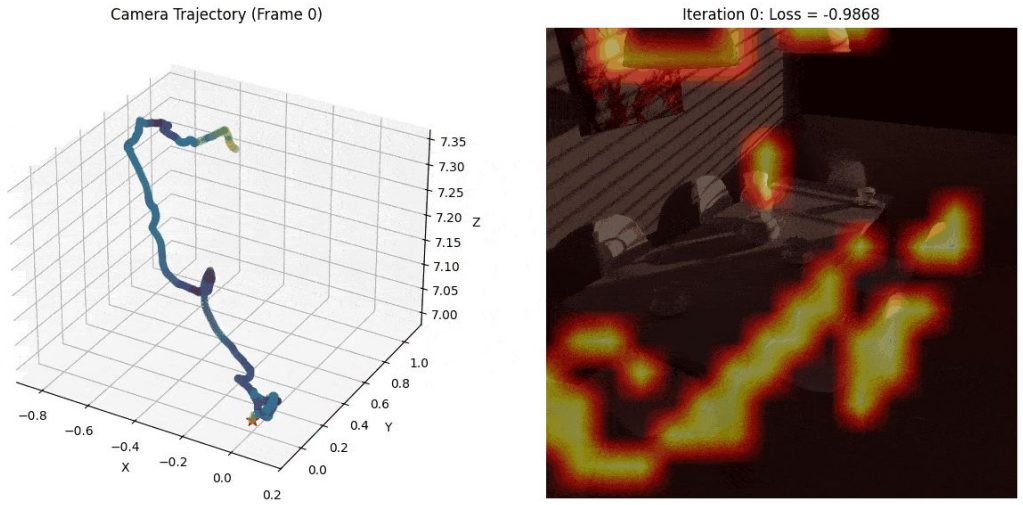




- Presented some clinical validations at SPIE, ISBI, and also ISUOG
- **Startup** based on this research (**Original Medical Research Labs**) raised **~$10 million** on the basis of this research!

Current Research

Current Research

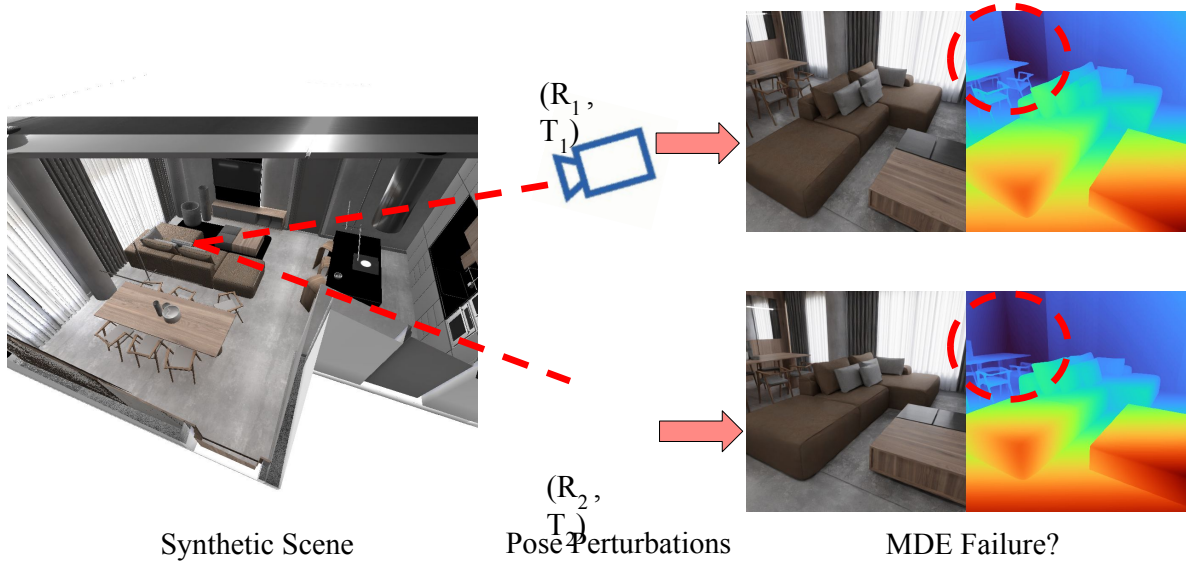


Adversarial 3D Scene Exploration

Character Centric Camera Motion Control for VDMs

Adversarial Scene Exploration

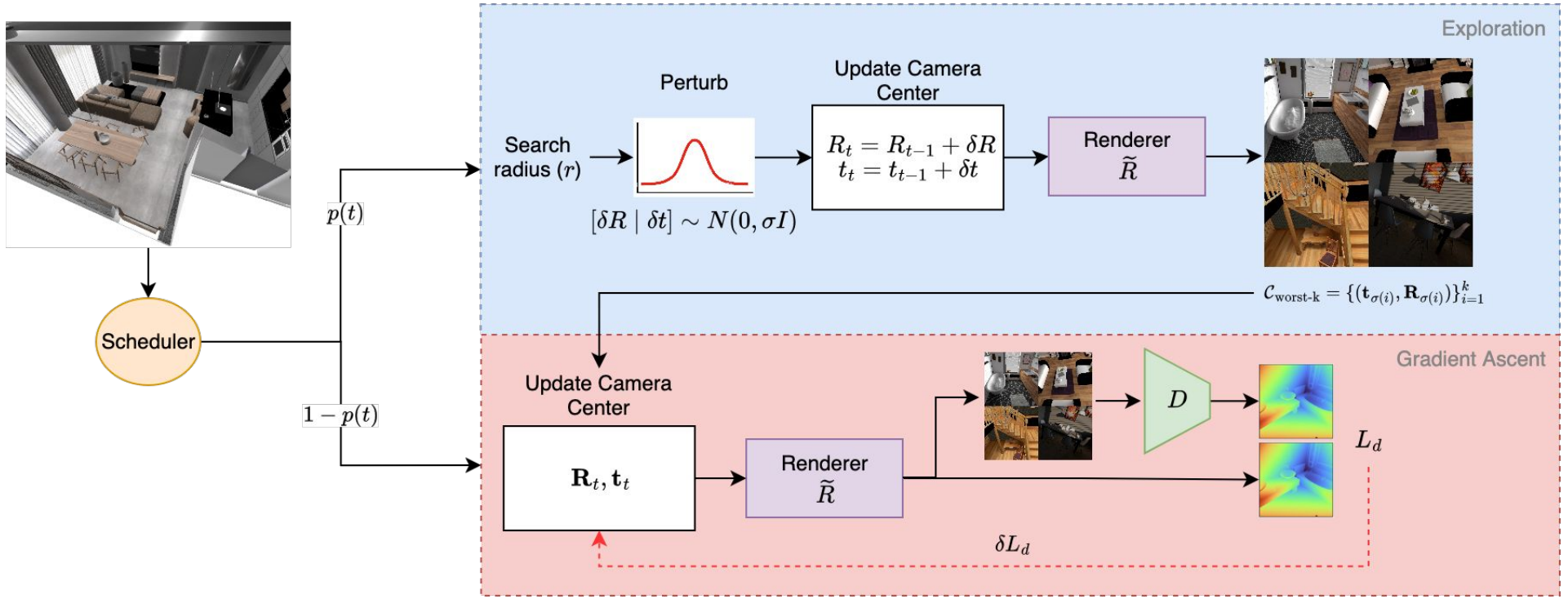
Debugging SOTA vision models is a tedious, large-scale task.



How can we evaluate these models in an automated, scalable way?

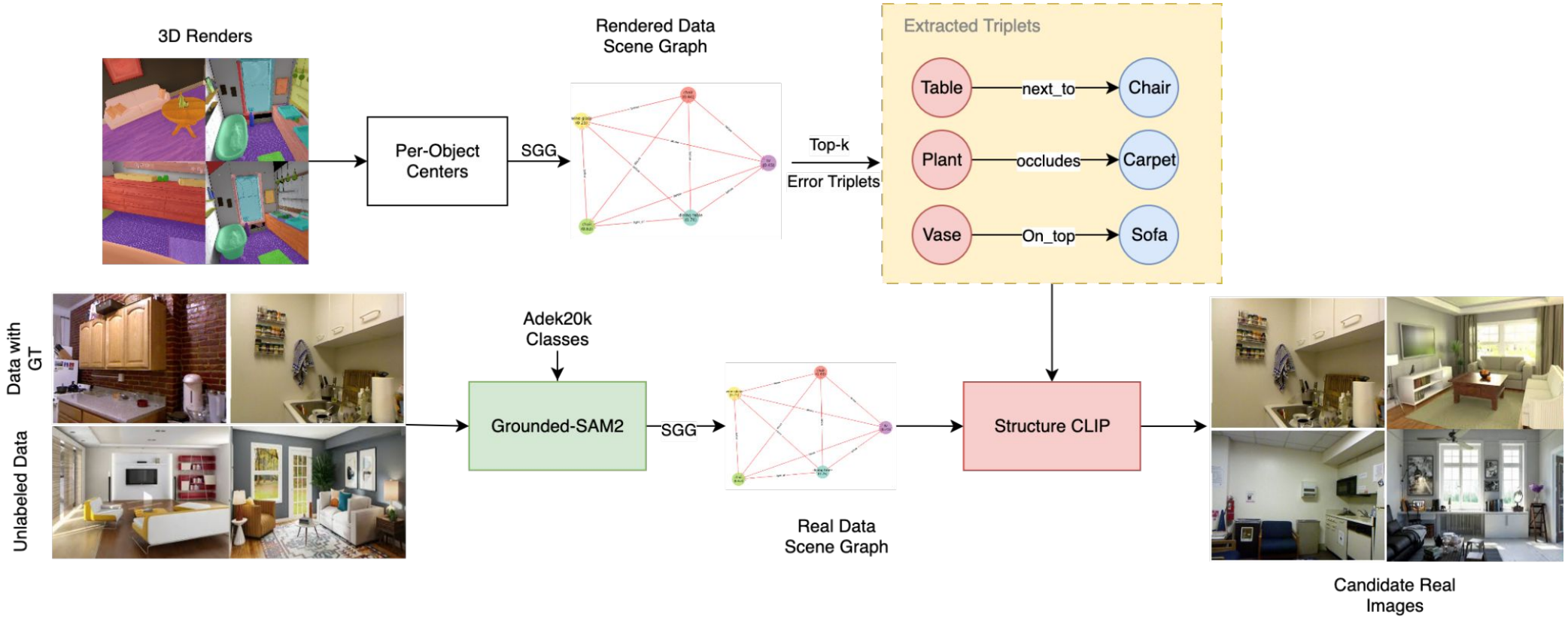
Adversarial Scene Exploration

- Scene Exploration - Explore the 3D scene to find random poses
- Gradient Ascent - Optimize in SE(3) to find local failures

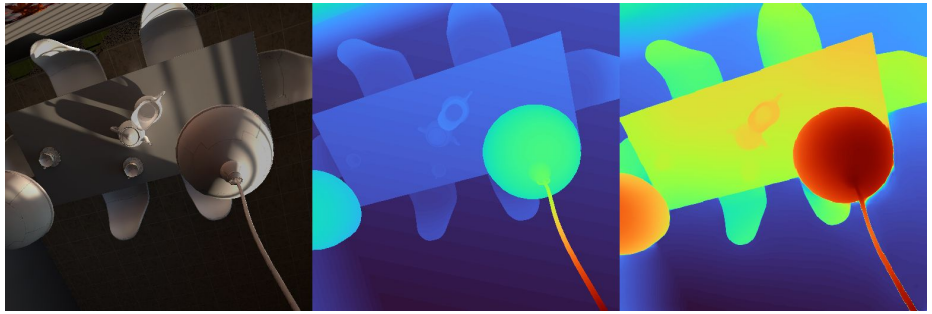


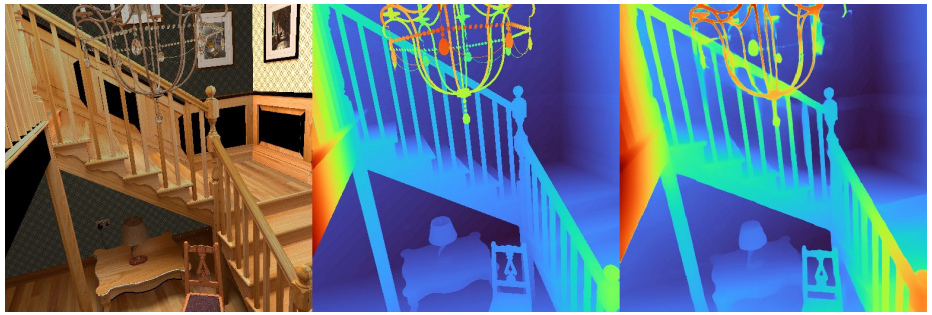
Adversarial Scene Exploration

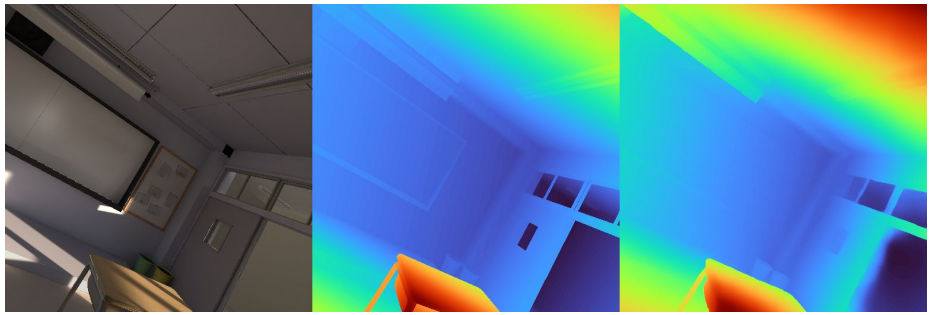
- Real data attribution is also key!



