Volumetric Animation and Manipulation
‘Do as I Do’ & ‘Do as I Say’

Sergey Tulyakov

Snap Research

Trajan's Column - Dacian War Comic Strip

Shows 155 scenes of 2 Dacian wars

Wings around 23 times

Trajan

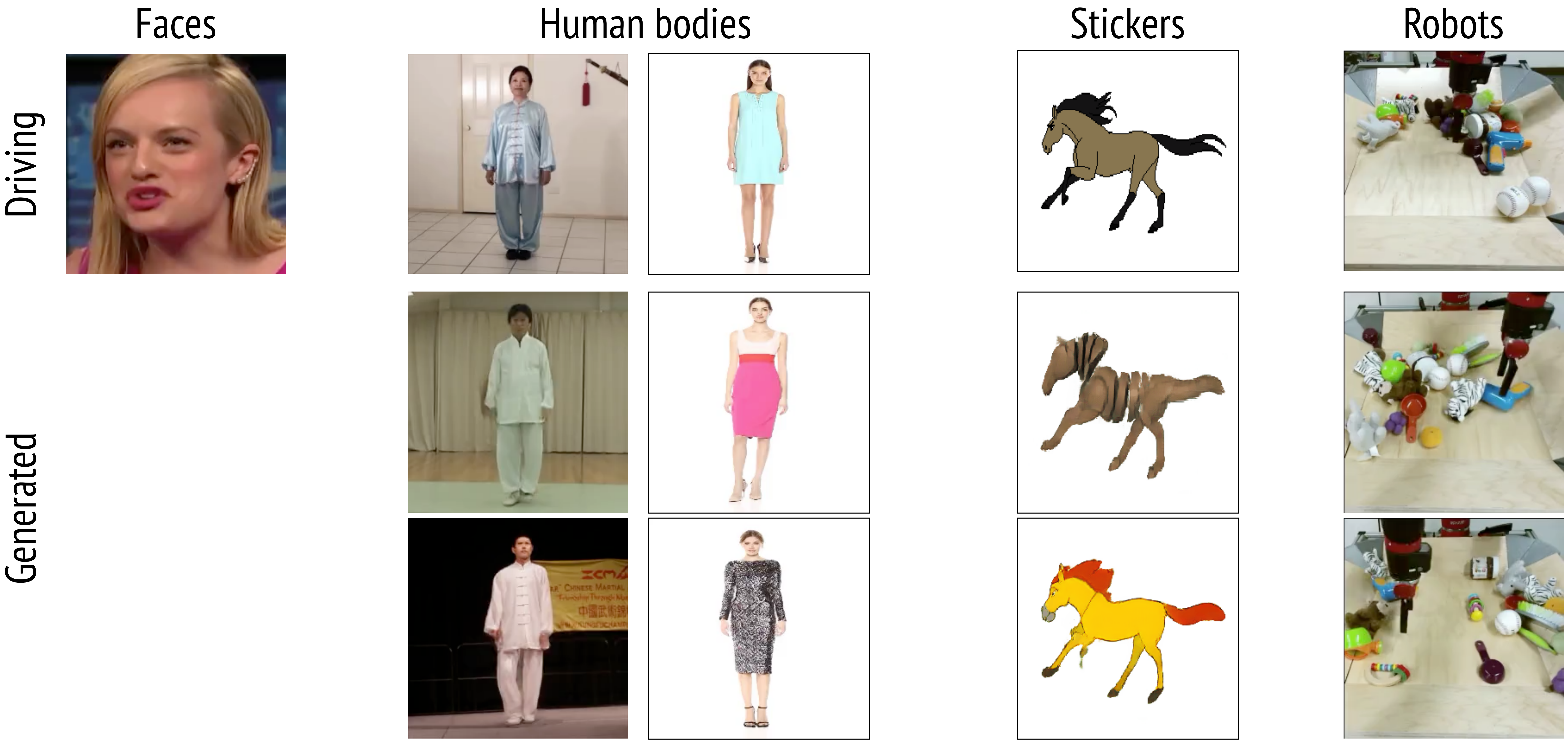
2000 years ago to create you need to have money and power

2000 Years Later

Less photorealistic, but serves the purpose very well

Video credit: https://www.youtube.com/watch?v=AaNjPAFD-UI

Present Day



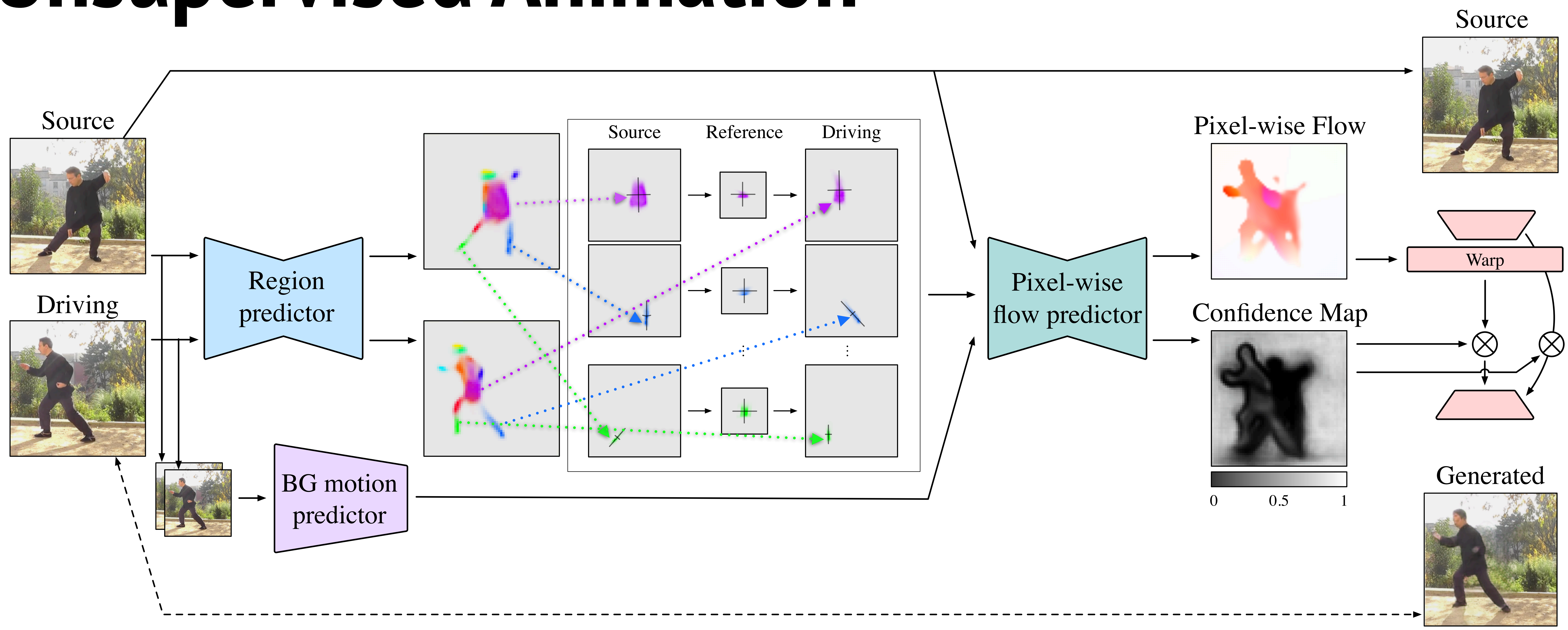
Each video is produced using a single input image

What is Needed for Animation?

- Location → Unsupervised Keypoints
- Orientation → Local Affine Transformations
- Missing content → Inpainting

Siarohin, Aliaksandr, et al. "**First order motion model for image animation.**" NeurIPS'2019

Unsupervised Animation



Siarohin et al. "**Motion Representations for Articulated Animation**" CVPR'2021

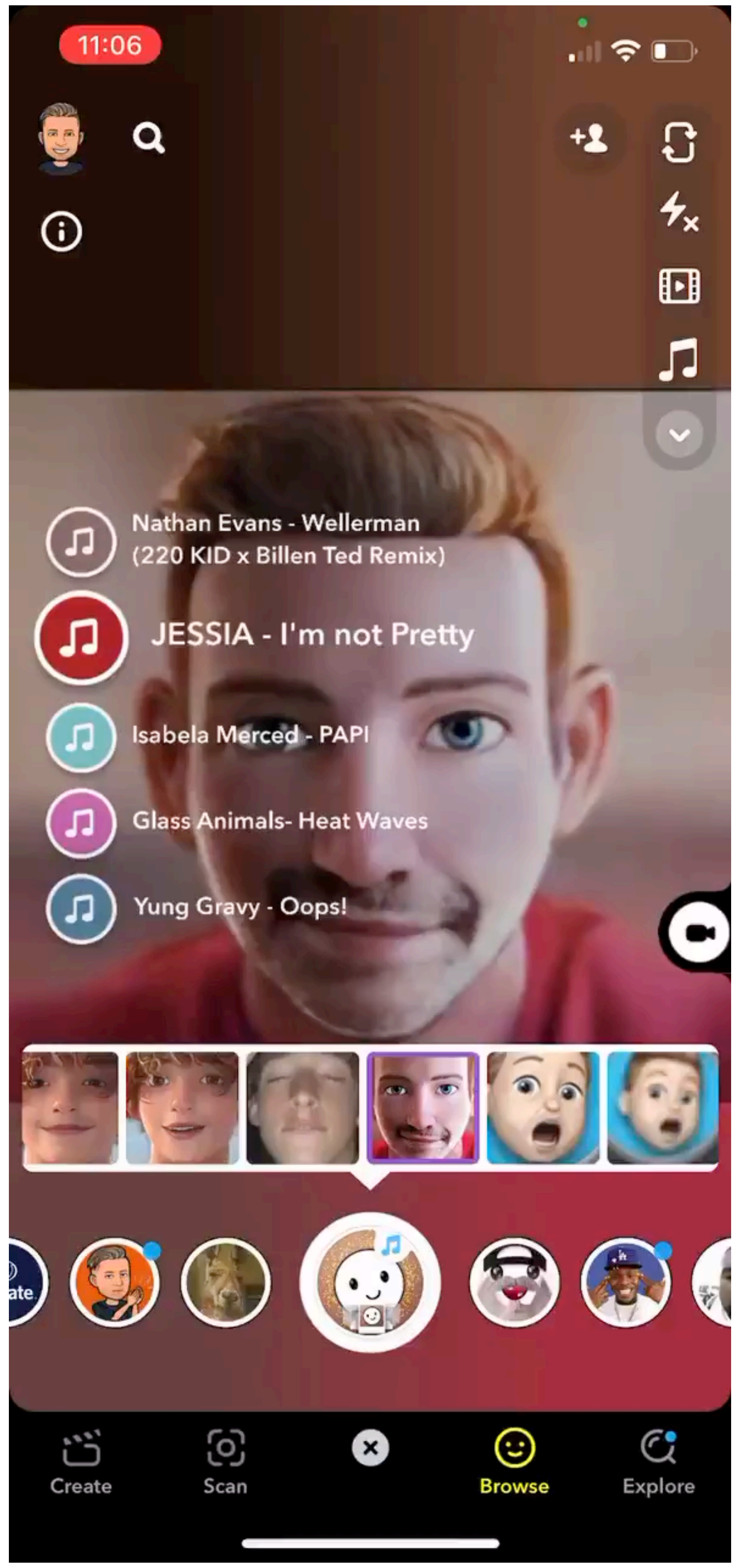
Siarohin et al. "**First order motion model for image animation.**" NeurIPS'2019

