3D-IntPhys: Towards More Generalized 3D-grounded Visual Intuitive Physics under Challenging Scenes

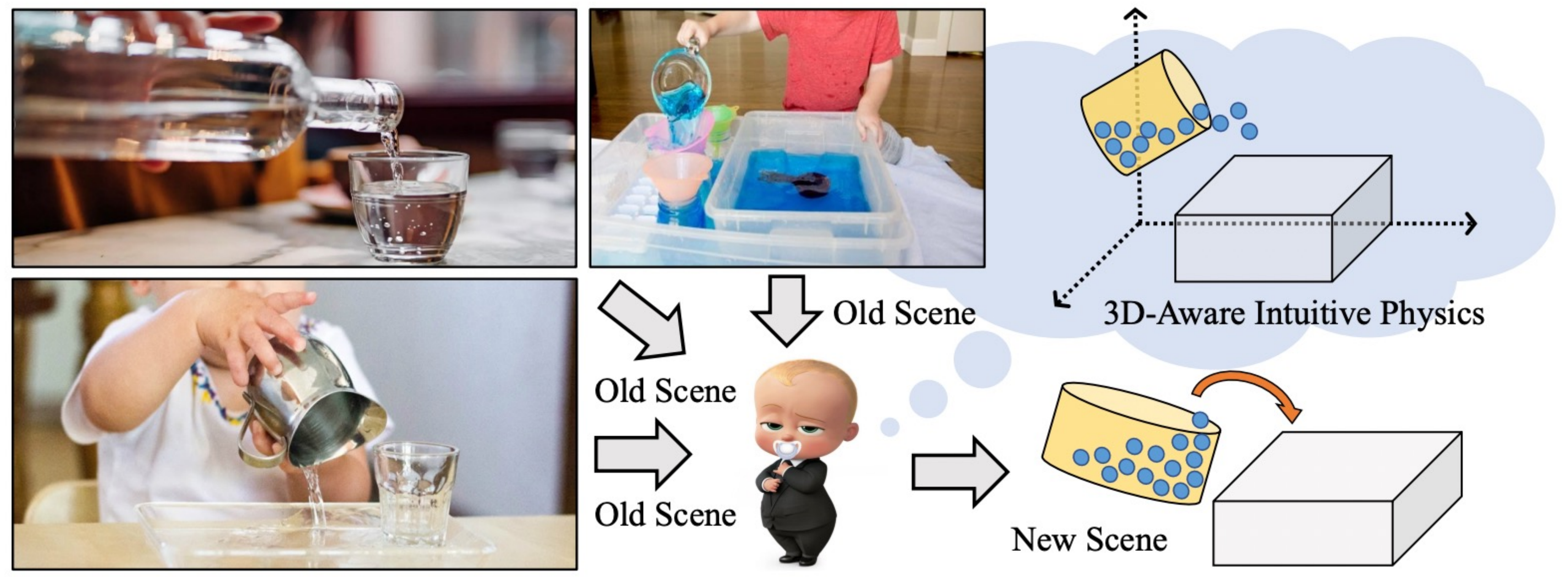
Haotian Xue[1], Antonio Torralba[2], Joshua B. Tenenbaum[2], Daniel LK Yamins[3], Yunzhu Li[3], Hsiao-Yu Tung[2]

1 2 3

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Introduction & Motivation:

- Humans have ability to gain strong intuition about the physical world around them, we can predict the movement of complex dynamics in 3D space without knowing the underlying dynamics:



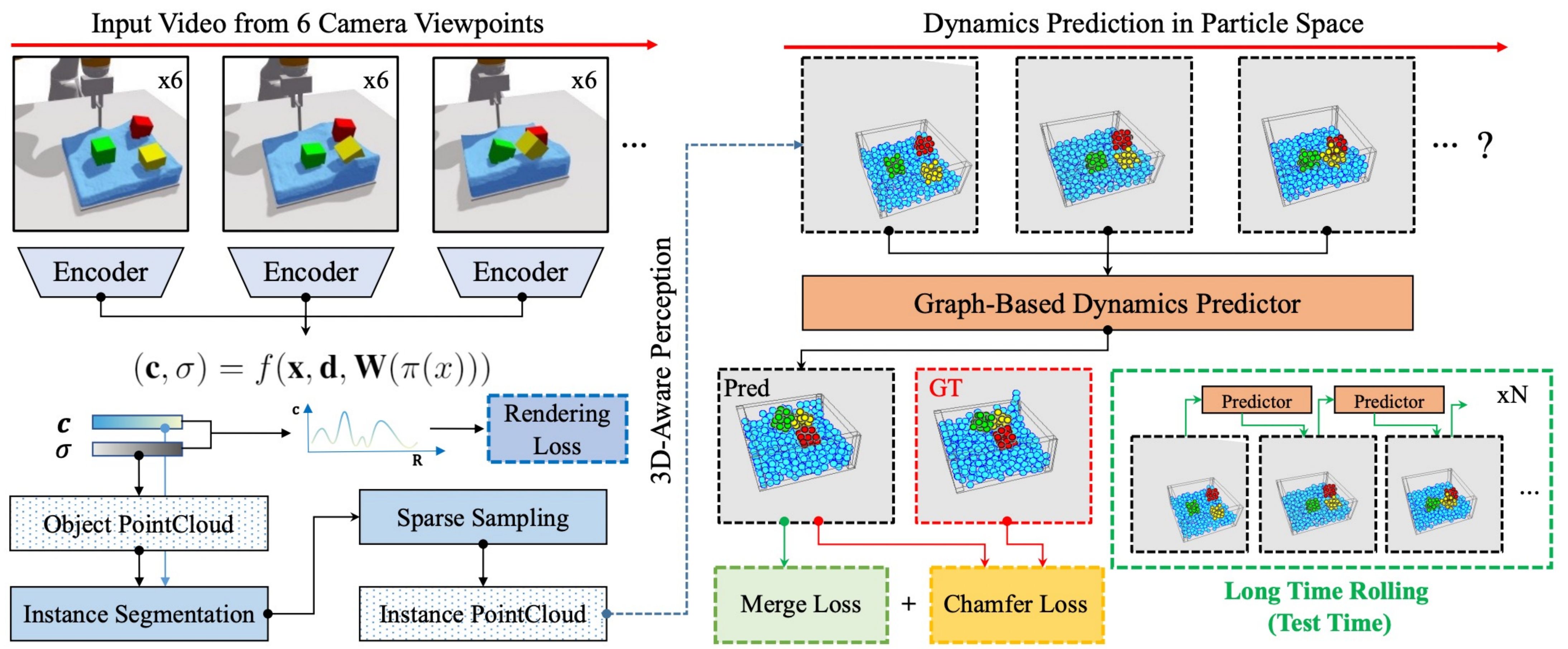
- Learning from visual inputs of old scenes, humans can generalize the acquired 3D-aware intuition to new scenes.
- Here we propose a novel framework to enable machine to learn such kind of 3D-aware intuitive physics from solely visual inputs.
- By imposing strong inductive bias, our methods can learn reasonable intuitive physics from visual inputs, it also has a strong generalization ability to unseen settings.

Limitations of Previous Methods:

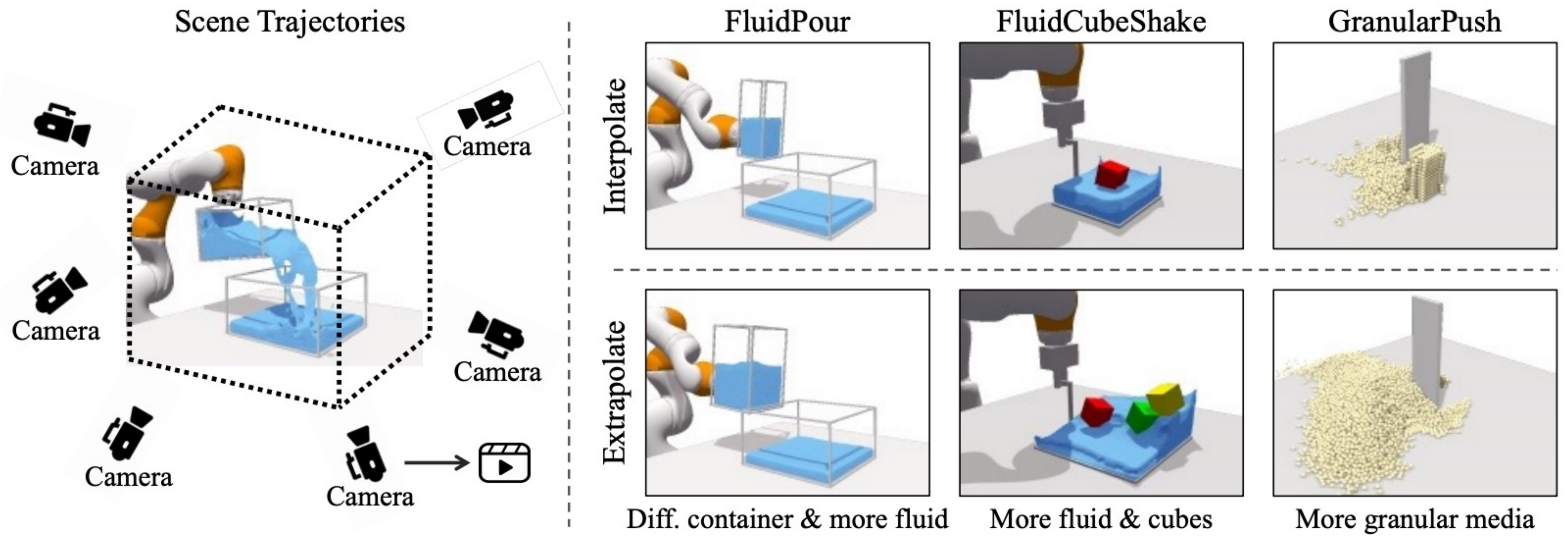
- Some of them focus on single view **2D prediction**, cannot deal with inputs from different views, making it hard to apply to 3D scenes
- Some of them require **dense 3D annotations** to learn particle-based intuitive physics, cannot learn from visual inputs
- Some of them do not use explicit 3D representation, instead they turn to use **global feature** to encode 3D scene, making it hard to generalize to unseen scenarios