Some Results





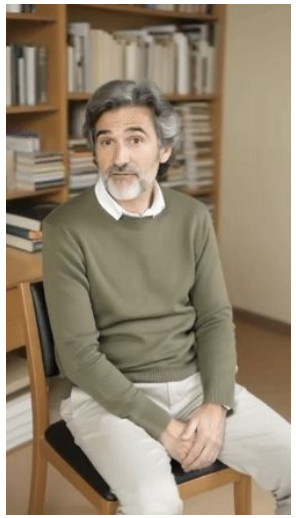


Character-Centric Camera Motion Control For VDMs

Carnegie Mellon University

What are Video Diffusion Models?

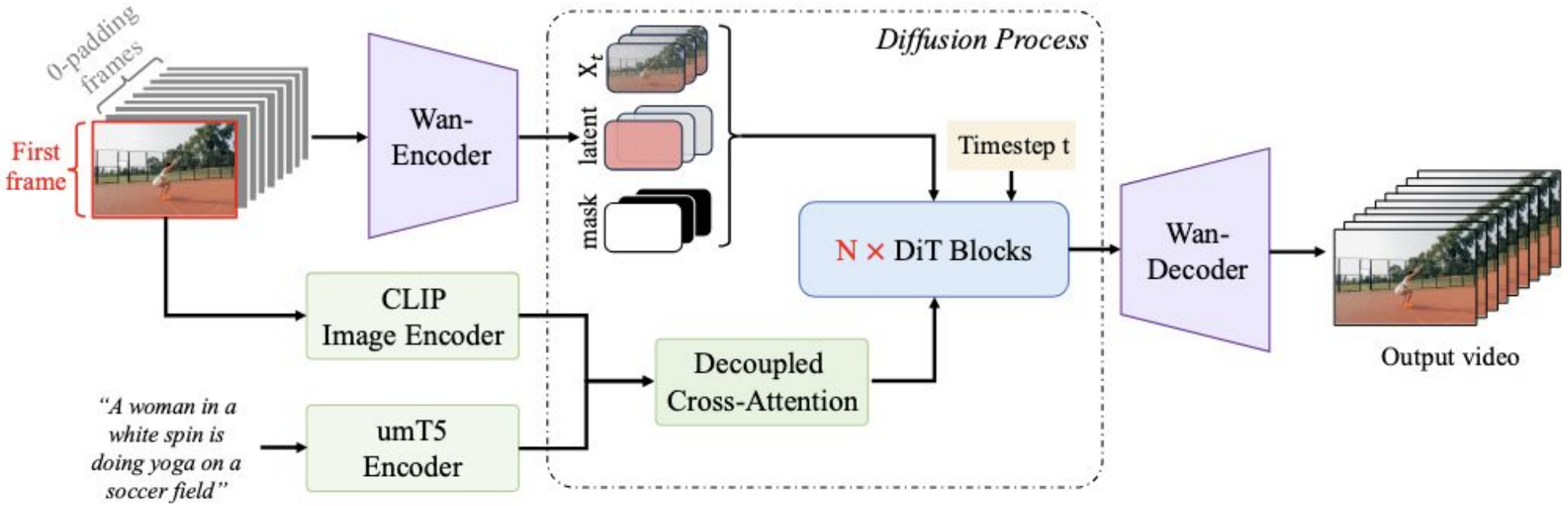
Image (or other modality)



Prompt

The person is talking passionately about the sea.

A simple overview



Modern Video Models

Sora2

Veo 3

Wan

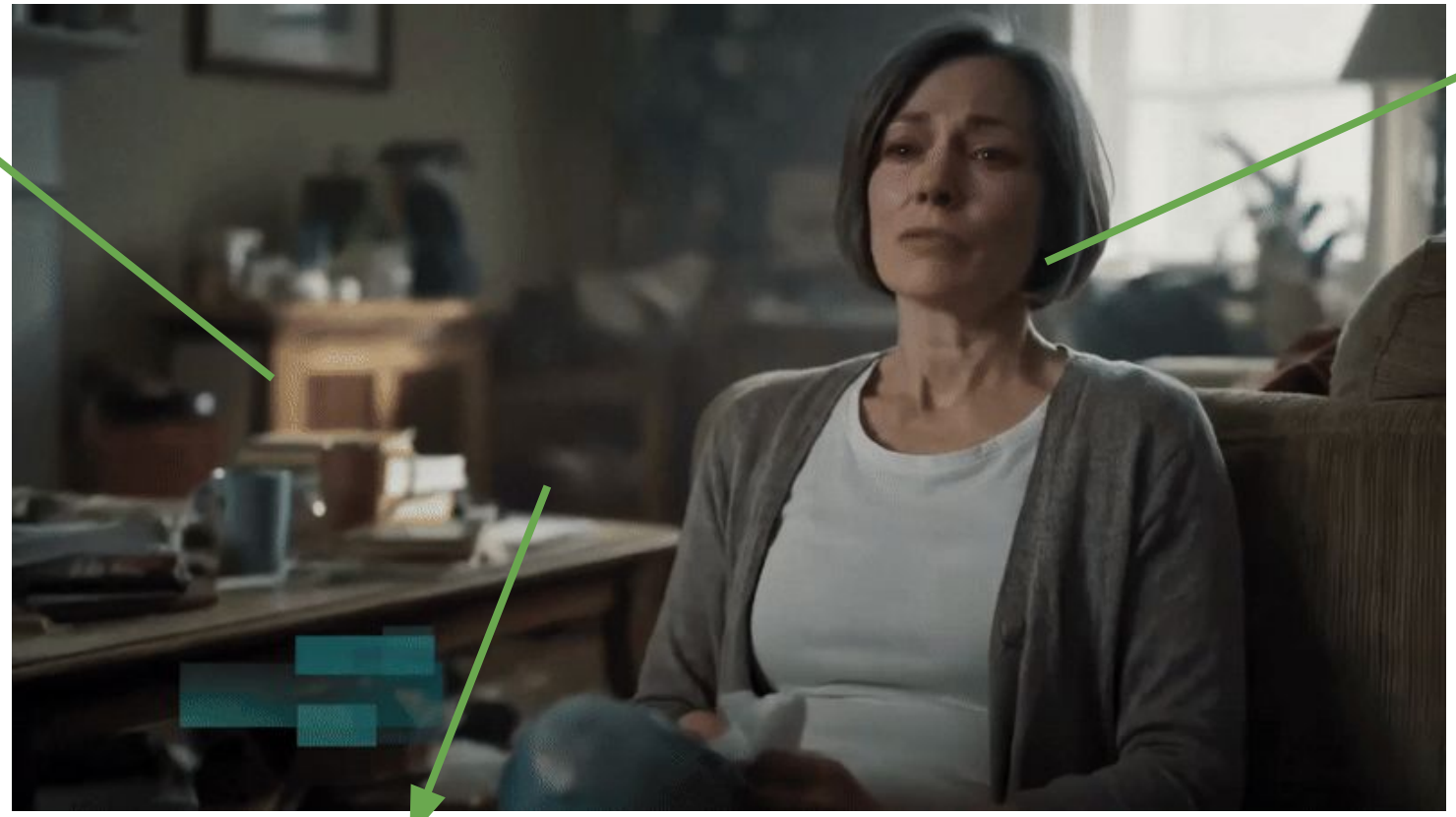
- Over the last 2 years, these models have significantly improved in **visual quality** and **temporal consistency.**

Also now multimodal!

What makes them look good?