Unsupervised Regions

Siarohin et al. "**Motion Representations for Articulated Animation**" CVPR'2021

Siarohin et al. "**First order motion model for image animation.**" NeurIPS'2019

Magic Karaoke Lens on Snapchat




AppName
File
Edit
Format
View
Window
Help
Tue 6:37 PM
Y2
Good1
dappyboi
2
DIRECTED BY
YASHA
IGHT SO
boom

Animating Bodies

Siarohin, Aliaksandr, et al. "**Motion Representations for Articulated Animation.**" CVPR'21

FAST COMPANY

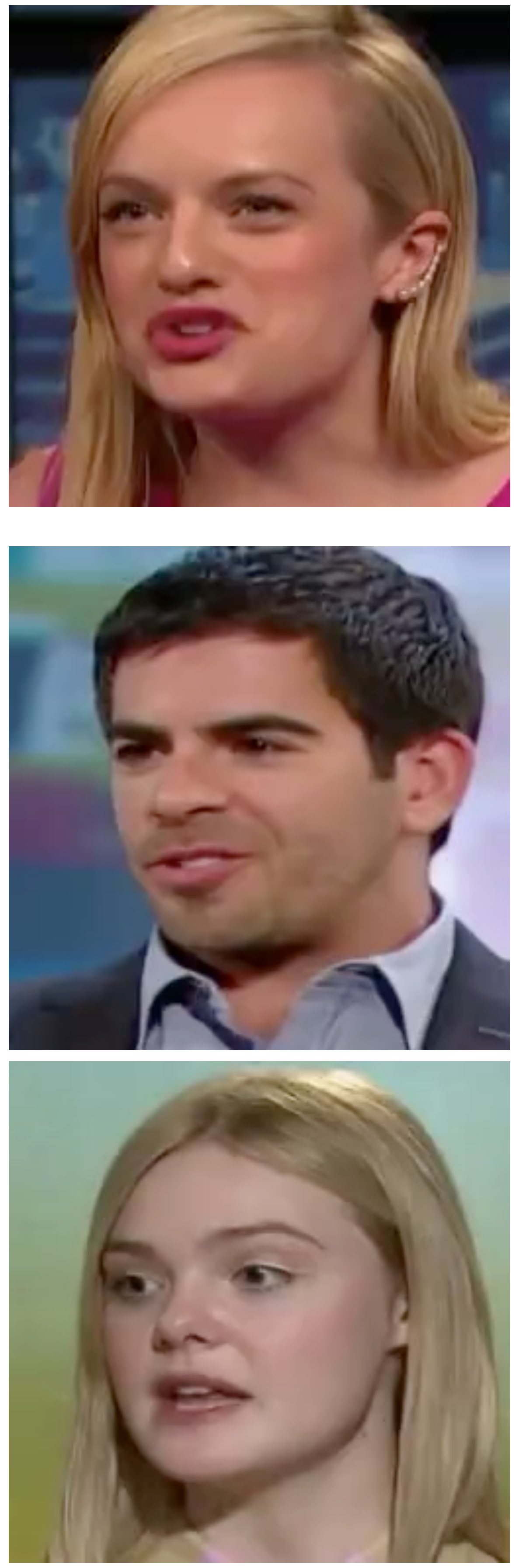
11-16-22

Exclusive: Try on your team's World Cup jersey with Snap's new AR filter

Snap's new Lens is a window into the future of augmented reality.

[Image: Snap Inc]

2D Animation

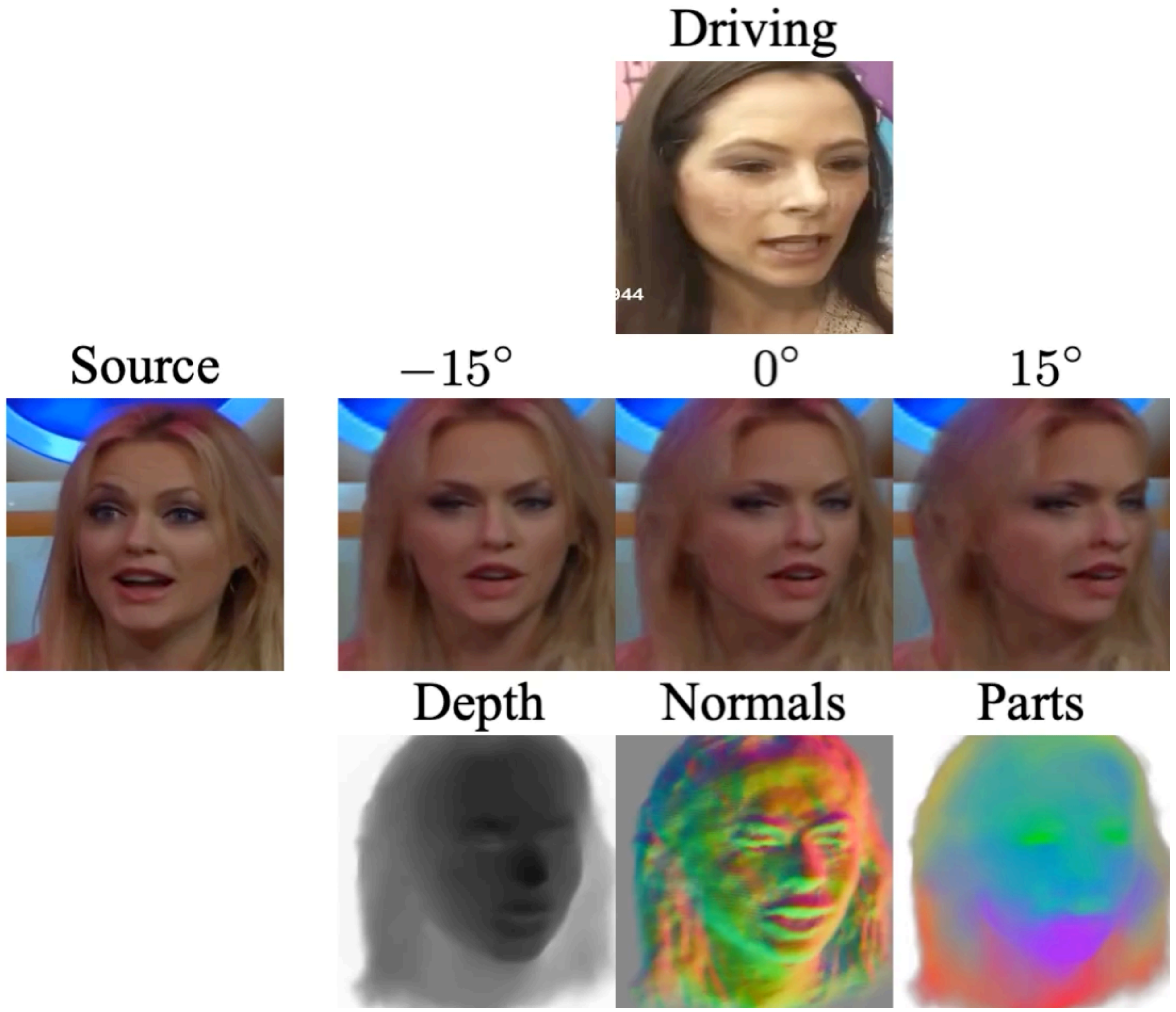


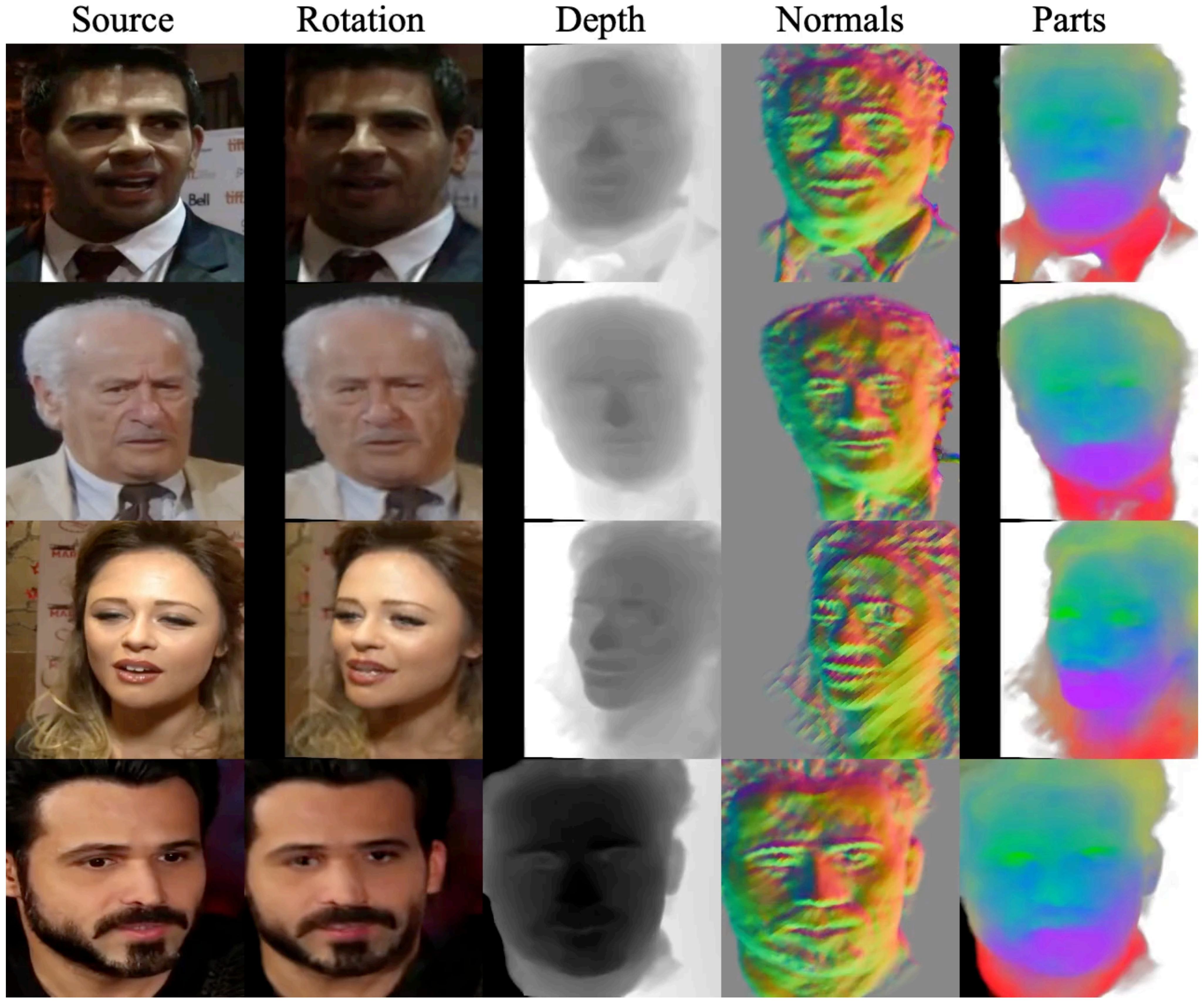
Extreme pose changes are hard to model in 2D

3D Animation is needed!

- Convolutional neural networks work very well with images and contain 2D inductive bias
- **Challenge:** how can we introduce inductive bias for 3D?
  - Animate Radiance Fields?
  - How to reuse 2D bias of 2D CNNs?