Methods:

- We propose **3D-IntPhys**, which can learn 3D-aware intuitive physics from visual inputs:

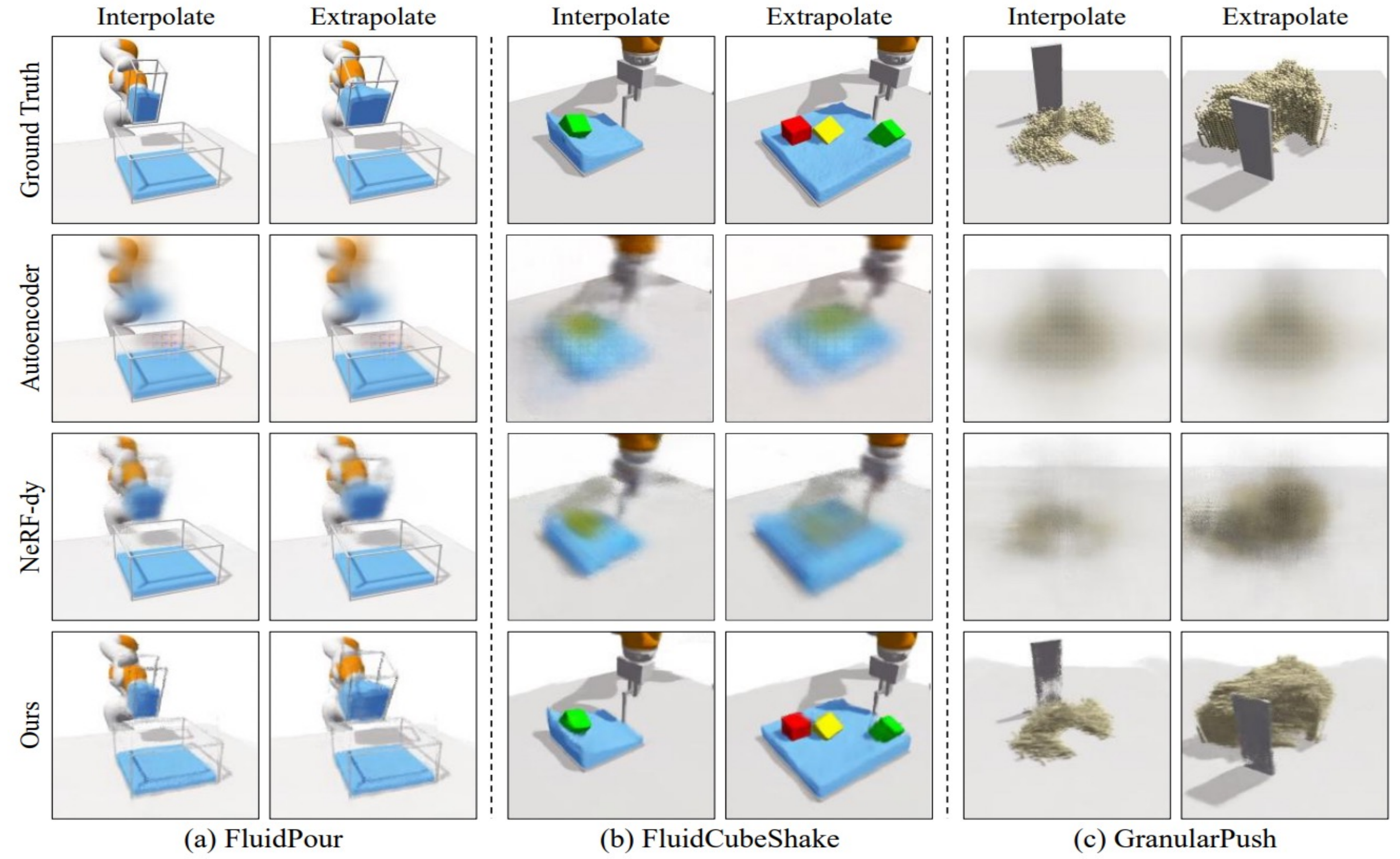


- Our method is composed of a **conditional NeRF-**style visual frontend and a **3D point-based dynamics** prediction backend, **imposing strong structural inductive bias** to capture the structure of the underlying environment.
- We first train conditional NeRF to reconstruct explicit 3D representation **from video subset**(self supervised, without additional 3D annotation), then we learn explicit 3D dynamics with **Chamfer Loss** and Merge Loss.