Dynamics

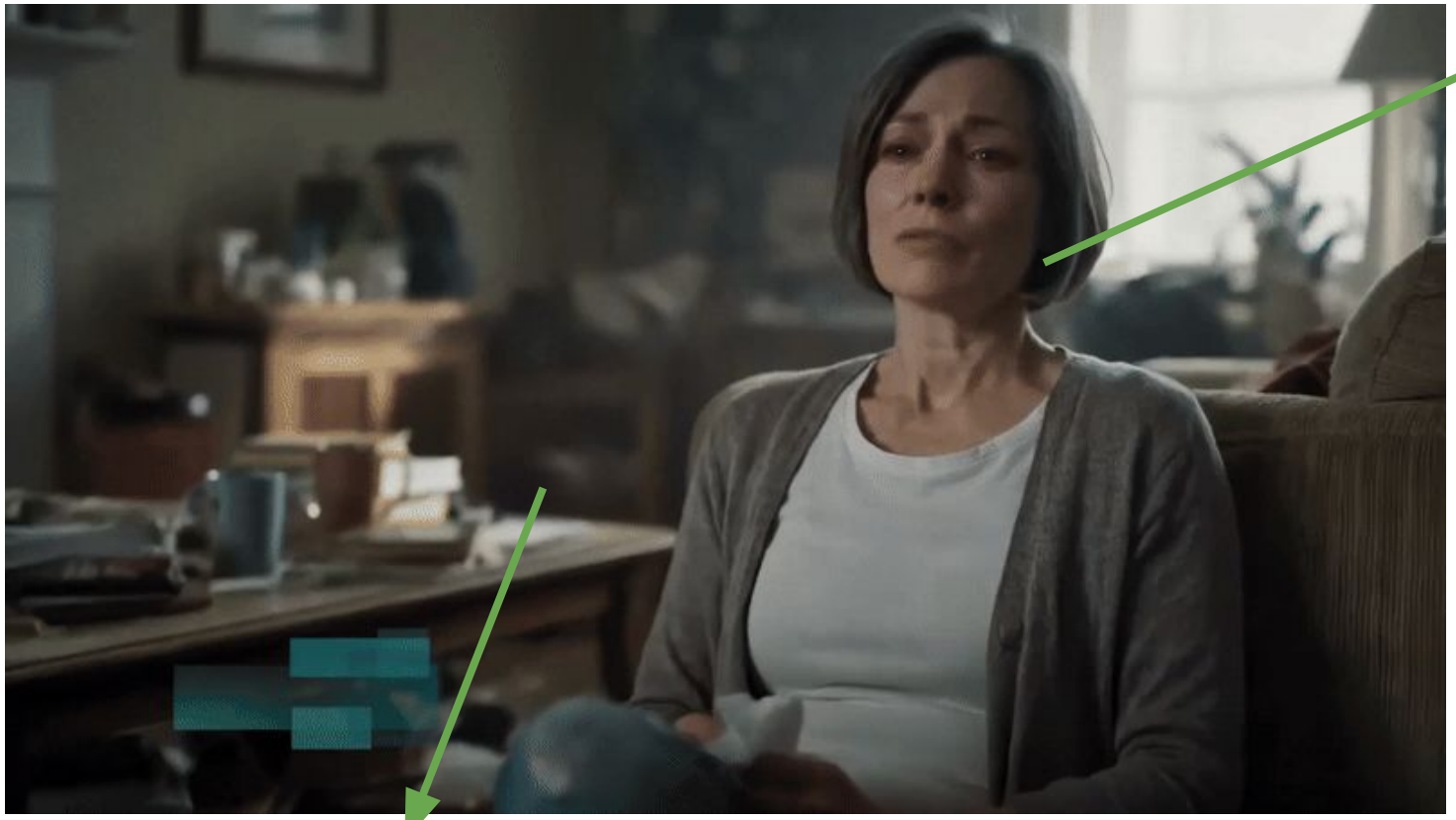
Physical Dynamics



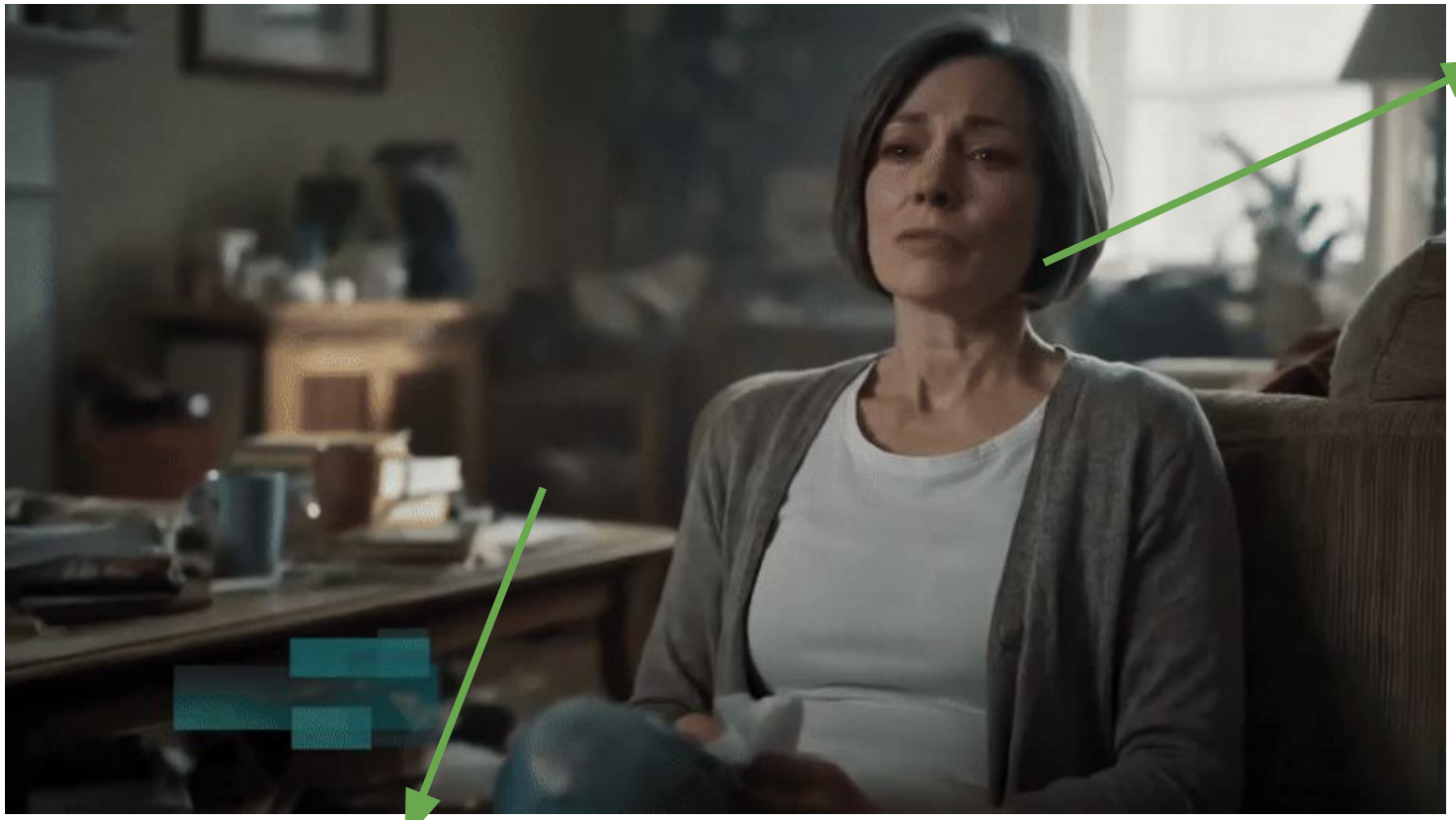
Natural Subject Dynamics

Global Dynamics (Camera Motion)

What makes them look good?



Current constraints



(Hardcoded) ~~Natural~~ Character Dynamics

Global Dynamics (Camera motion not rigged to character)

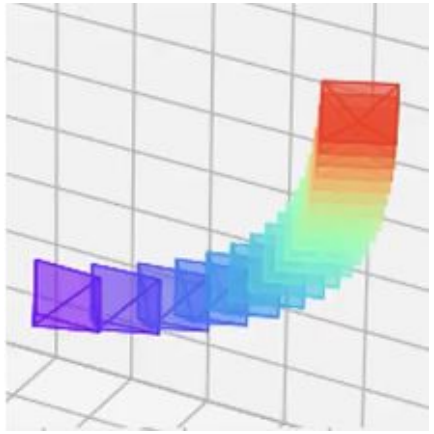
Research Statement

How can we develop a user-friendly system for modelling character centric camera motion?

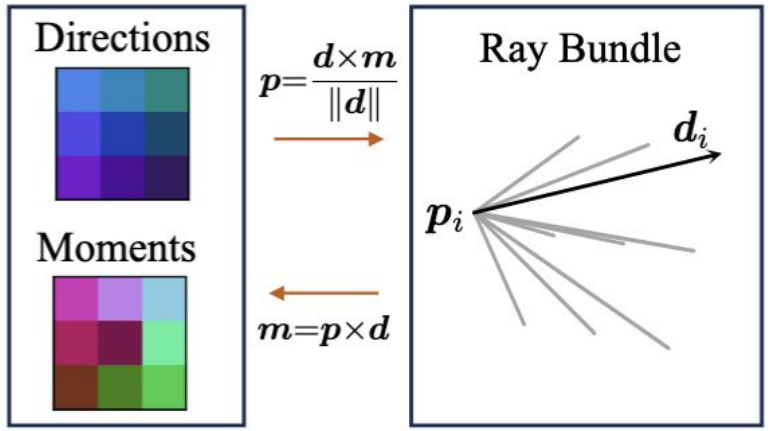
Good Dynamics for Character Videos

What is a good representation?

We need a representation that is geometric, able to model complex motion and is **user friendly.**



Pose Matrix?



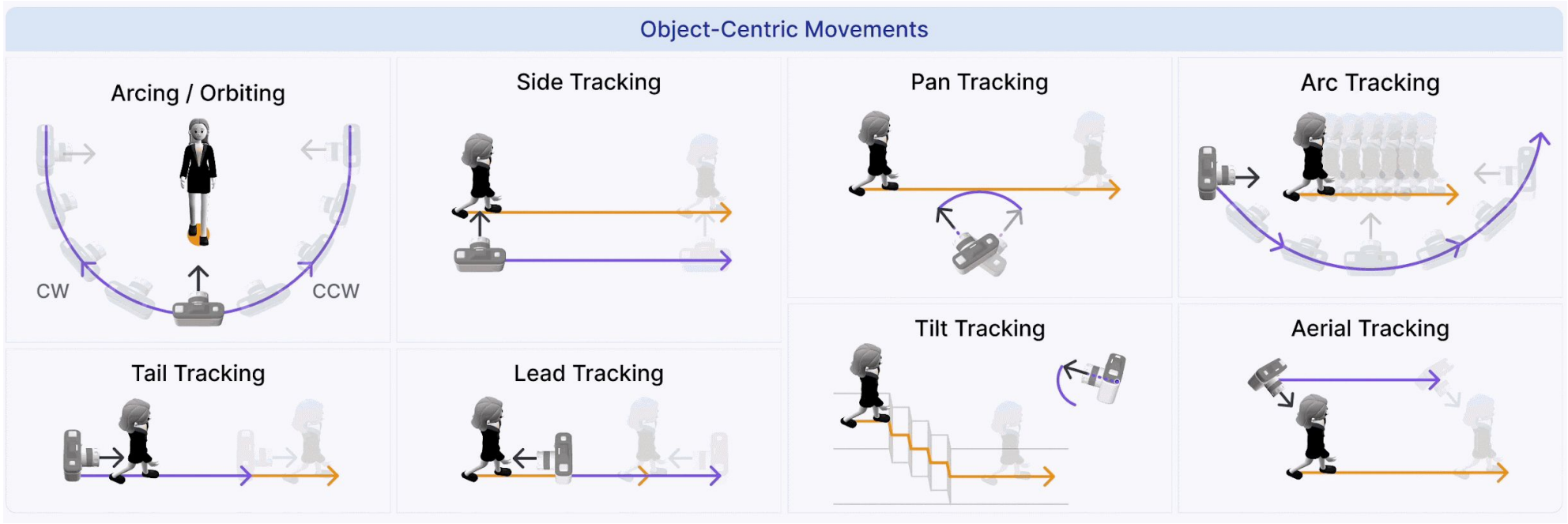
Plucker Embedding?

Point Clouds?

Good Dynamics for Human Videos

How can we model object centric motion?

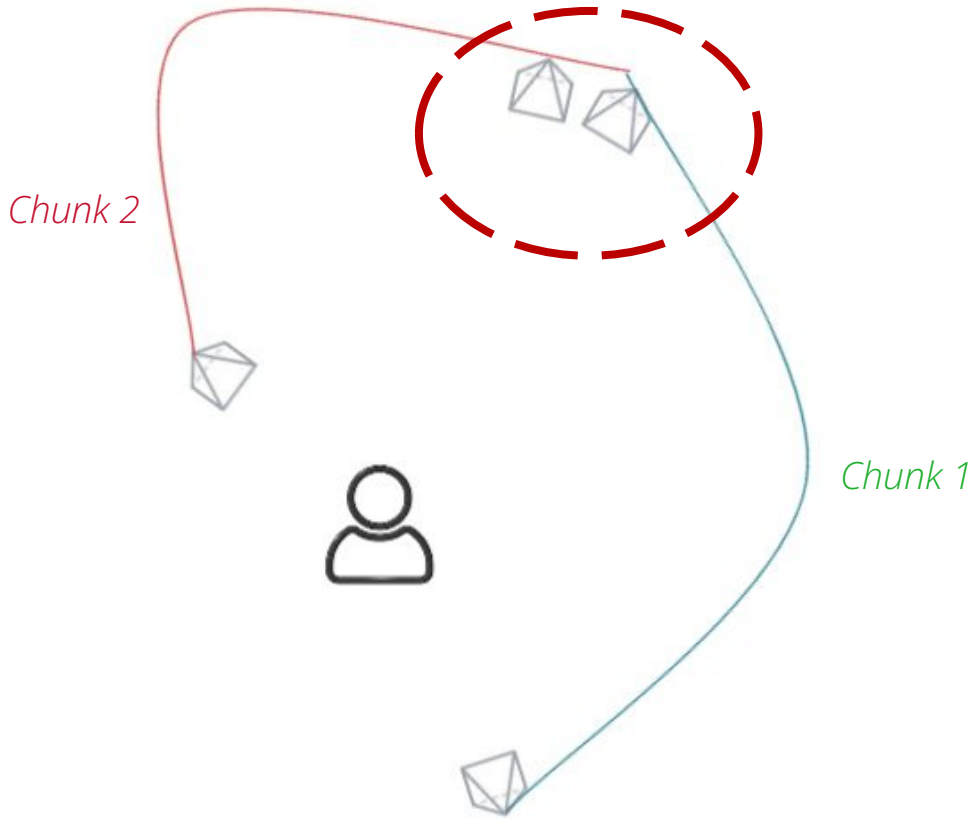
We need to regularize camera trajectory to be rigged to the subject.



Good Dynamics for Human Videos

Long-range Temporal Consistency

We need to ensure stable long trajectories without artifacts.