UVA: Unsupervised Volumetric Animation

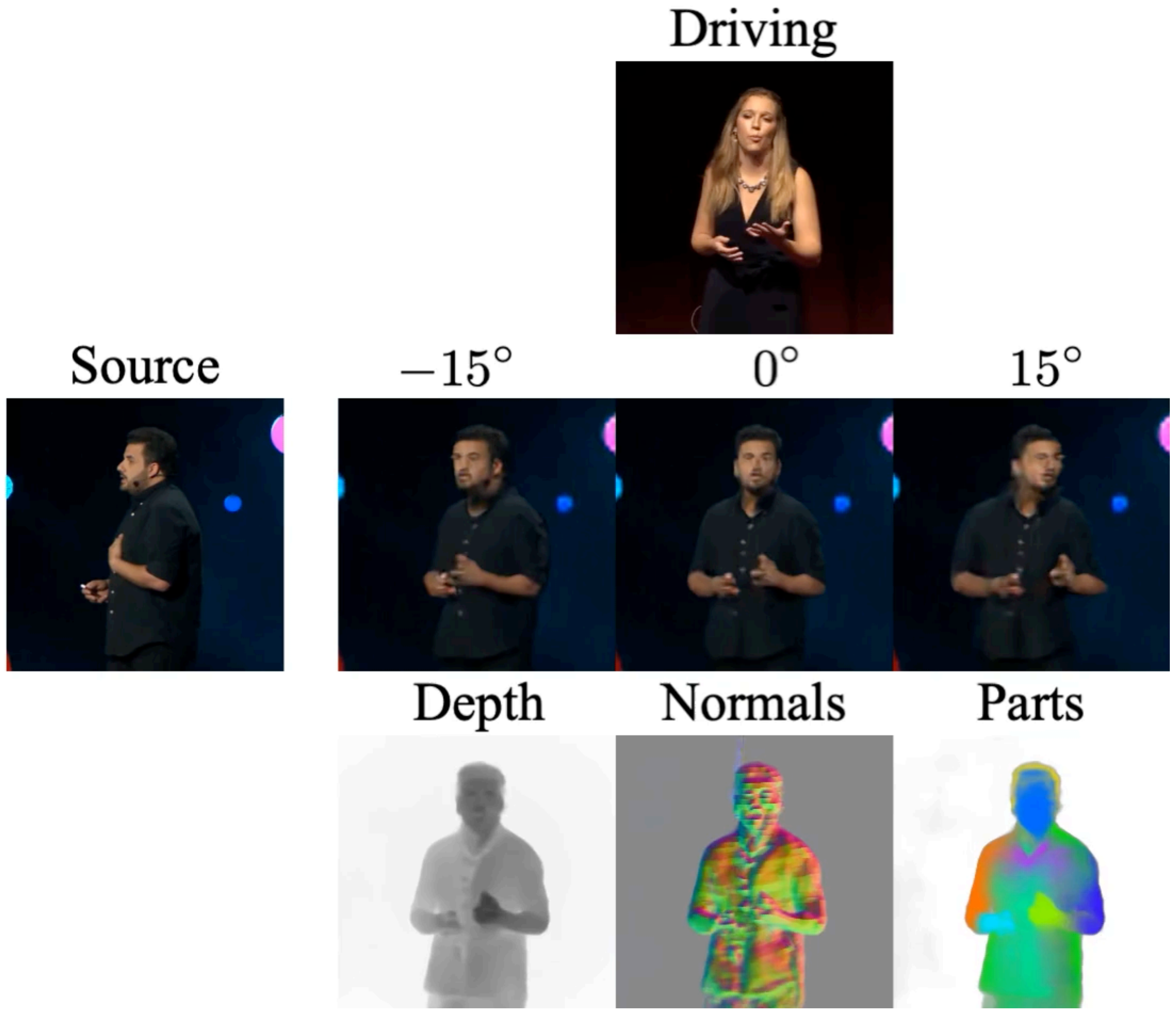


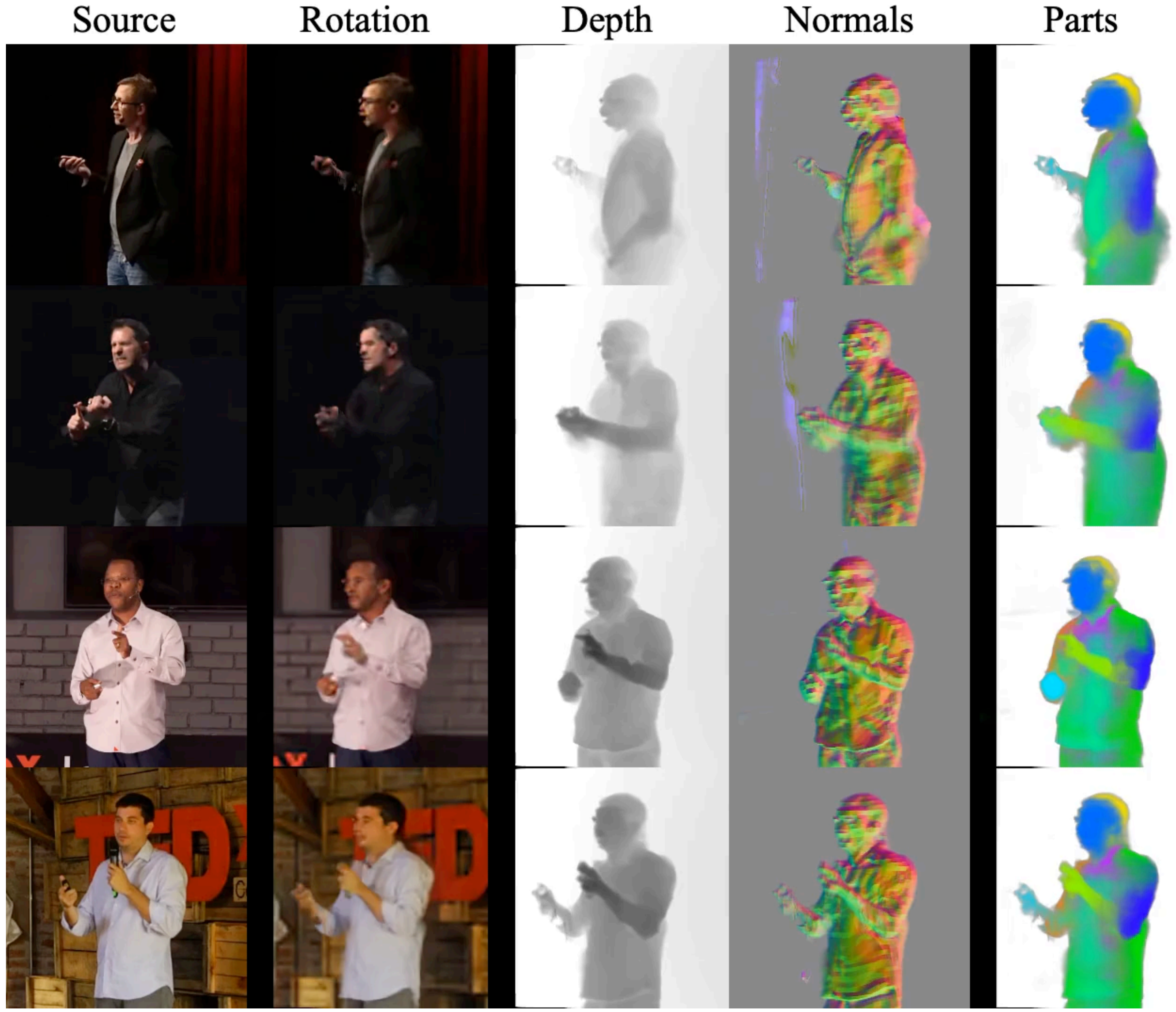


Trained on monocular videos only, UVA reconstructs 3D shape, pose, and articulation parameters