- We generate **multi-view dataset** of three common scenes: **Pour water**, **Shake water and cubes**, **Push granular materials**, include interpolate and extrapolate settings:

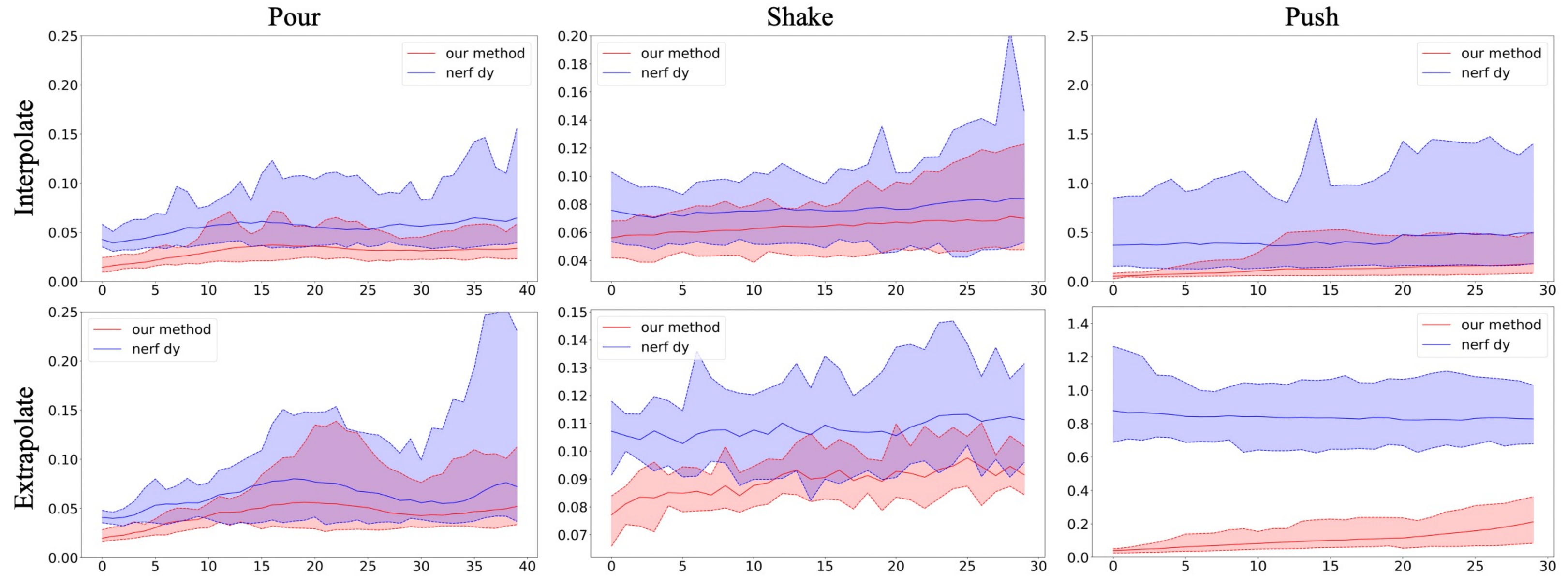


Results:

- 3D-IntPhys has a strong visual head which can generalize better than baseline methods (e.g., NeRF-dy and AE)



- 3D-IntPhys can make long-horizon future predictions in 3D space by learning from raw images, significantly outperforming baseline methods (NeRF-dy), both in interpolate and extrapolate settings:



More video results can be found here 👉