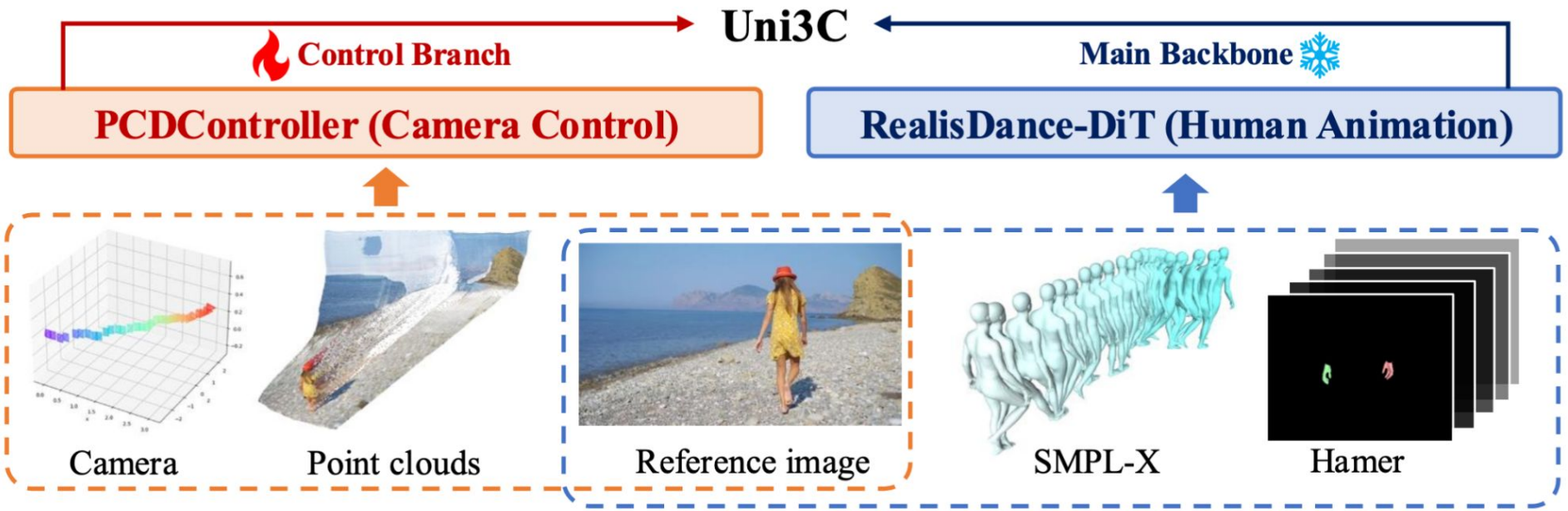
Inconsistent at the seams!

Our requirements

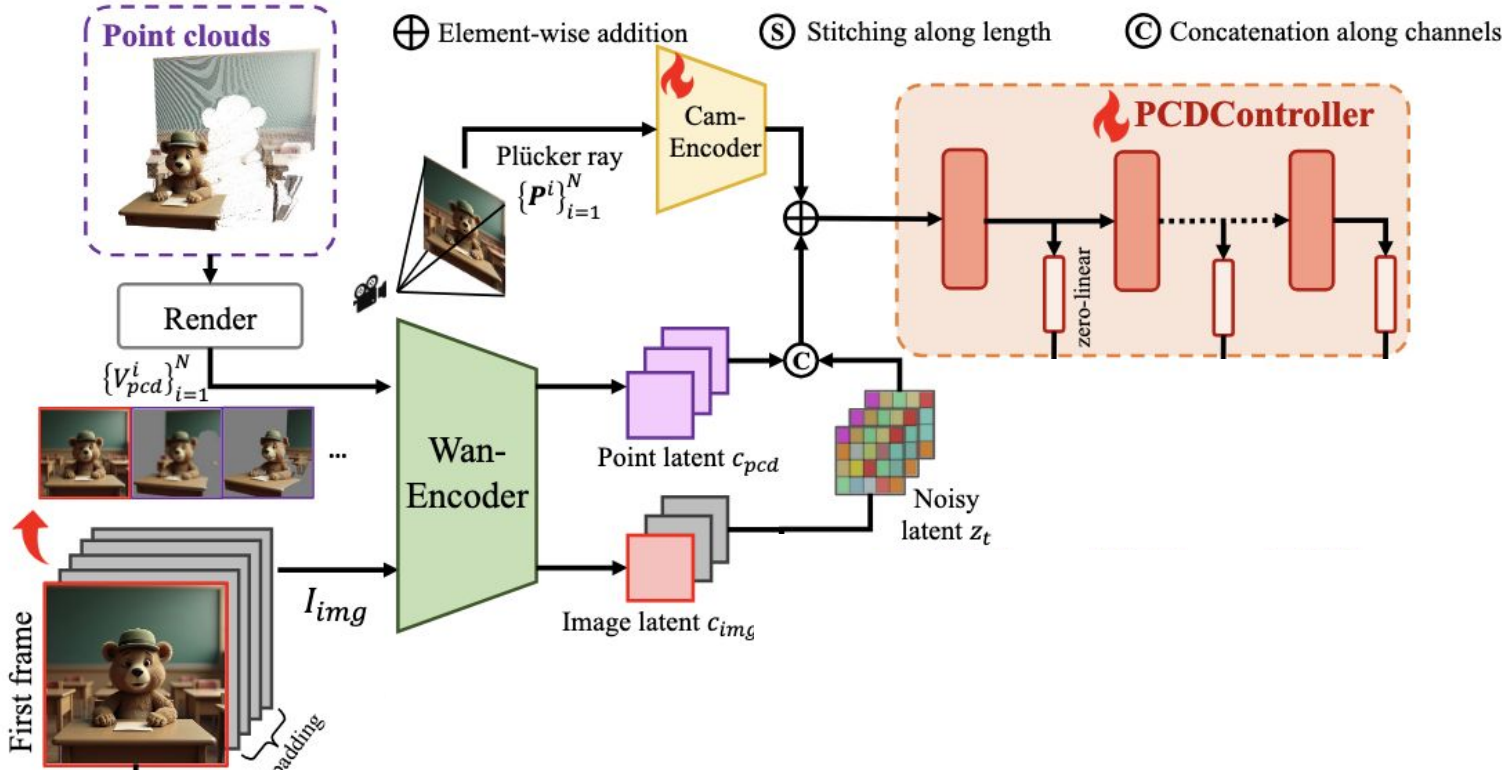
What is a good representation?	We need a representation that is geometric and is able to model complex motion.
How can we model object centric motion?	We need to regularize camera trajectory to be rigged to the subject.
Long-range Temporal Consistency	We need to ensure stable long trajectories without artifacts.

Background: Uni3C

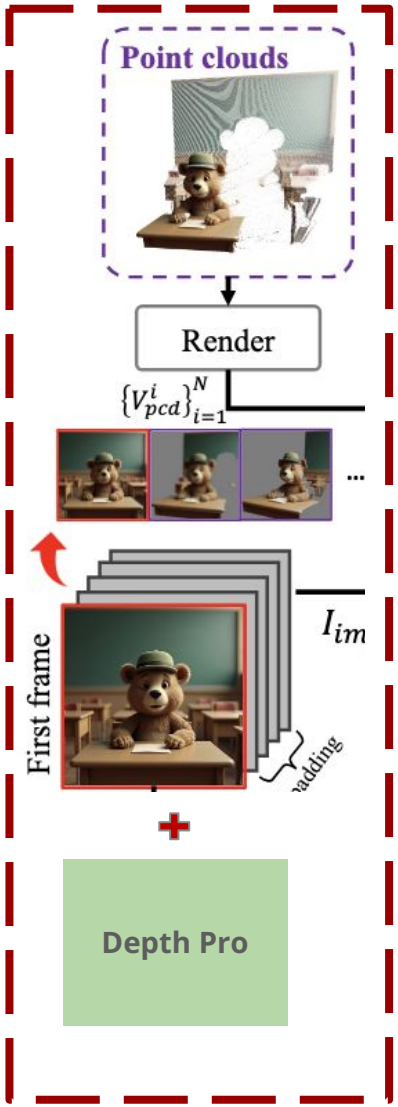
- Targets subject-centric motion and global motion
- Camera representation: Point Clouds
- Subject representation: SMPL-X



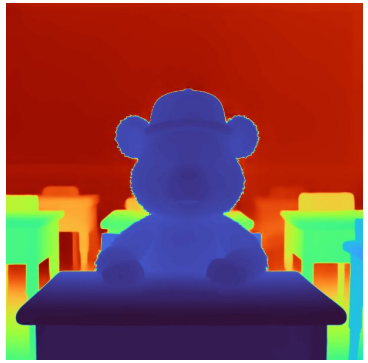
Uni3C's Camera Control



Representation



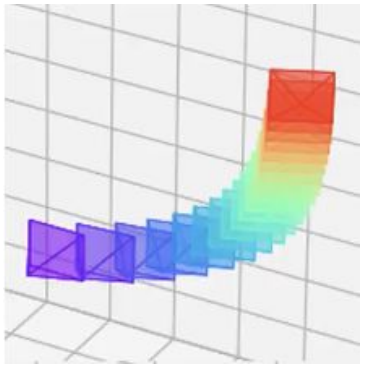
Input Frame



Metric Depth

Representation = Point Clouds!

Camera Pose

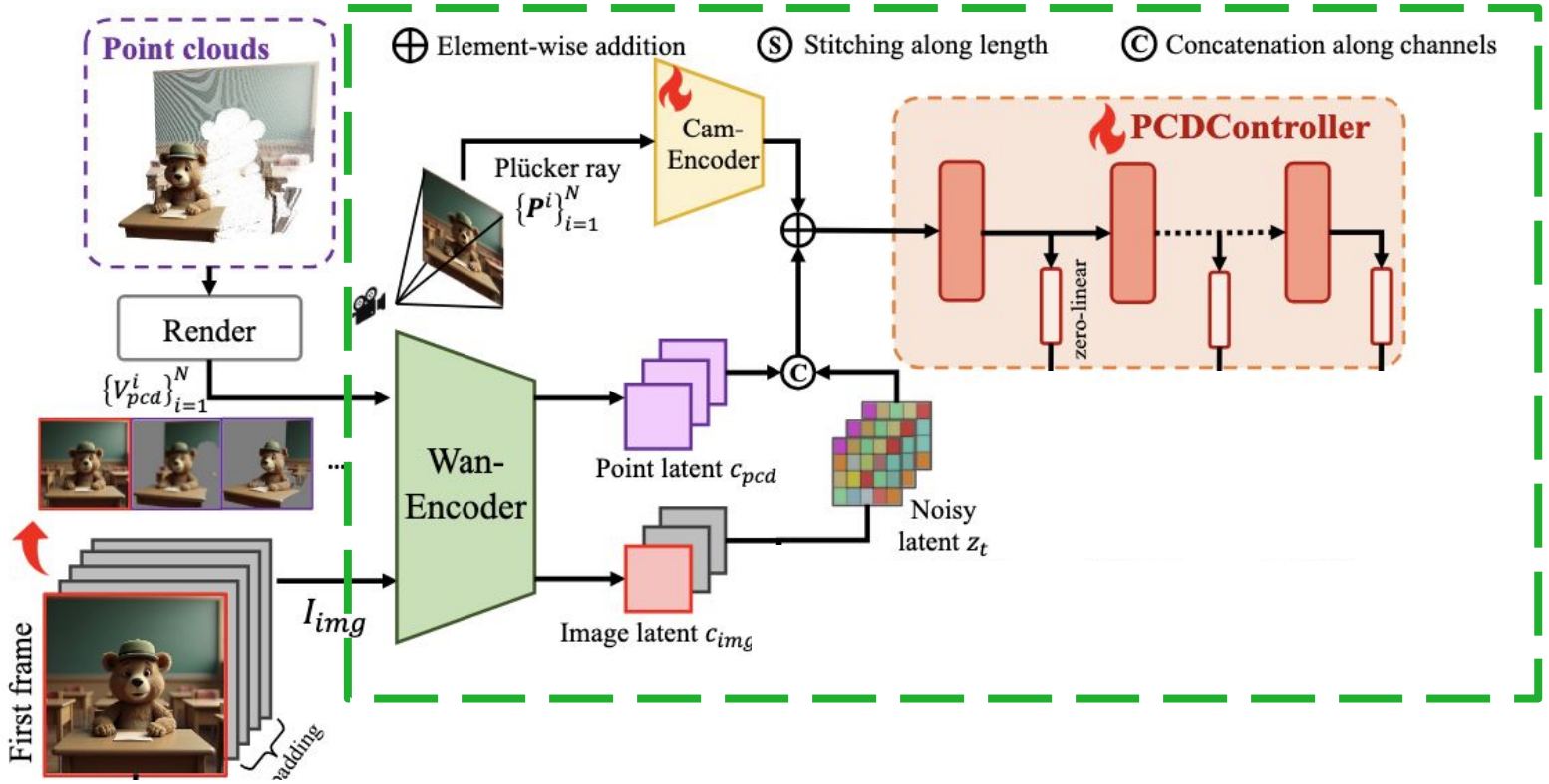


Control Video

Good metric prior.