Siarohin et al. "**Unsupervised Volumetric Animation**" CVPR'2023

UVA: Unsupervised Volumetric Animation

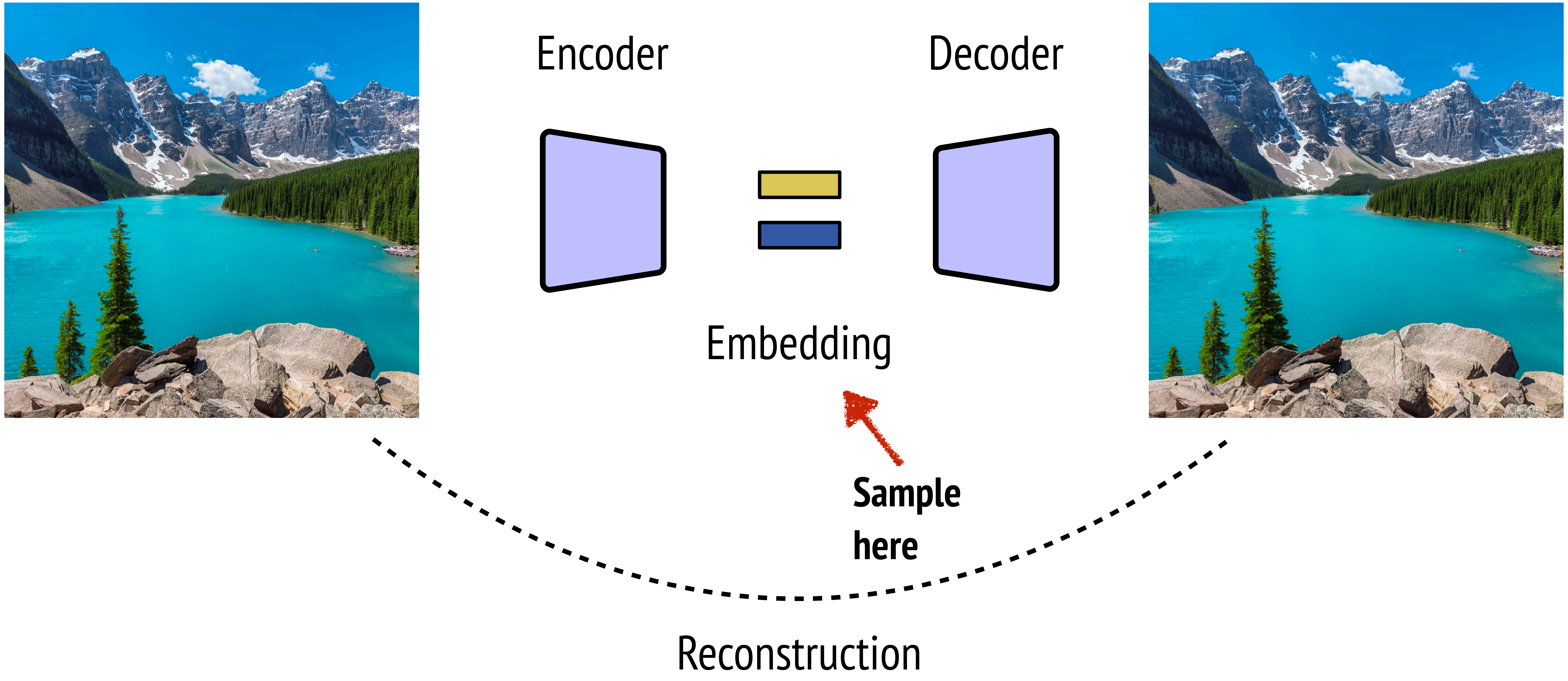




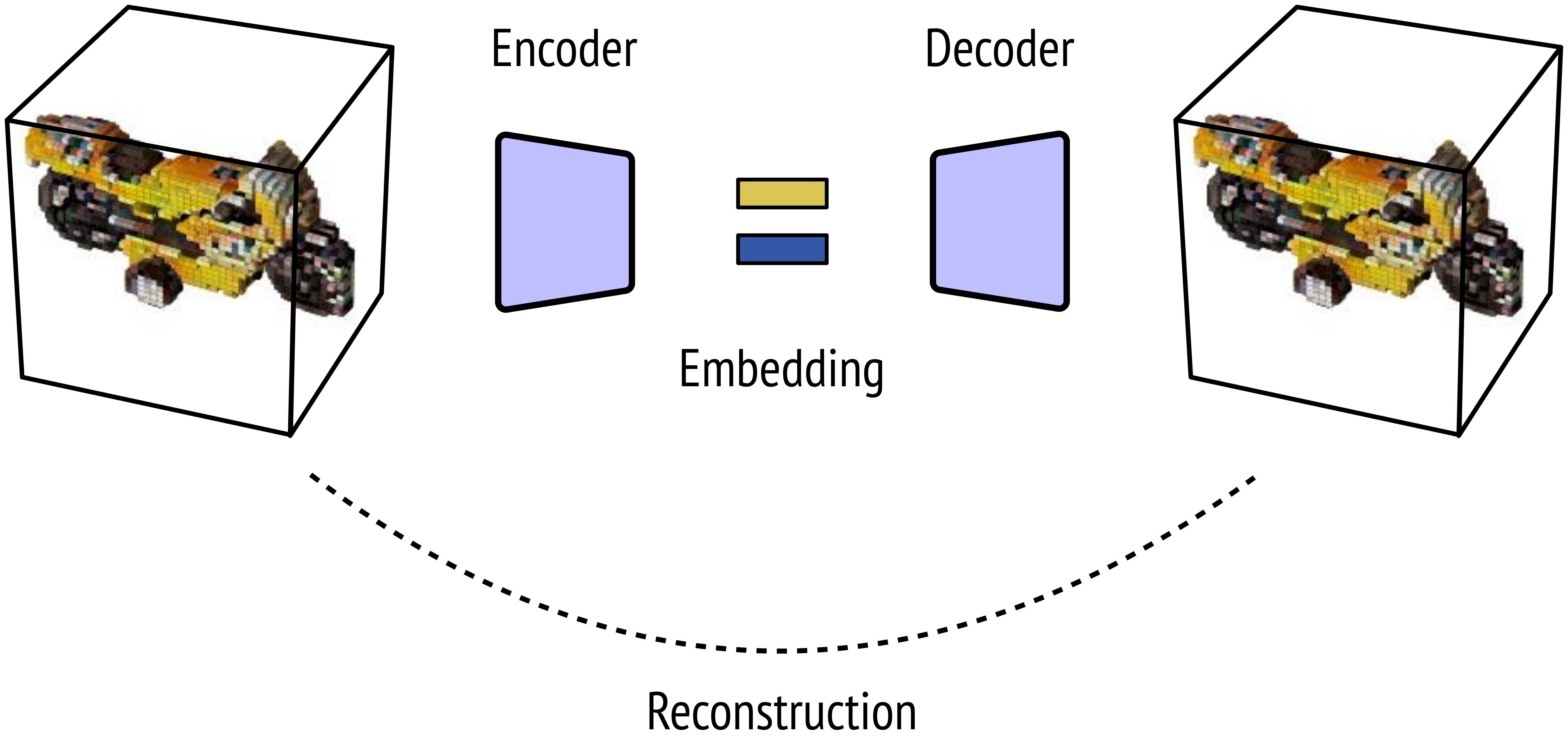
Trained on monocular videos only, UVA reconstructs 3D shape, pose, and articulation parameters

Siarohin et al. "**Unsupervised Volumetric Animation**" CVPR'2023

Generating Images: Reconstruct-then-sample

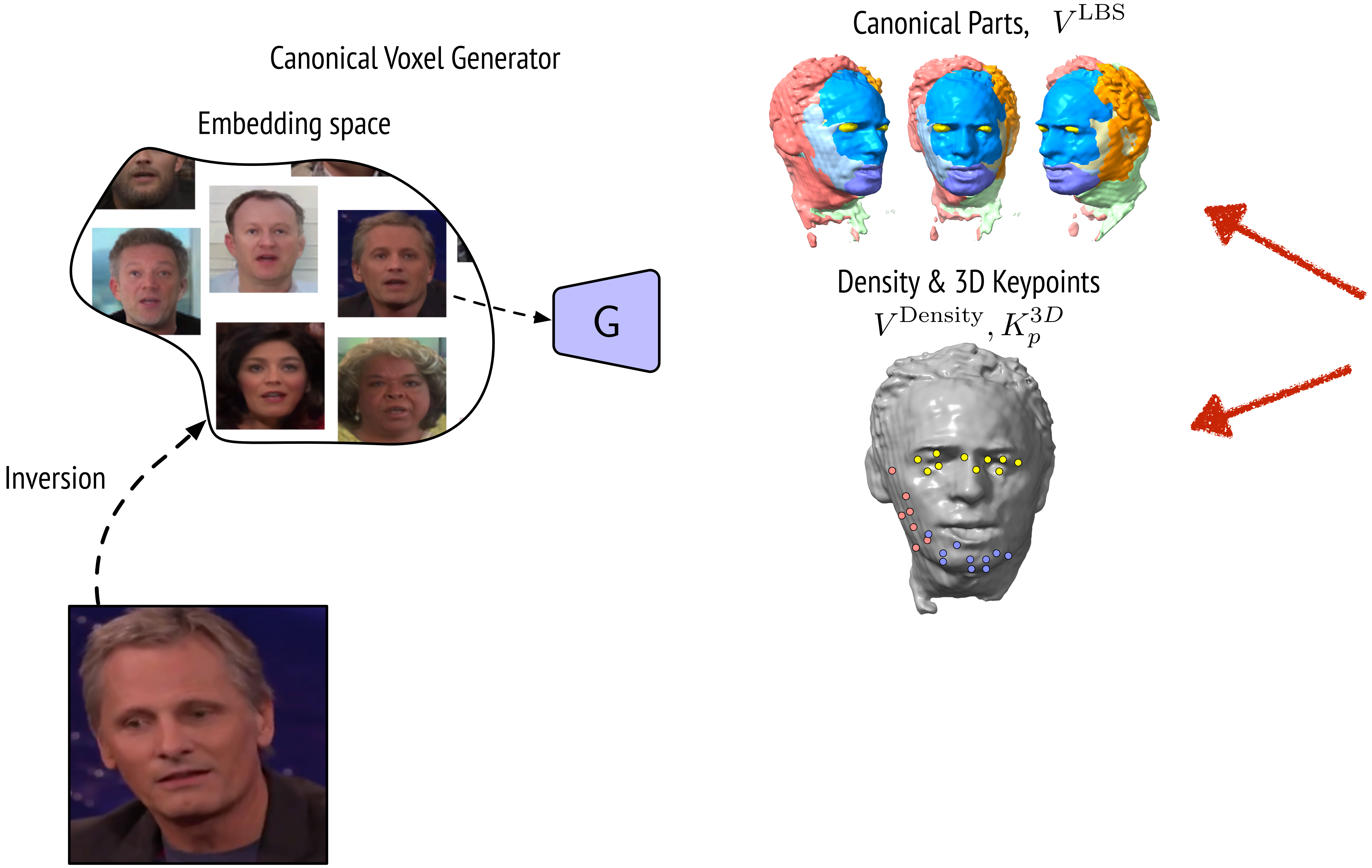


Auto-Encoding 3D Objects?



There is just not enough data

UVA: Canonical Generator

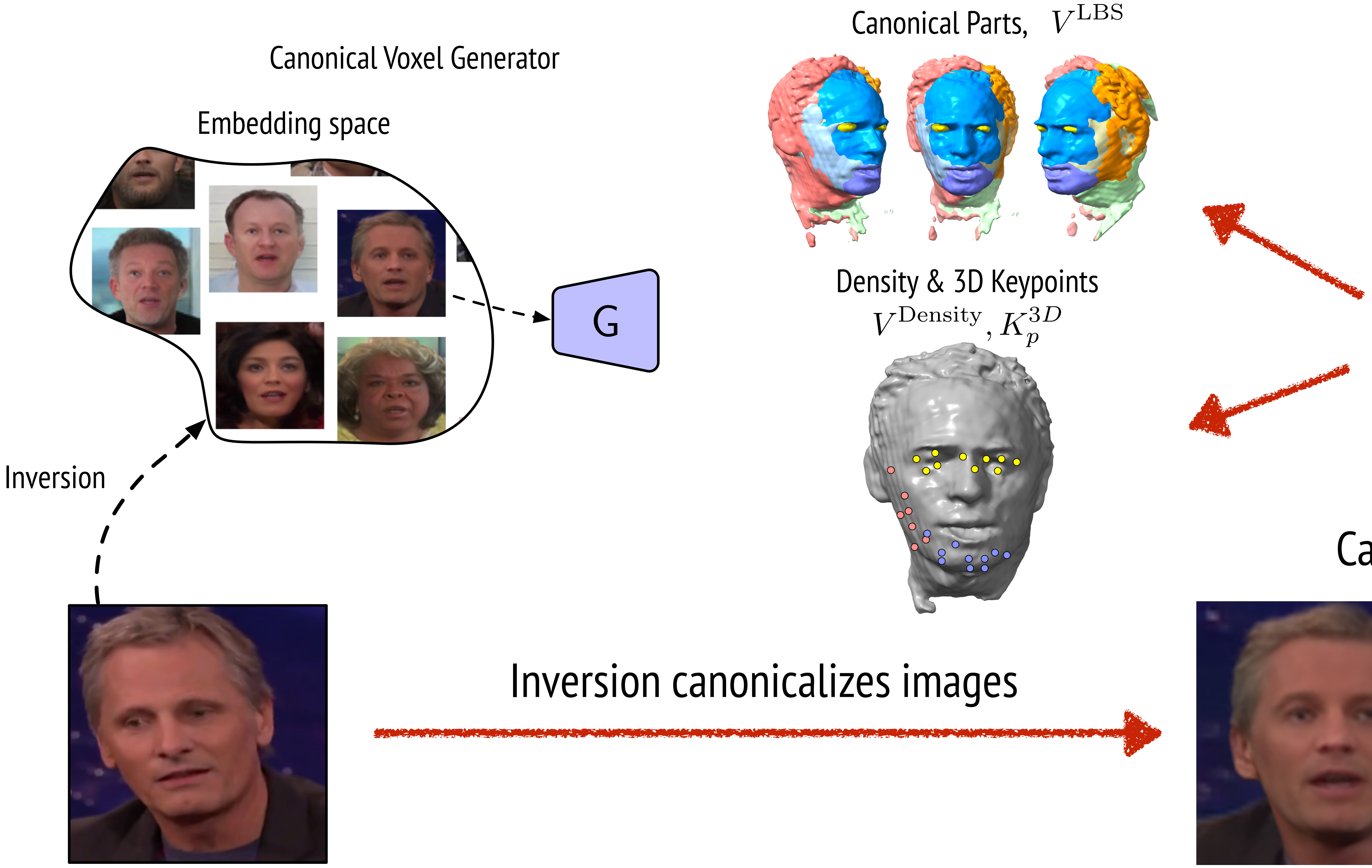


Pose, expression, size is the same for each object

G generates several things in <u>canonical</u> space

- Density and color, as before in NeRFs
- Learned 3D keypoints
- Learned 3D parts

UVA: Canonical Generator

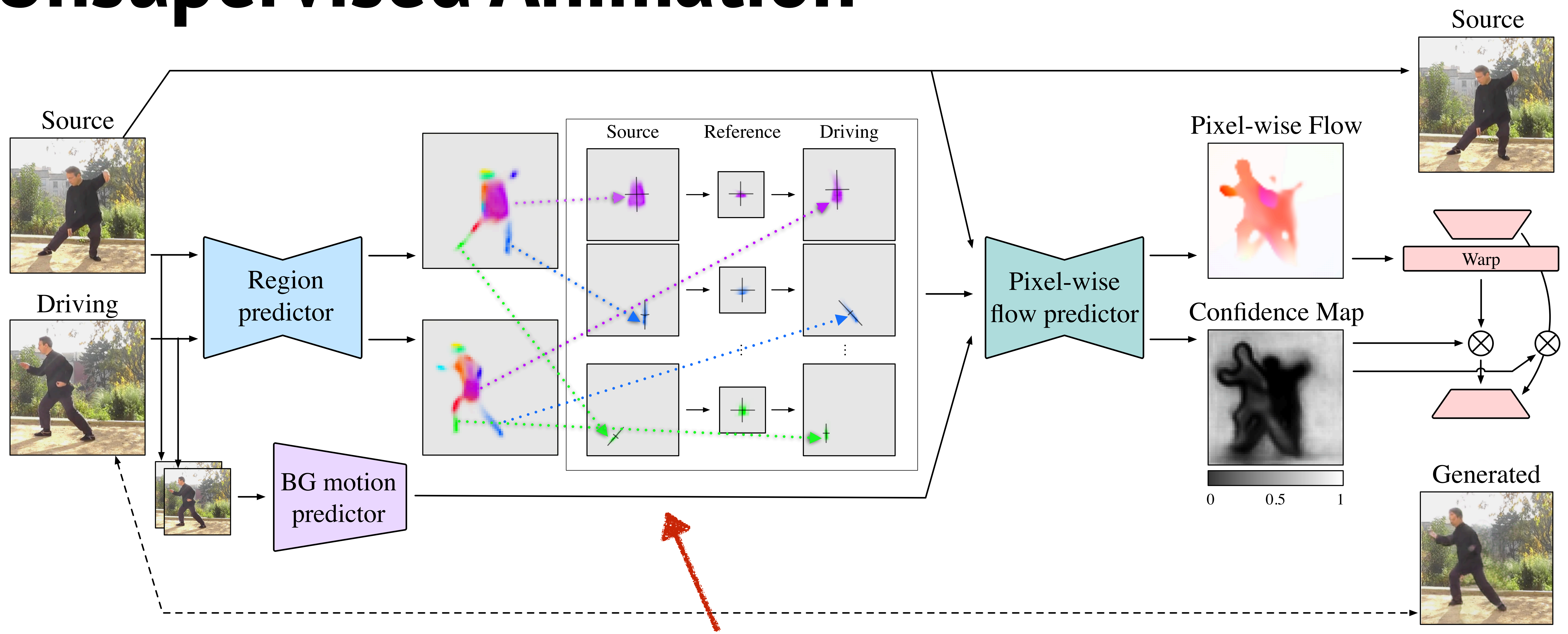


G generates several things in <u>canonical</u> space

- Density, as before
- Learned 3D keypoints
- Learned 3D parts

Canonical volumes in canonical pose and camera

Unsupervised Animation

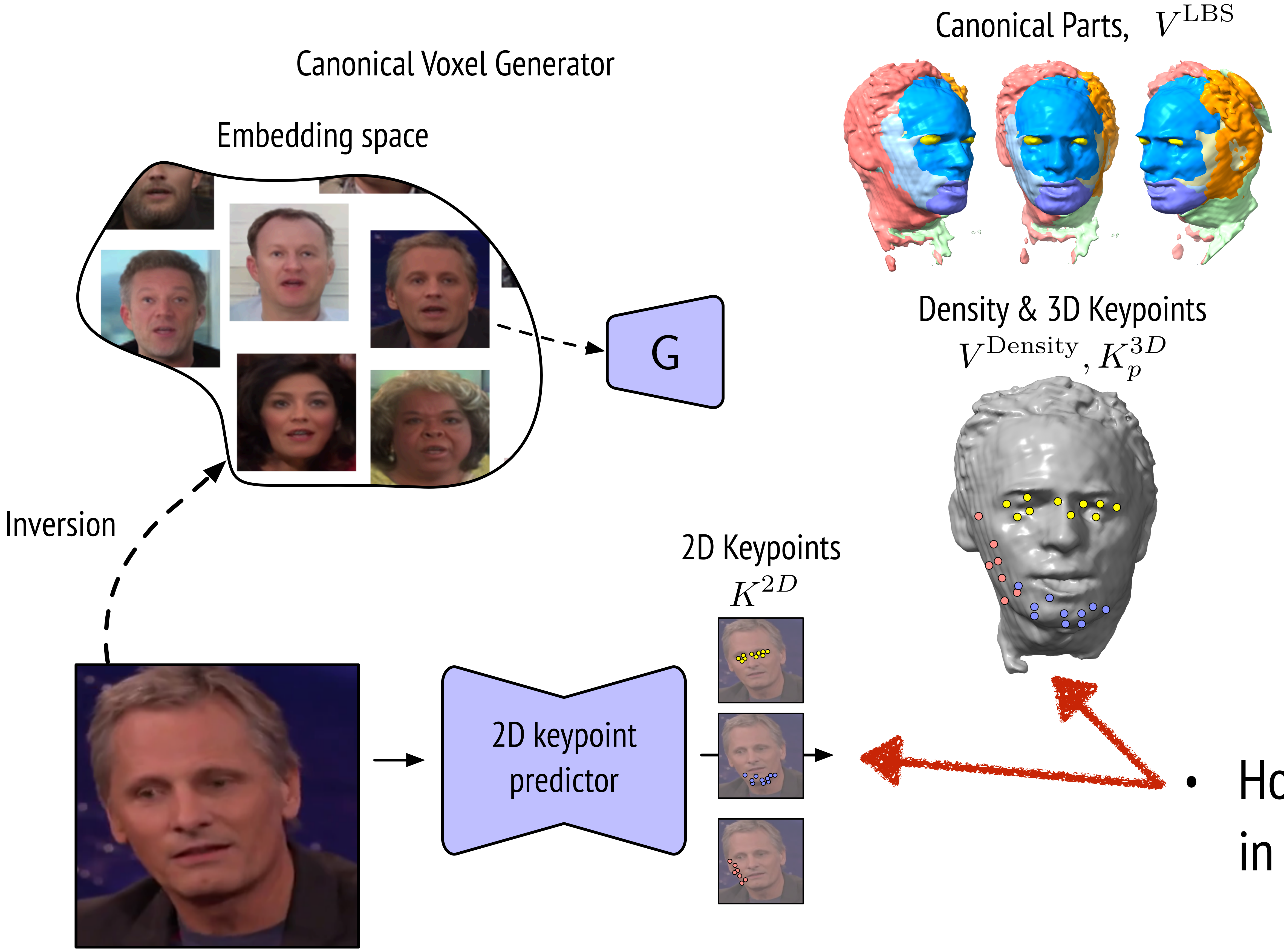


Unsupervised parts, learned automatically during training

Siarohin et al. "**Motion Representations for Articulated Animation**" CVPR'2021

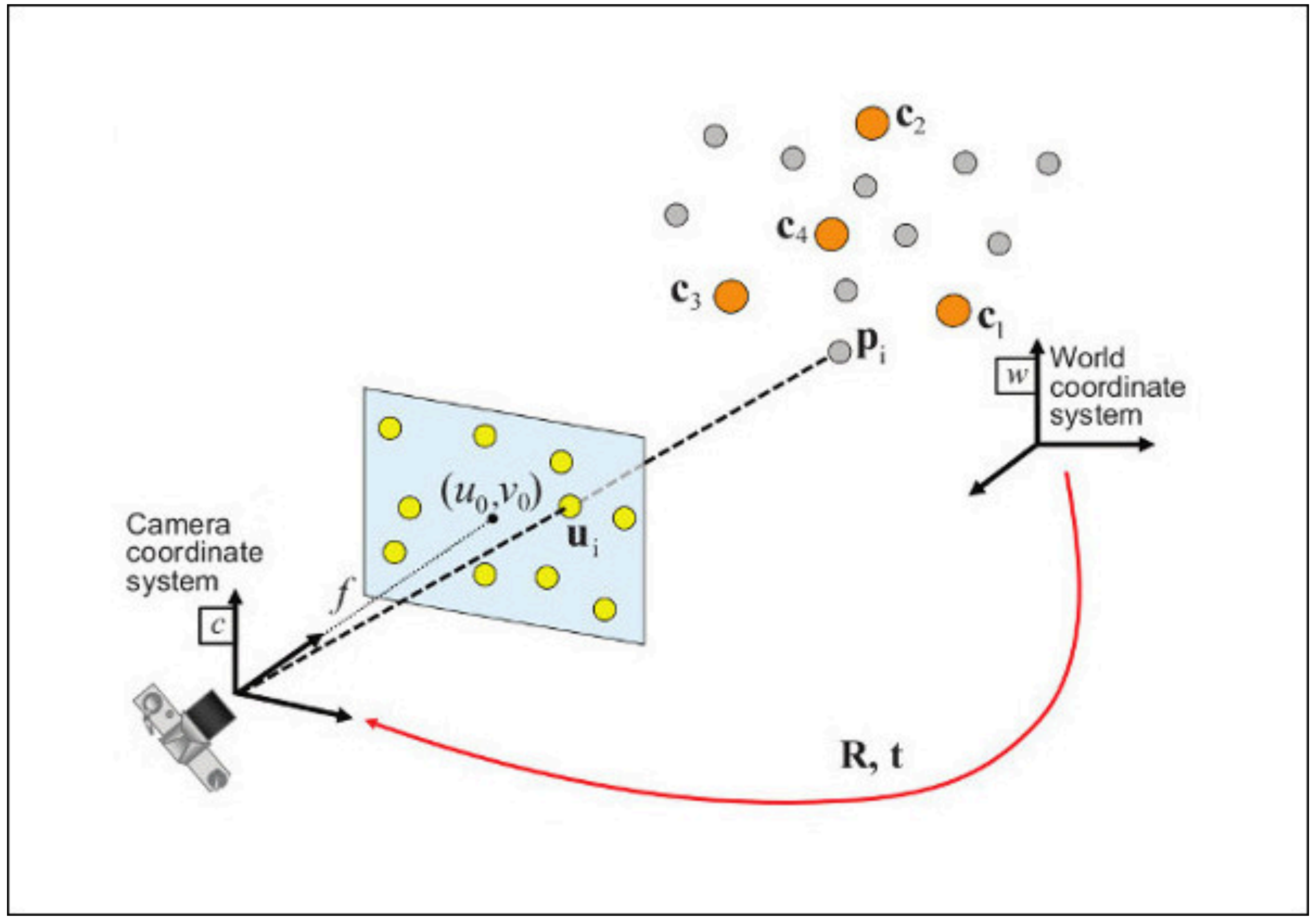
Siarohin et al. "**First order motion model for image animation.**" NeurIPS'2019

UVA: Animation & Rendering



- Given canonical 3D points and their 2D counterparts find camera pose
- How to connect 2D and 3D in a differentiable manner?