Render video easy to embed!

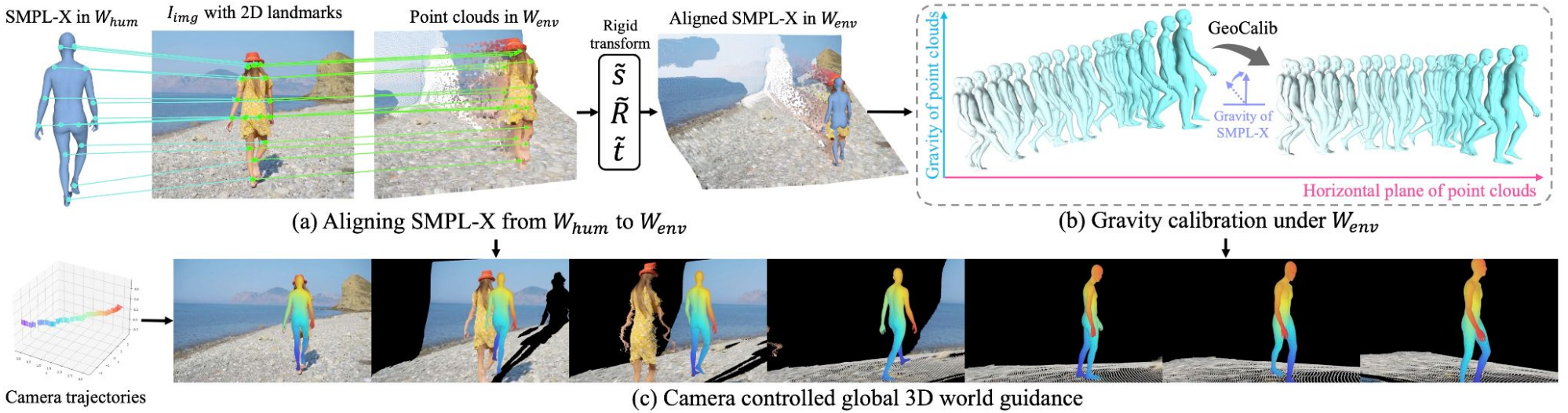
Uni3C's Control Architecture



Control Mechanism

- Use **video embeddings** of the point cloud render video
- Controlnet architecture
- Plucker is **optional** in practice (infact can introduce noise).

Subject Motion



- Extract SMPLX
- Do a rigid transformation to project joints onto rendered point cloud per frame
- Enforce horizontality of SMPL-X motion

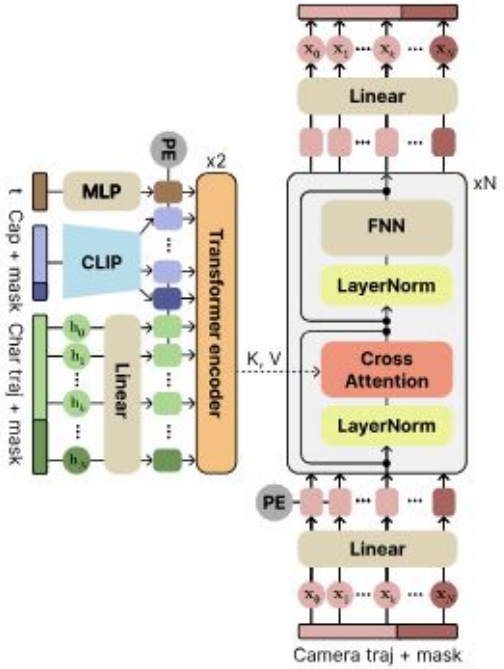
Limitations

- Good metric prior
- Convenient rendering representation
- Models both subject and camera

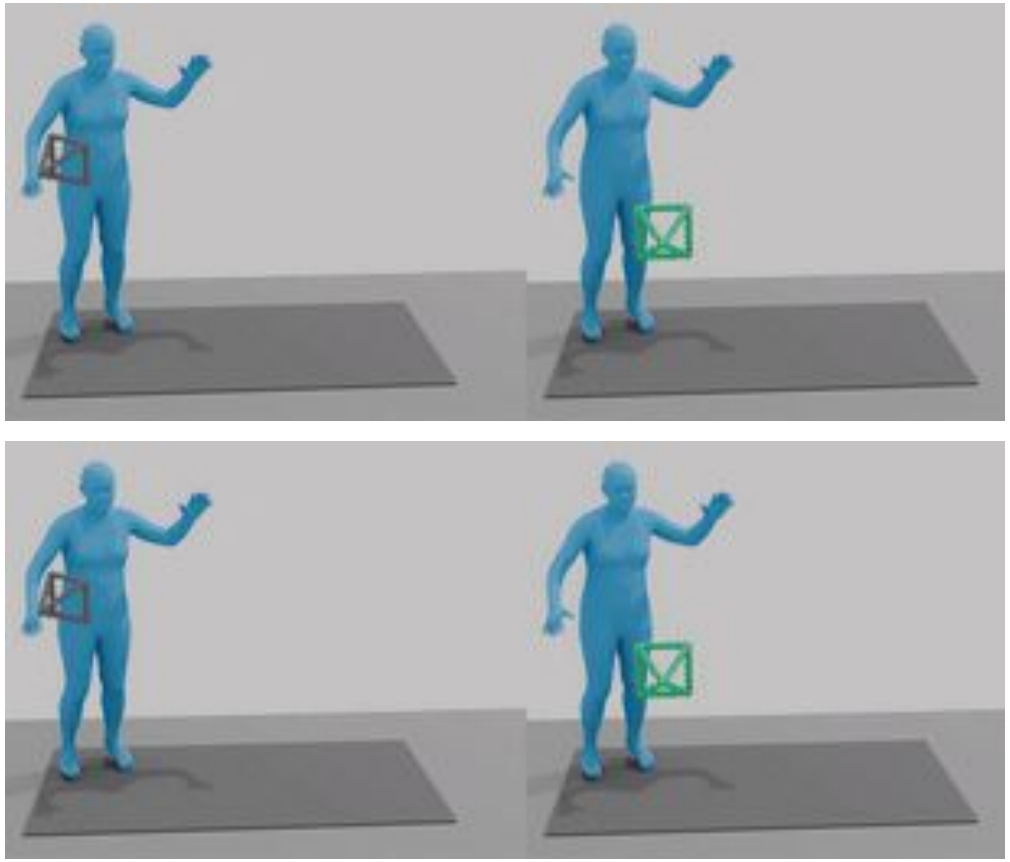
- Manual rigging to couple camera and subject
- Limited to 81 frames
- Overfits to SMPL

Background: E.T. the Exceptional Trajectories

- Essentially a video diffusion model
- Conditions on input SMPLX trajectory to generate camera poses



(c) Director C

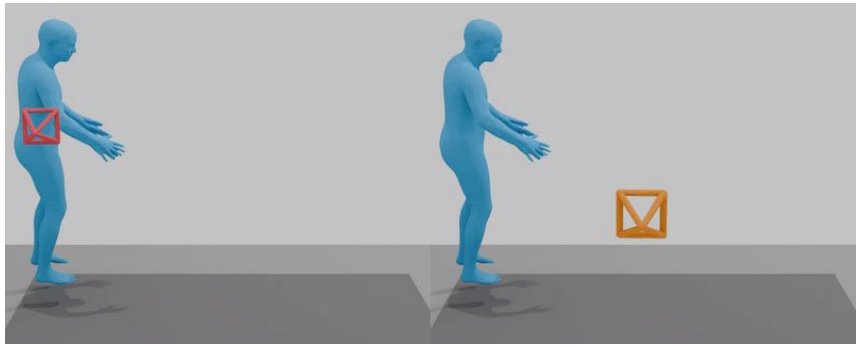


To recall: All is not well

Uni3c

- Manual rigging to couple camera and subject
- Limited to 81 frames
- Overfits to SMPL motion

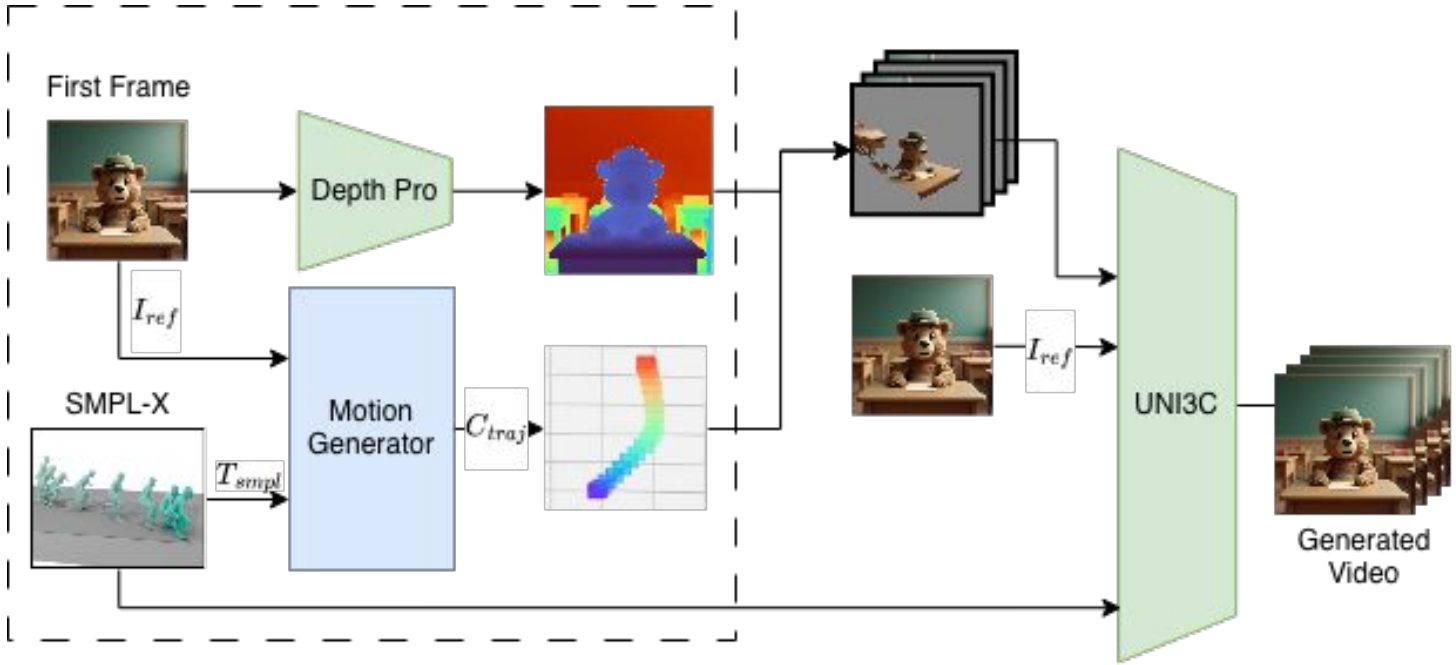
ET



- Sensitive to jitter in the SMPL-X model
- Not subject agnostic due to dependence on SMPL-X
- Not user friendly since you have to pose the smpl model

Combine ideas from both!

The Initial Solution



But a few questions

- How do we define SMPL for arbitrary characters?
- How do users generate SMPL trajectories for arbitrary inputs?
- How do we go beyond 81 frames?