Perspective-n-Point Algorithm



Intrinsics — Projection — Rotation — Translation

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix}$$

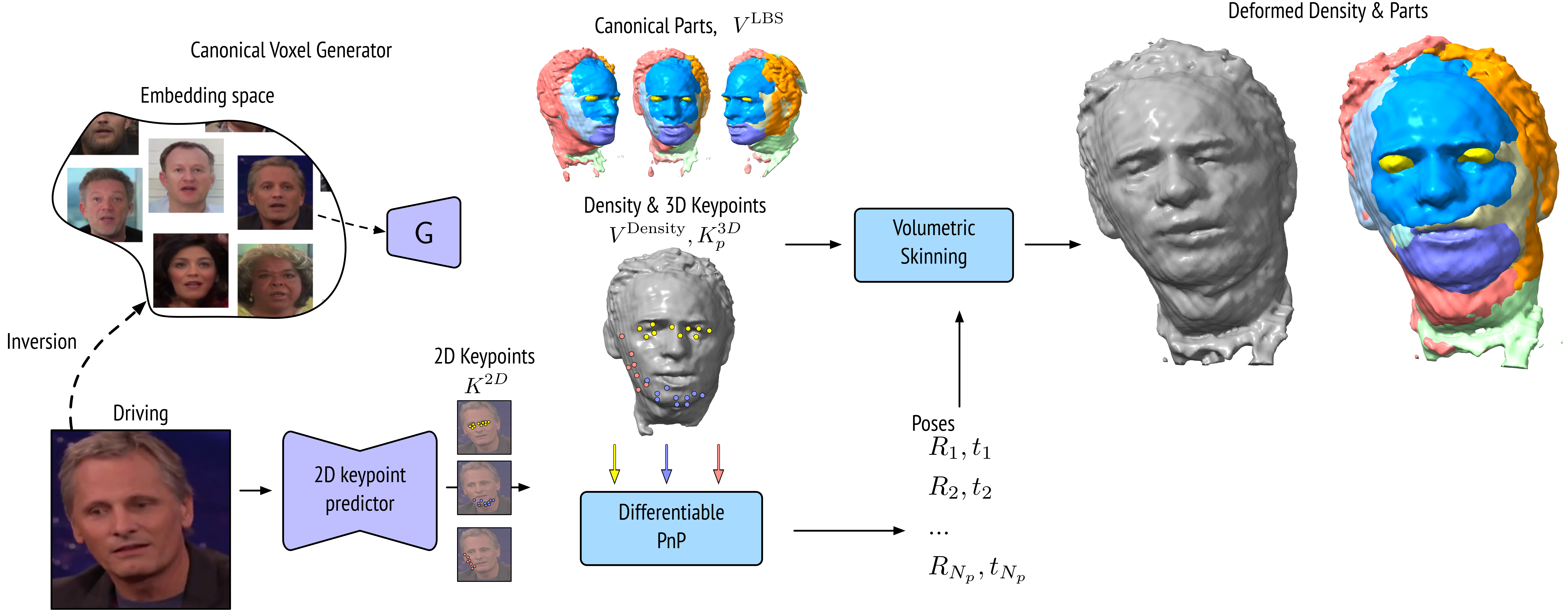
Points in the camera plane

3D points in the world coordinate system

Given points in the world coordinates and their 2D projections, find the camera pose

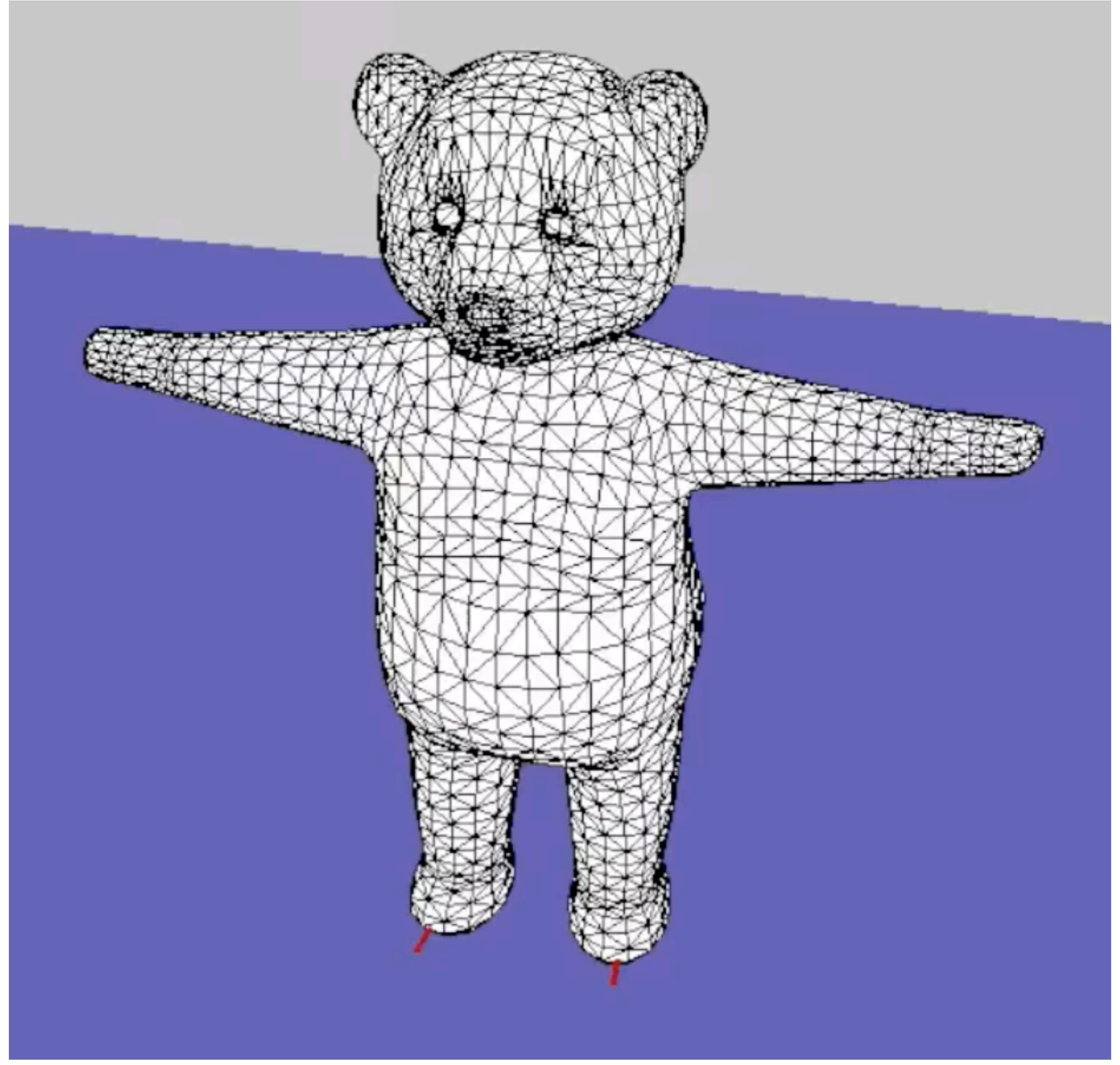
Lepetit et al., IJCV'2009

UVA: Animation & Rendering

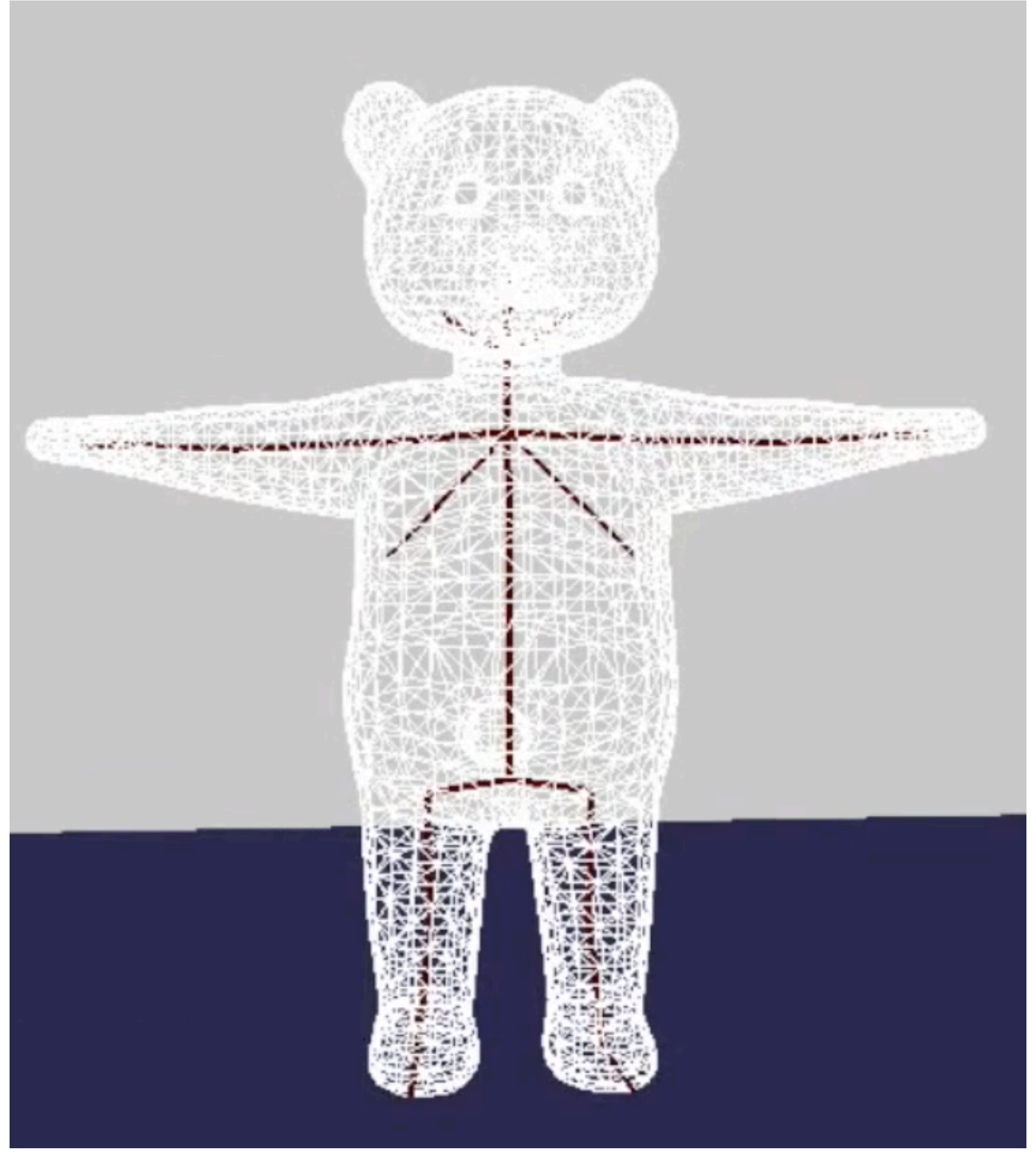


Volumetric Skinning

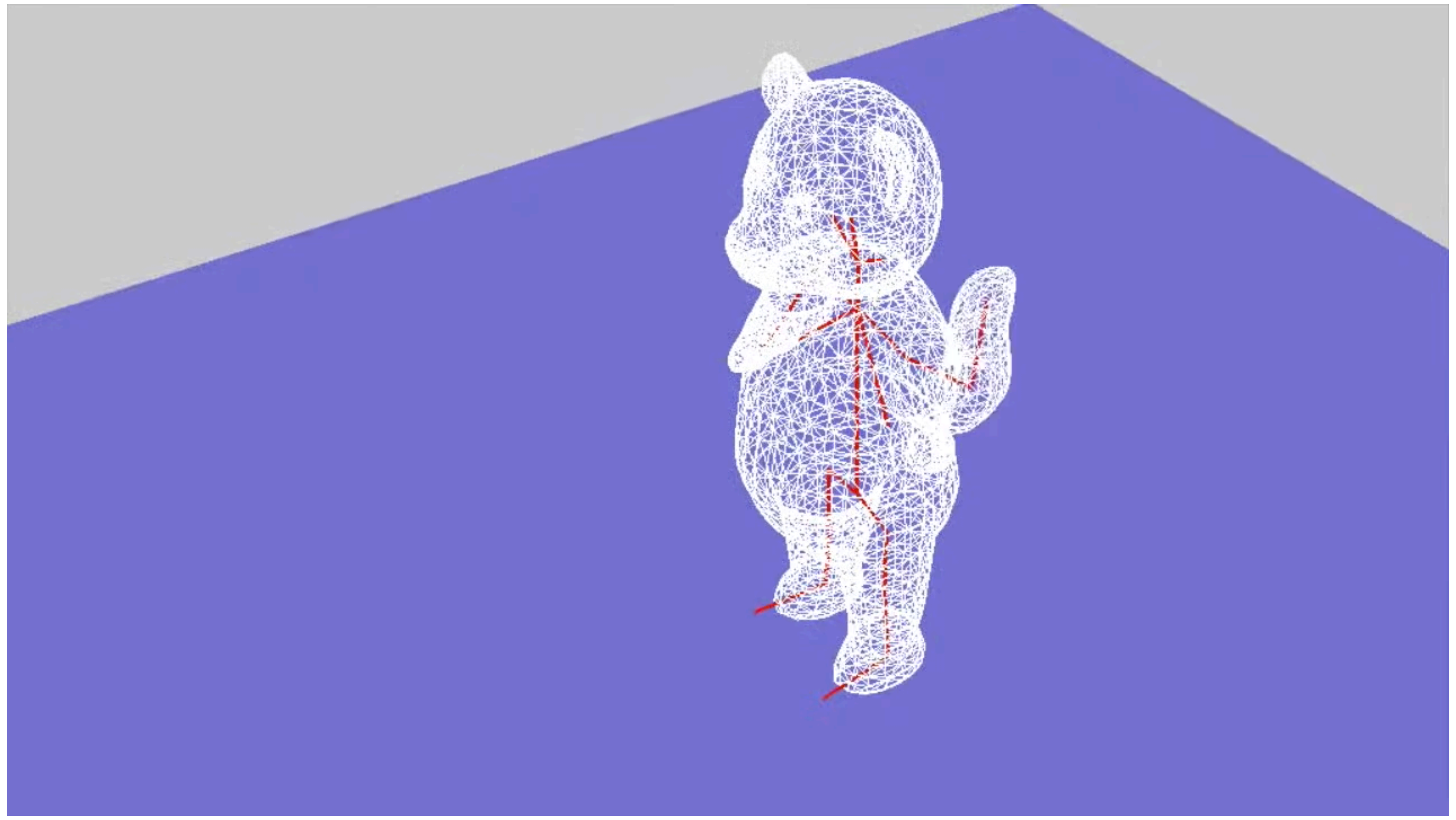
Canonical Pose



Bones



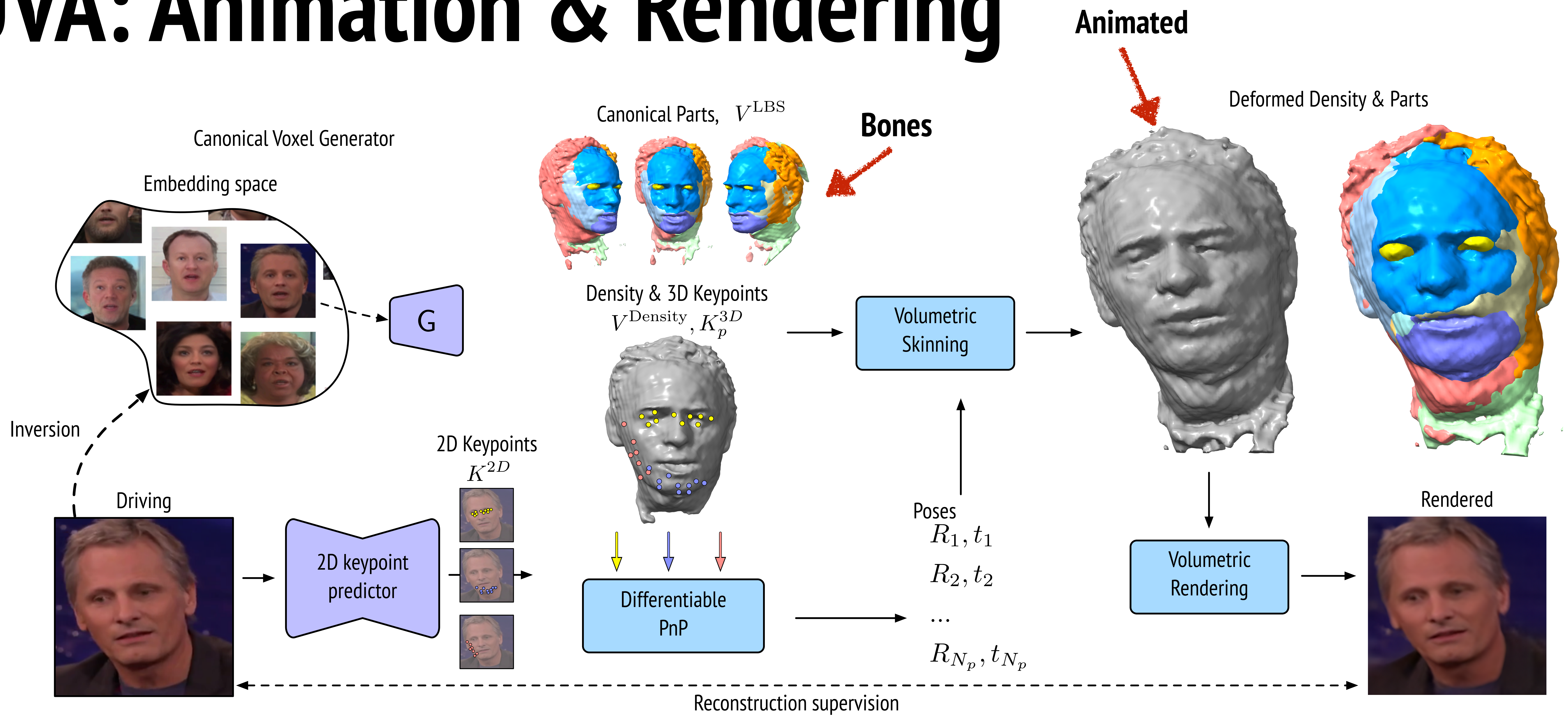
Animation & Skinning



Each vertex on the mesh moves according to the motion of the closest bones

video credit: https://www.youtube.com/watch?v=GPFUVgy-7DE

UVA: Animation & Rendering