Goals

1. **Addressing manual motion inputs**
2. **Addressing better subject motion generation**
3. **More consistent long video generation**

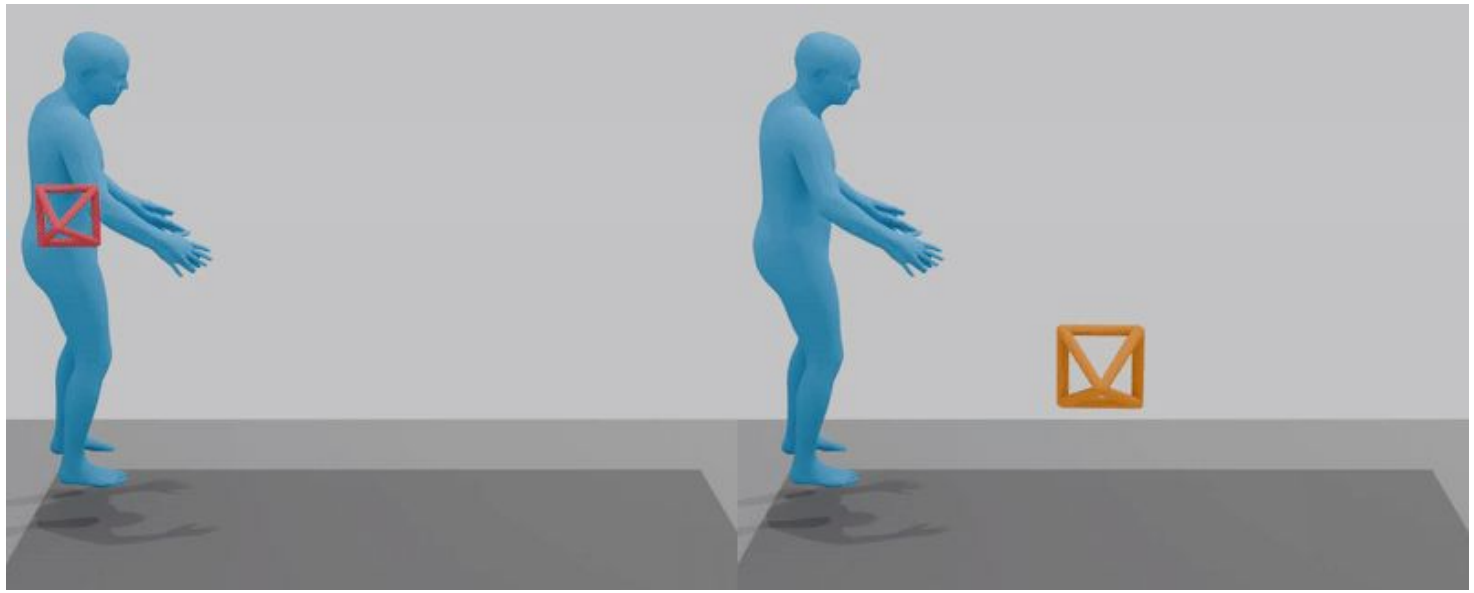
Goals

1 — 2 — 3

Addressing manual motion inputs

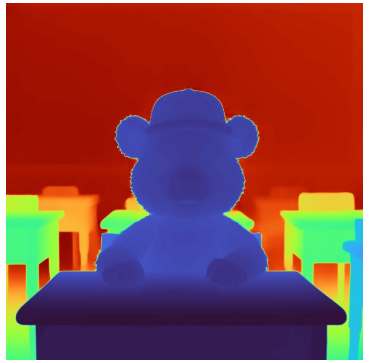
Addressing the dependance on SMPL-X

More consistent long video generation



Addressing Manual Motion Inputs

Input Frame

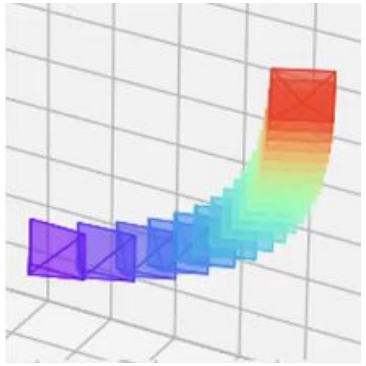


Metric Depth

Trajectory Prompt

While the character stays still, the camera [**arcs to the left** / **orbits to the left**] and then stops.

Camera Pose



Control Video

Does a plug and play solution work?

Unfortunately no…

- **ET and UNI3C** have different scene scales!

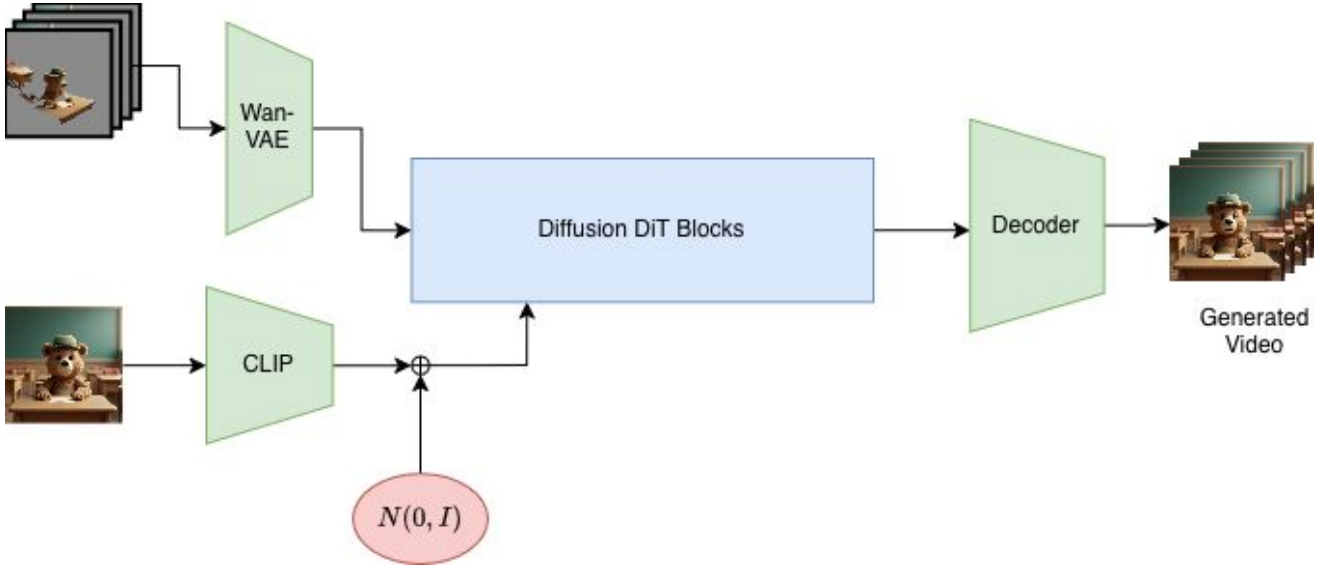
*Camera [**arcs to the left**] while the person talks.*

*Camera [**trucks to the right**] while the person runs along a wall.*

Solving with synthetic data

- Scene scale is pre-defined!
- Additionally, easy to generate human centric videos with large scene diversity

But what about the sim-real gap?



- Noise reference latent to reduce overfitting
- Fine-tune at low learning rate

Did this help?

- Yes! While the trajectories are not perfect, there is better consistency between prompt and generated motion
- However, trajectory is still unstable and jittery due to SMPL-camera rigging

With no fine-tuning

"Camera [zooms in slowly] with the woman talking passionately"

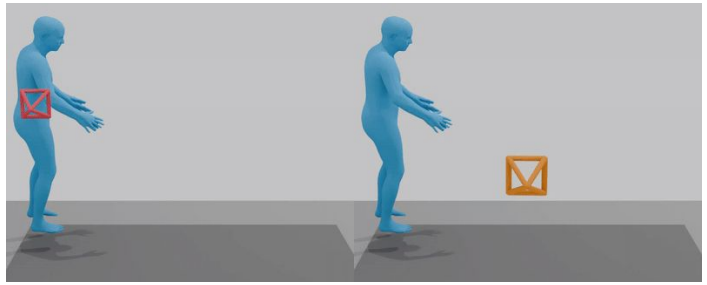
With fine-tuning

"Camera [zooms in slowly] with the woman talking passionately"

Goals

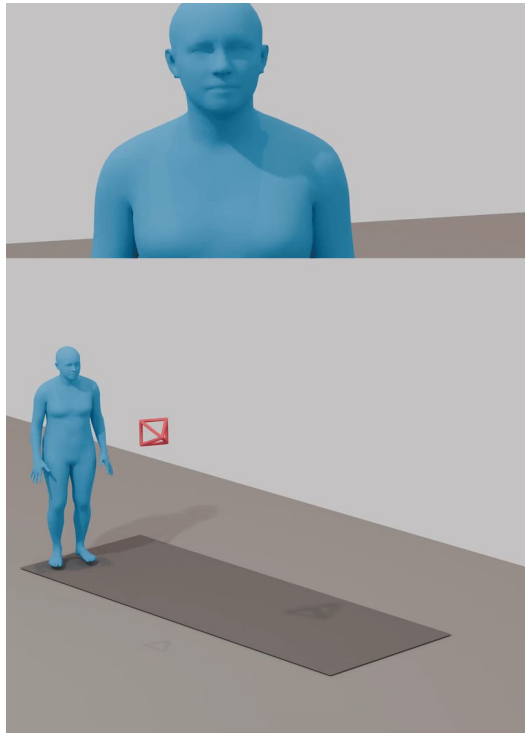
1

Addressing manual motion inputs



2

Addressing the dependance on SMPL-X



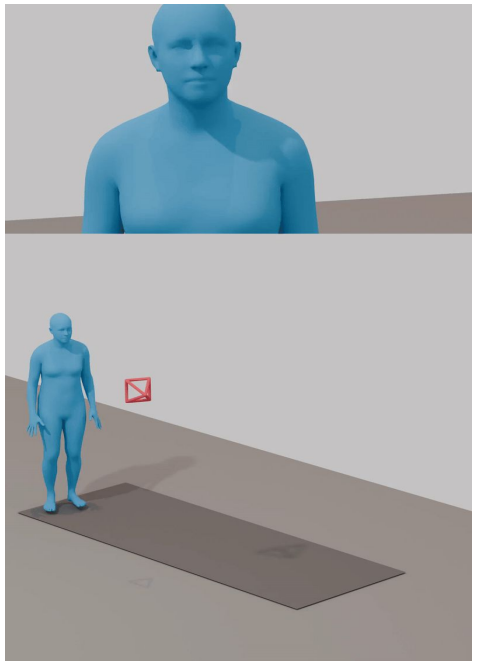
3

More consistent long video generation

Why is SMPL-X a problem

Great when the character is human

Camera motion is largely inaccurate

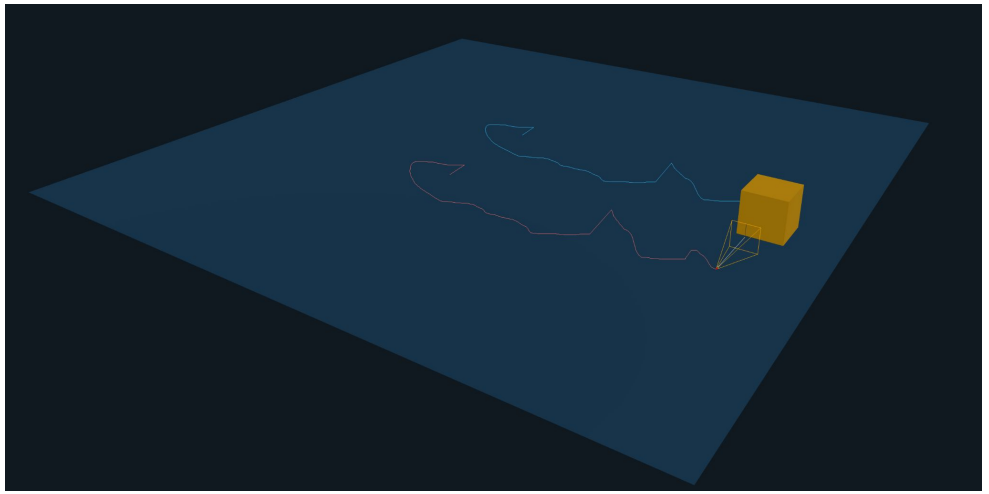


We want a conditioning signal that's not too granular

Also should be easy for the user

Condition on 3D tracks

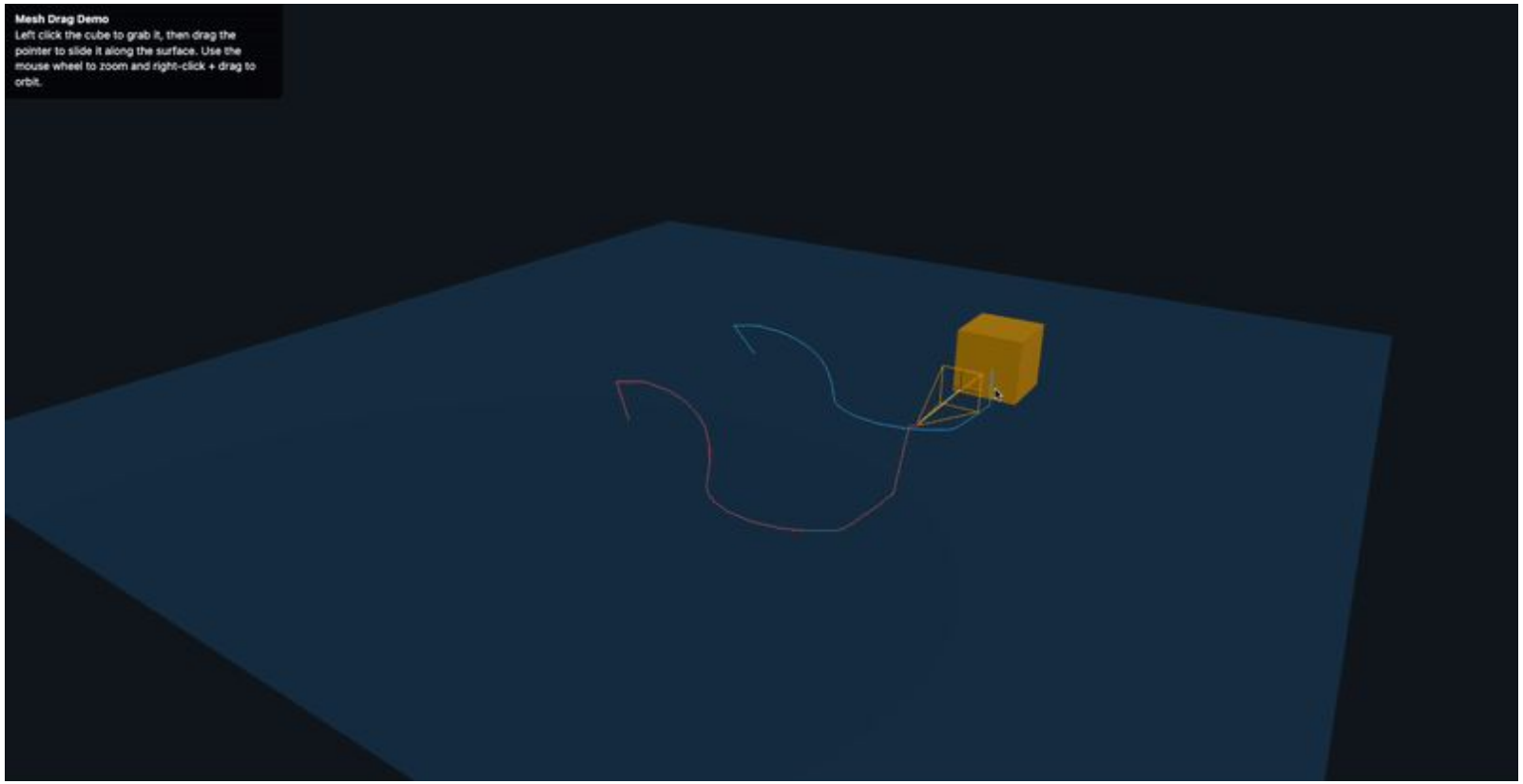
- Similar papers like Motion Prompting, DragAnything, GoWithTheFlow, but **uplift** to 3D!



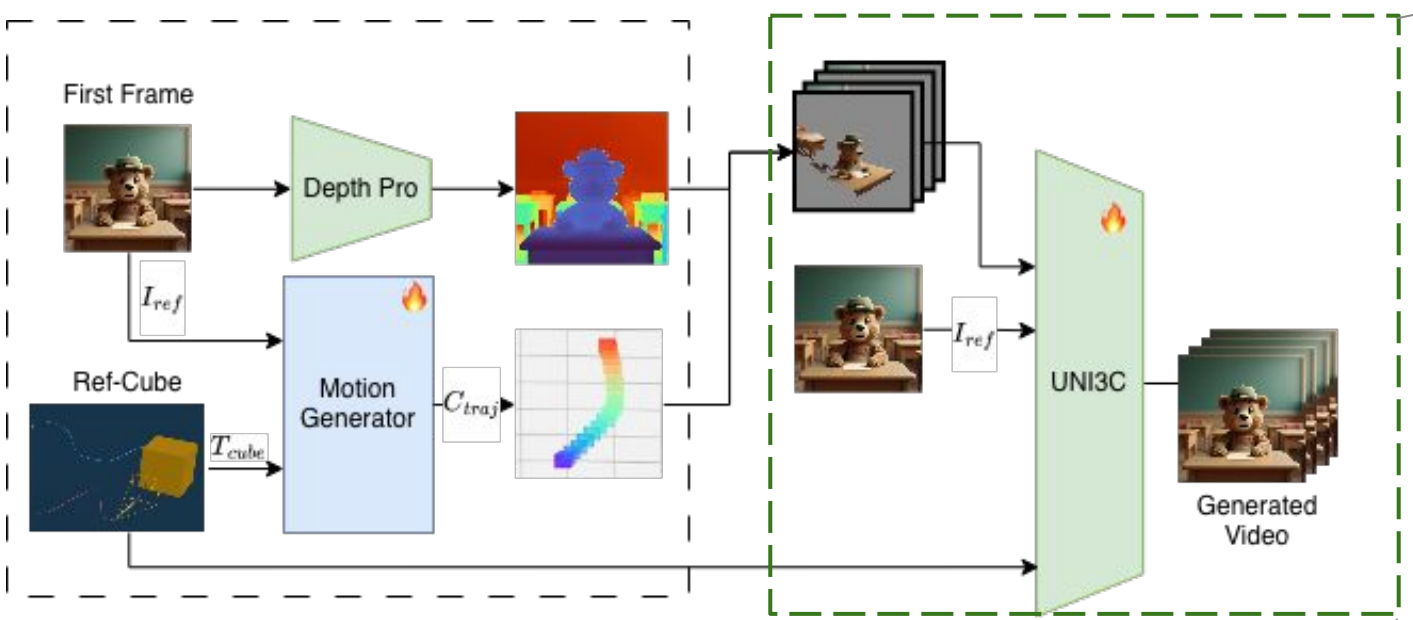
Use 3D motion tracks plus a reference collider for condition

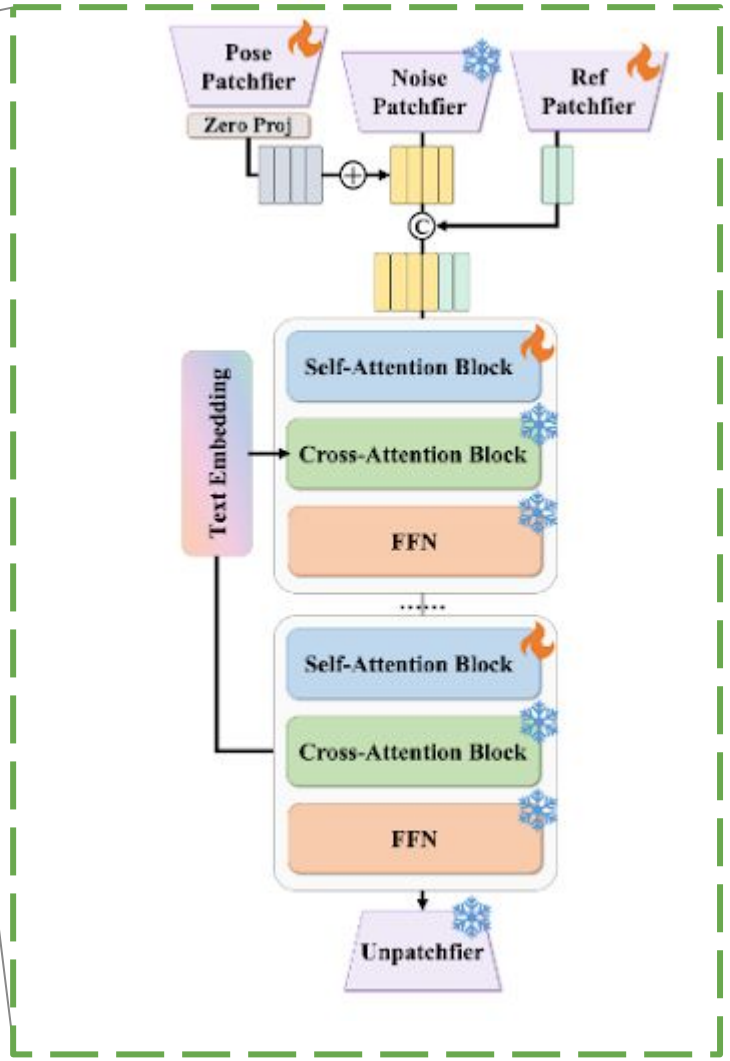
Also easier for the user

- No inconvenient SMPL-X posing required. Just need to move the cube around
- Later projected into scene PCD to use as a conditioning signal



Updated Architecture





- Only fine-tune the self-attention blocks
- **Temporal concatenation**, empirically found to do better than channelwise
- Fine-tuned the temporal self-attention blocks to adapt to temporal concat.

Results

"Camera [arcs to the left] with the woman talking"

"Camera [arcs to the left] with the woman talking

"Camera [arcs to the left] with the man talking"

Also works better for non-humans

Before

"Camera [trucks to the right] with the cat walking in a park

After

"Camera [trucks to the right] with the cat walking in a park

Caveat: Takes about 100k / 81 frames / 64 batch over 16 GPUs

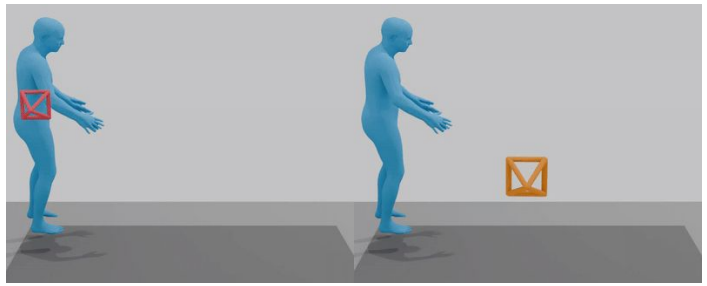
What about reshooting?

- Similar to the previous setting, we condition the video model on a set of reference **video latents** instead of an image latent
- Allows us to **reshoot** and correct camera trajectory after the fact as well
- Ex: Below from Recammaster's pipeline

Goals

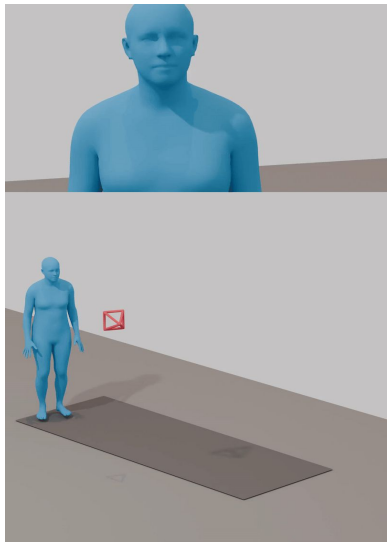
1

Addressing manual motion inputs



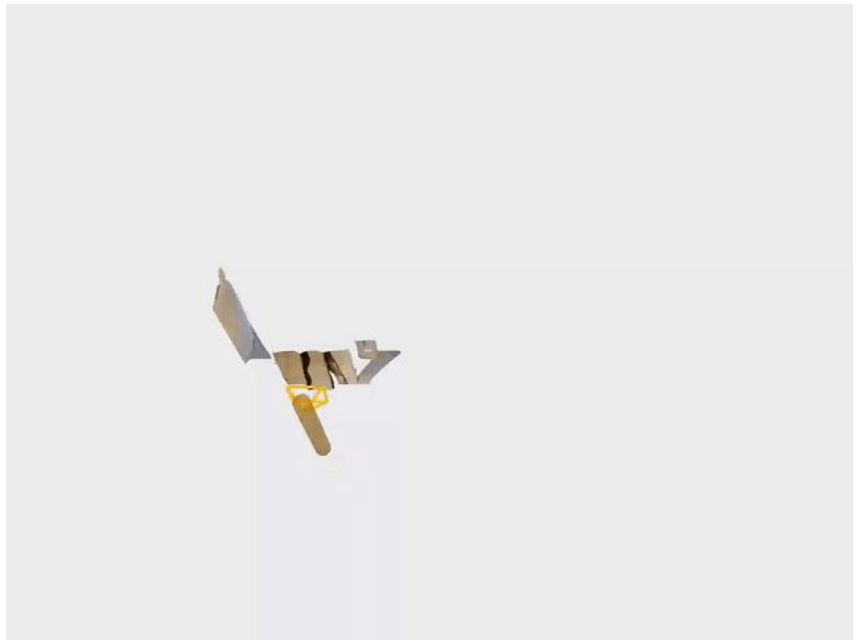
2

Addressing the dependance on SMPL-X

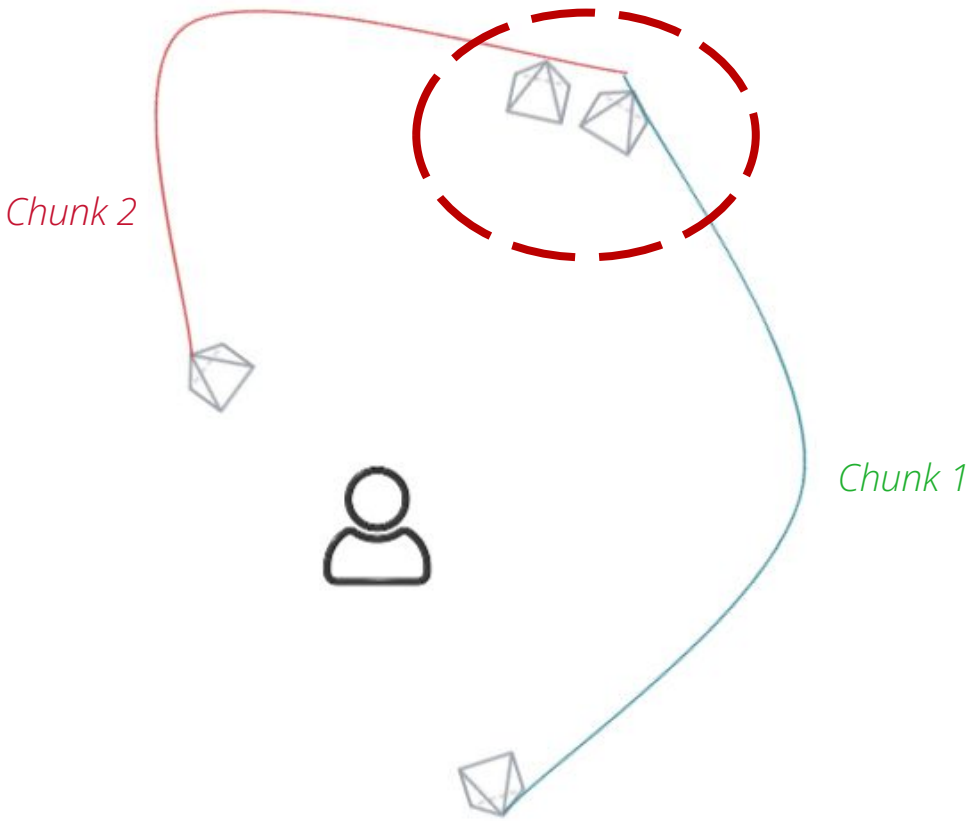


3

More consistent long video generation



Long Range Consistency



The Challenge

- Content consistency
- Error accumulation

The solution to content

A point cloud cache!

- Memorize already generated content
- Store the metric reprojections (using DepthPro) back into the point-cloud.
- Generate conditioning PCD off of this cache

But error accumulation remains unsolved

Future Directions

- Can we use methods like diffusion forcing to reduce error accumulation?
- Can we apply this system to generate character consistent worlds and shoot full-sequence videos?
- Can we leverage these models to model physical phenomenon?

Collaboration

Massive scope = multi-team effort to make things production ready

Data Engineers

- Video data sourcing
- Data validations to ensure processing pipelines were working correctly

ML Engineers

- Helped productionize the code and make it more efficient
- Helped with good coding practices and with clean codebases

Frontend

- Key in identifying gaps in the user interfacing aspect of the work

- Learn how to forecast better
- Task delegation to speed up iterations
- Learn how to let people do what they're good at

3D-VLMs

Marriage of ideas!

Can we answer 3D-VQA questions using 2D VLMs?

Where are the two different size tables located in the room?

In what part of the room is the brown armchair located?

Where is the paper towel dispenser located?

Where is the window located?

Marriage of ideas!

Can video models help with this?

Imagine forward

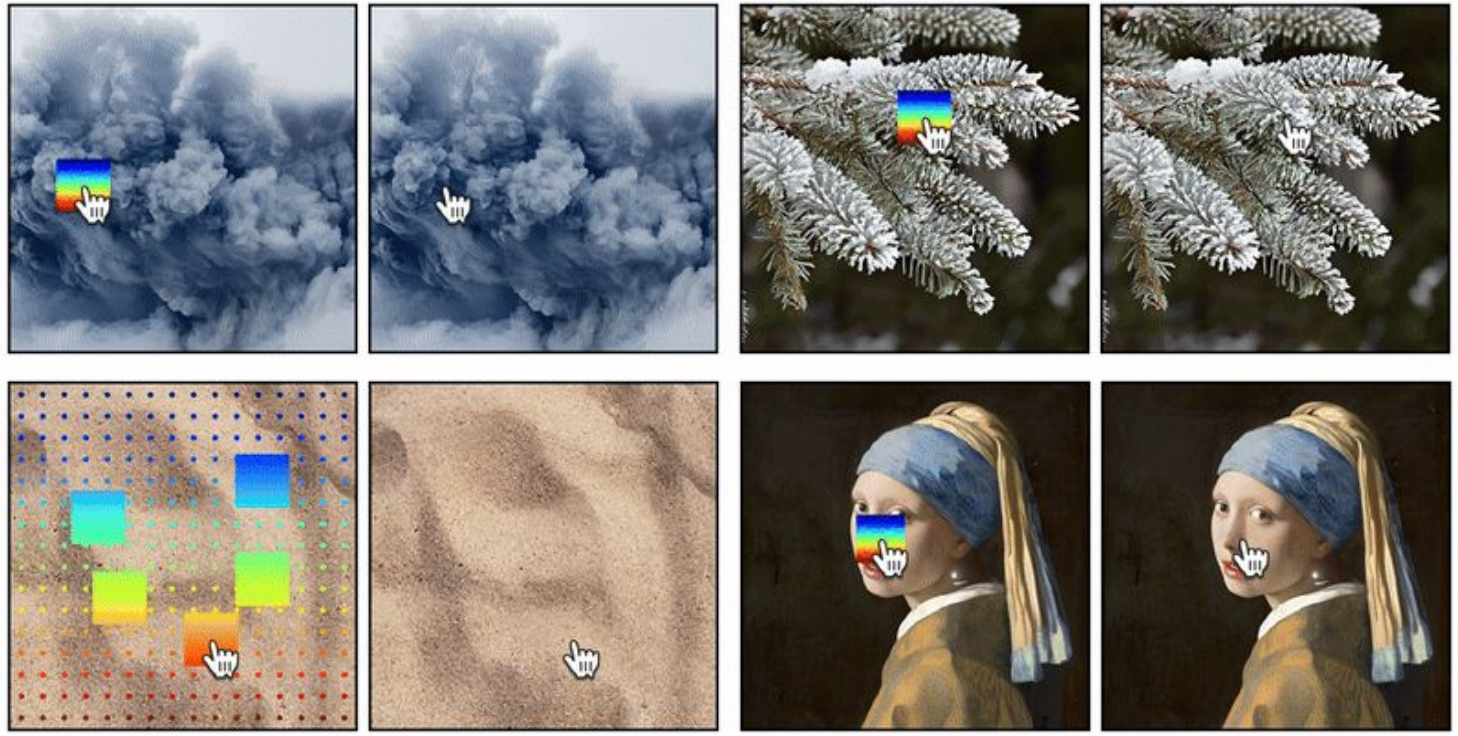
Then answer questions

We generate detailed views of the scene from a single image

Future Trends/Vision

Video models have emergent properties

- Models trained on optical tracks can simulate physical phenomenon

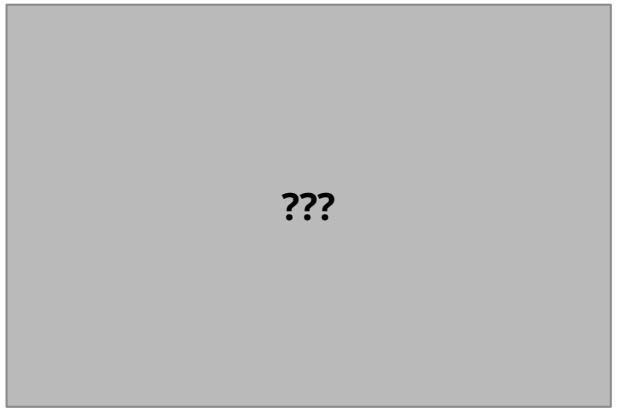


But they lack true physical grounding!

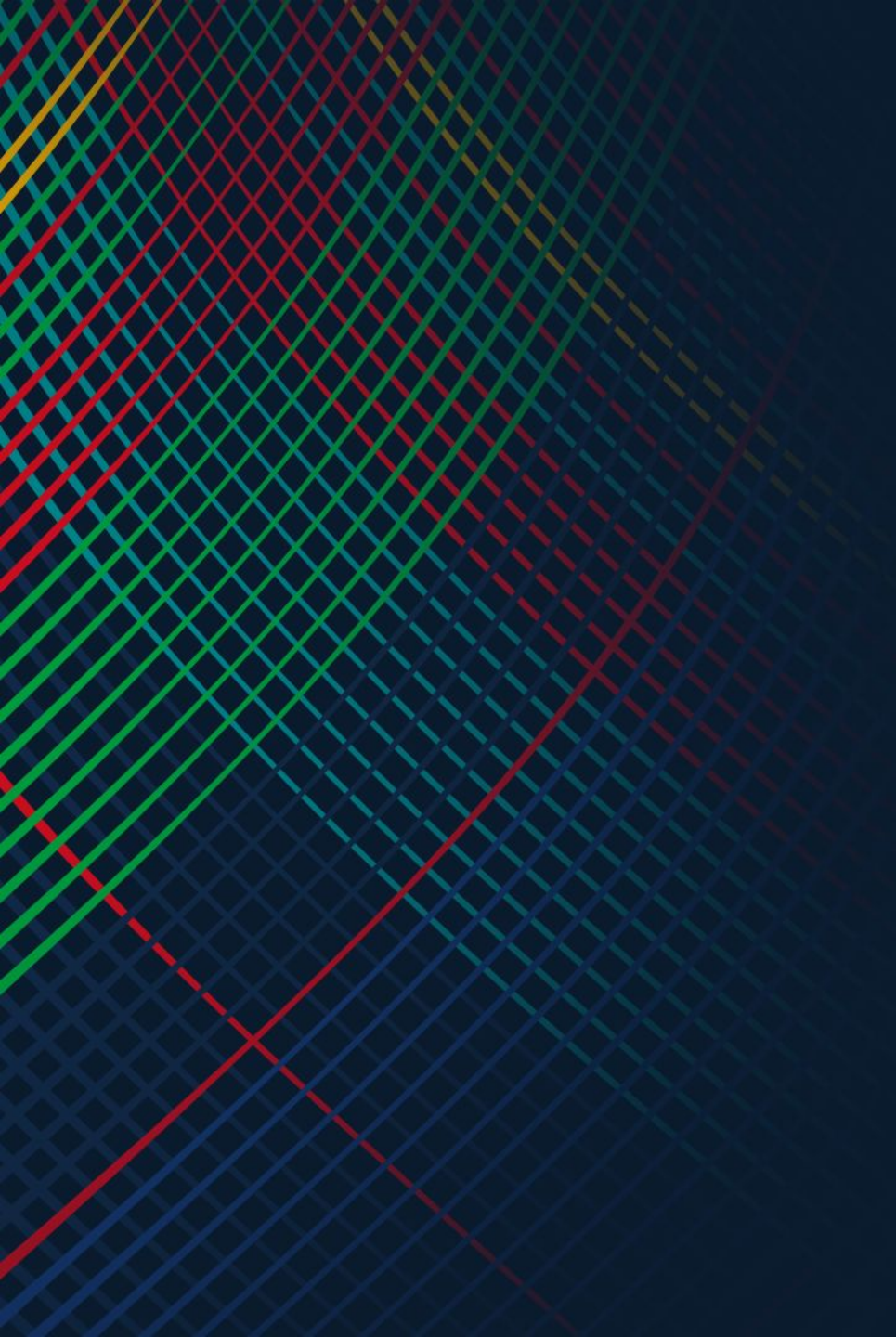
Video models that simulate physical phenomenon

Scene dynamics

Physics



Biological Systems



Thank you!