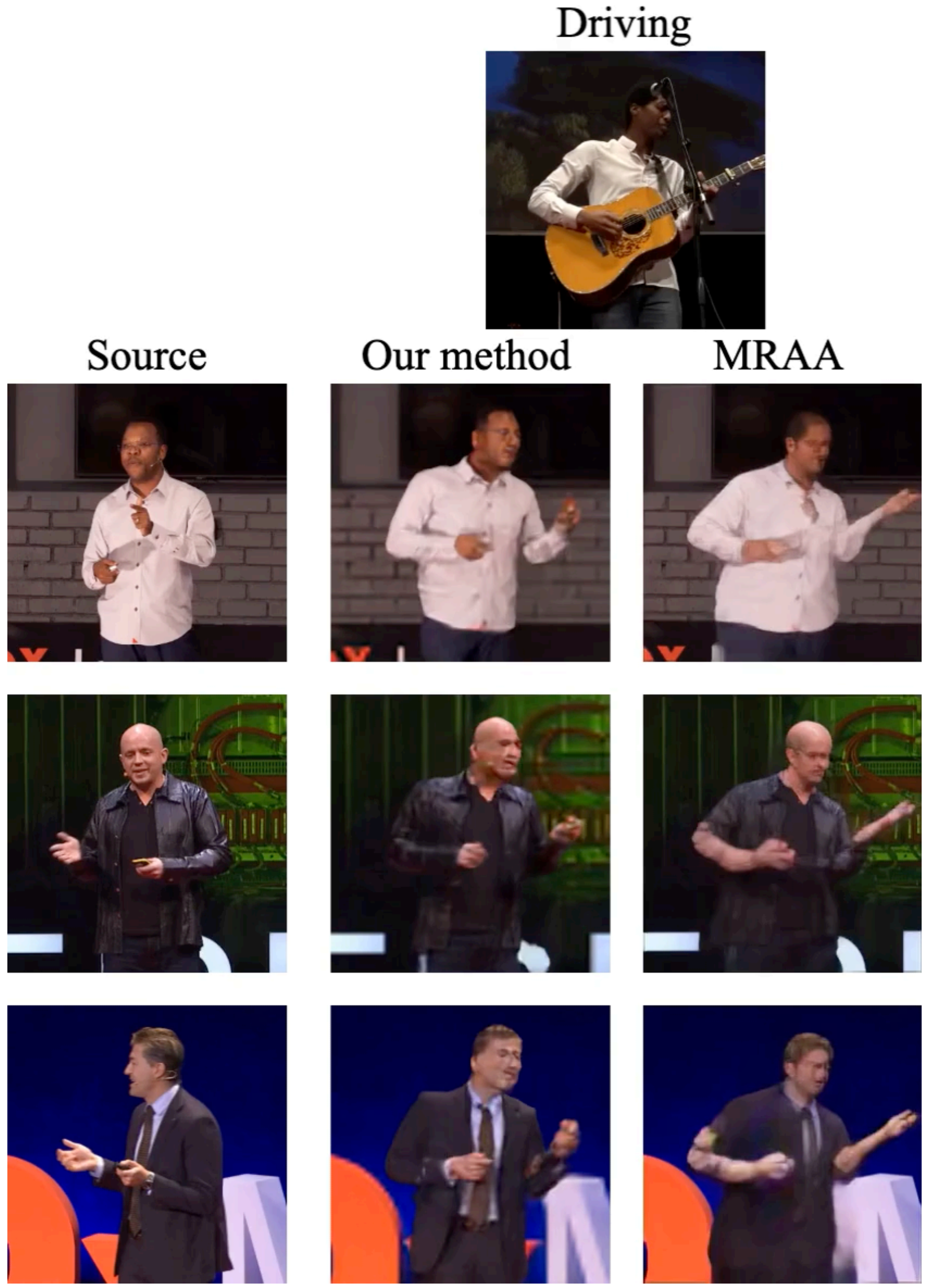


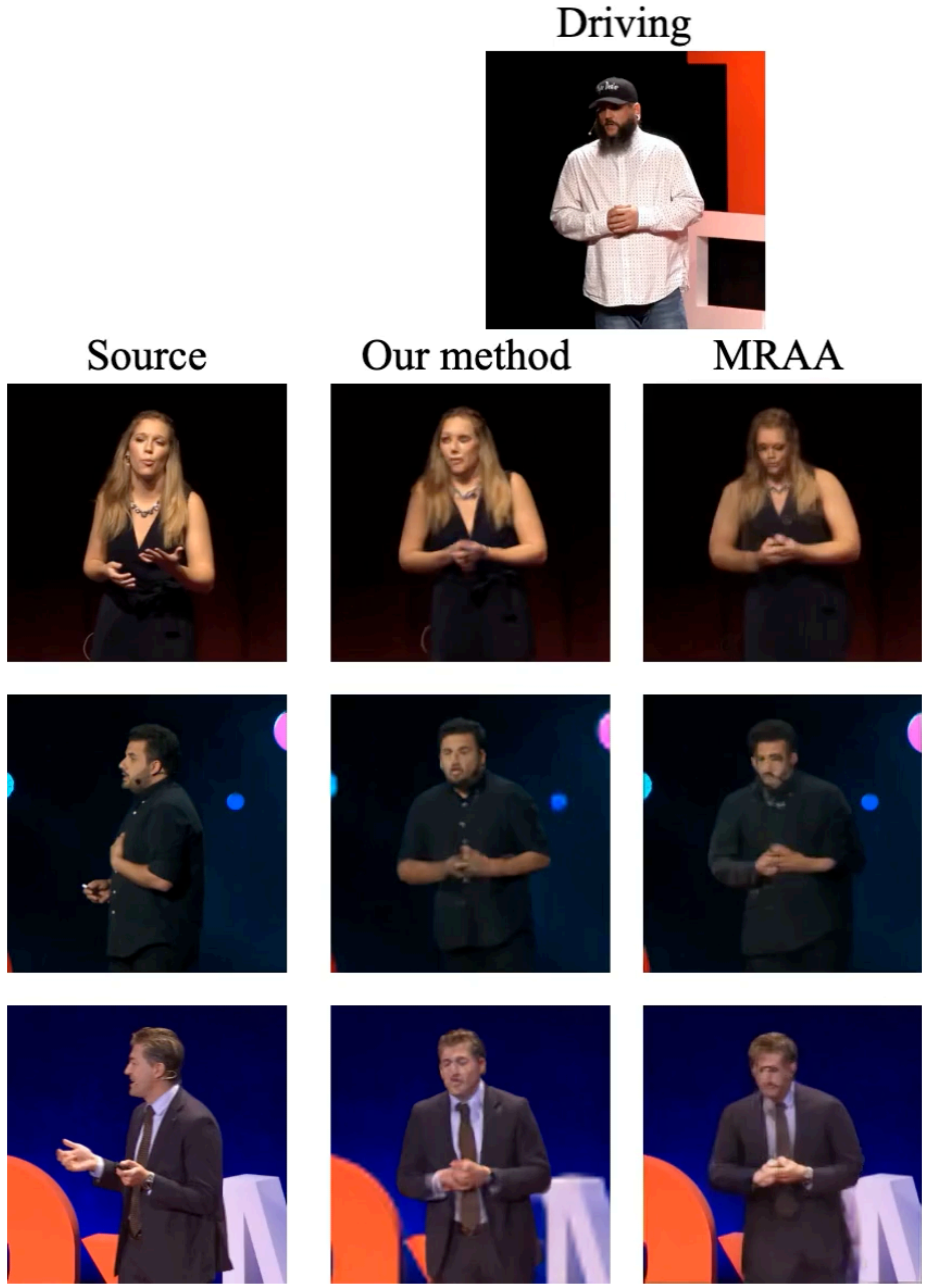
We don't need no supervision (apart from reconstruction)

2D vs 3D Animation: Faces

3D animation better preserves identity, better animates, shows consistent rotations

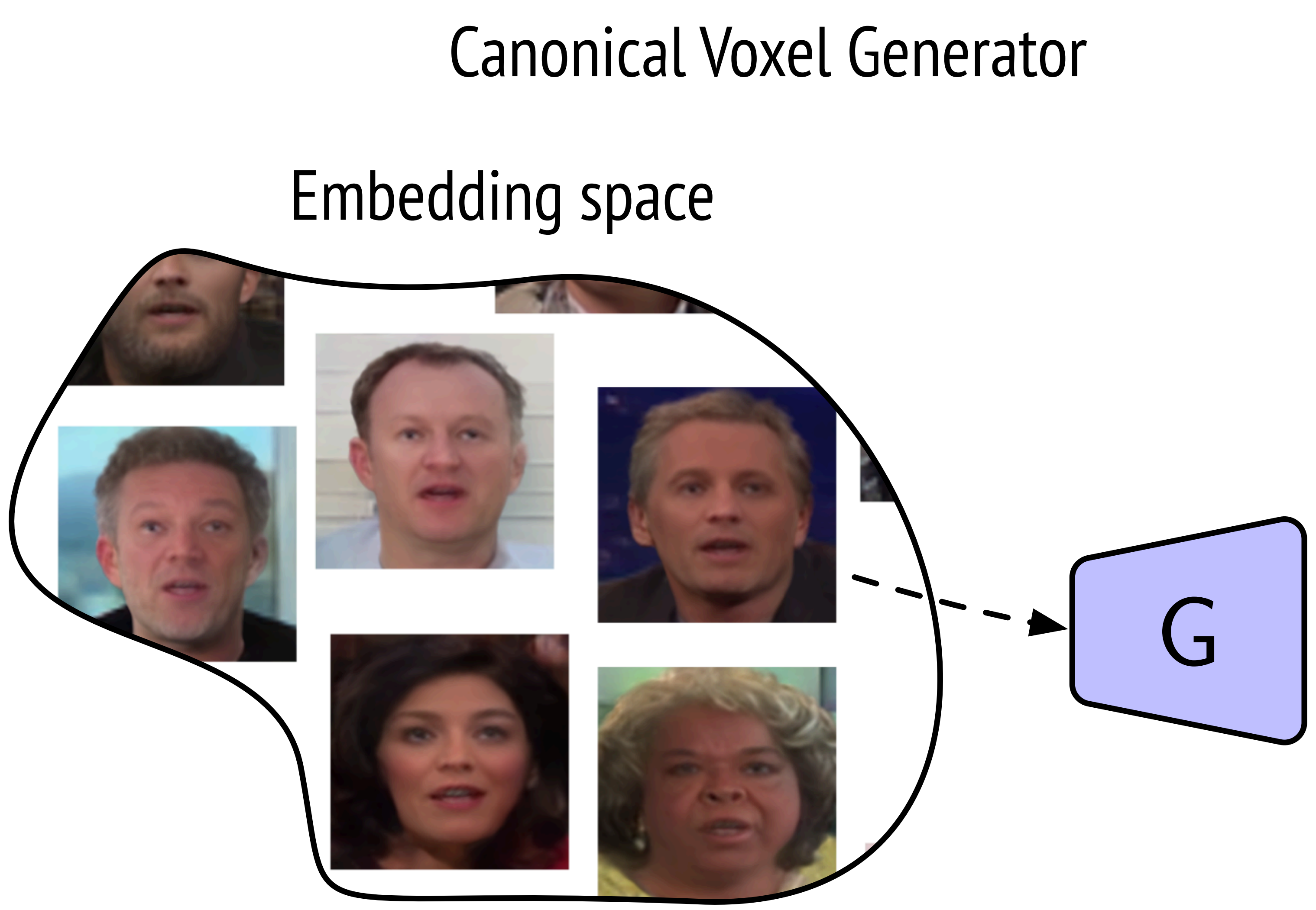
2D vs 3D Animation: Bodies



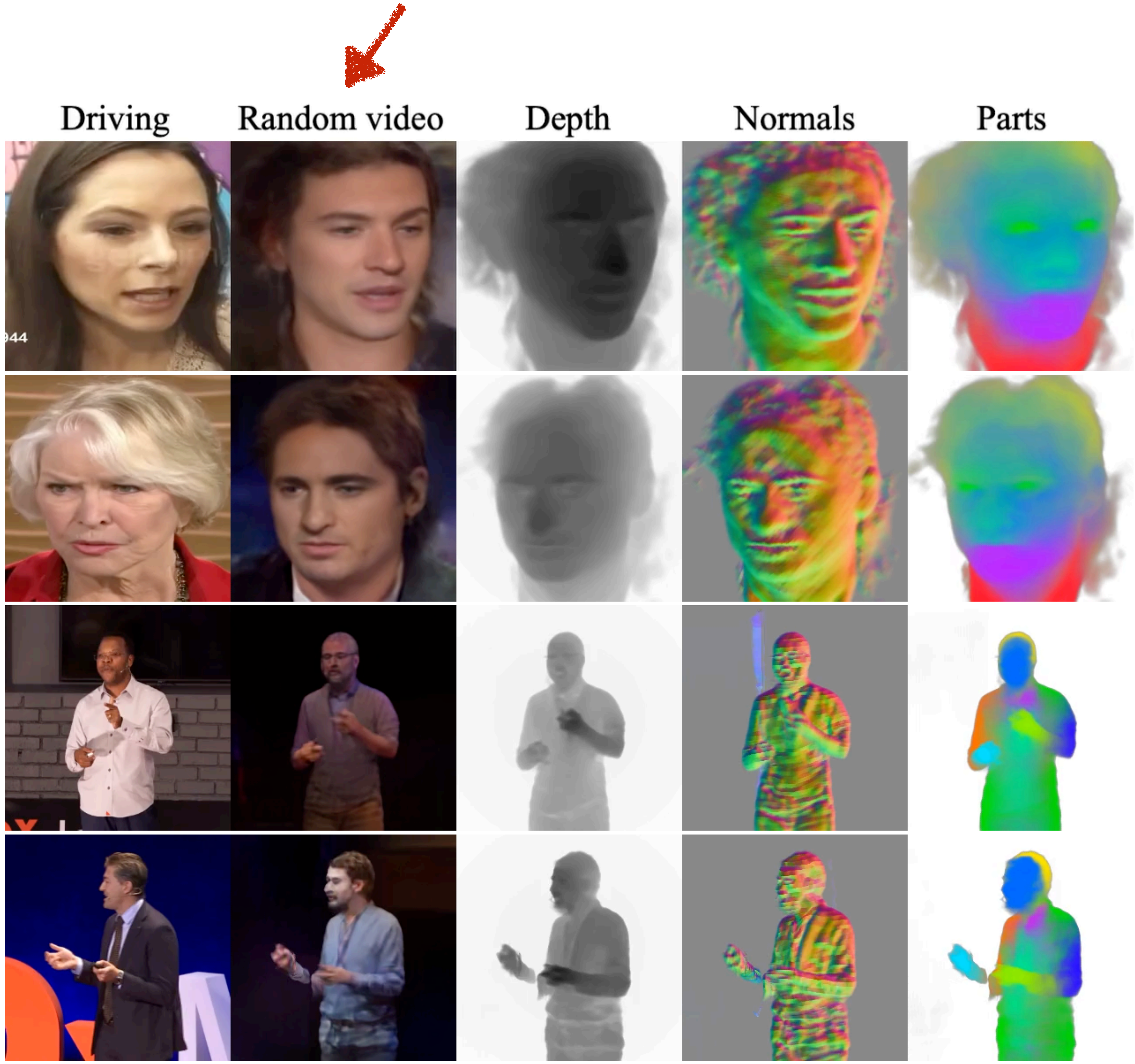


3D animation better preserves identity, better animates, shows consistent rotations

3D Video Synthesis



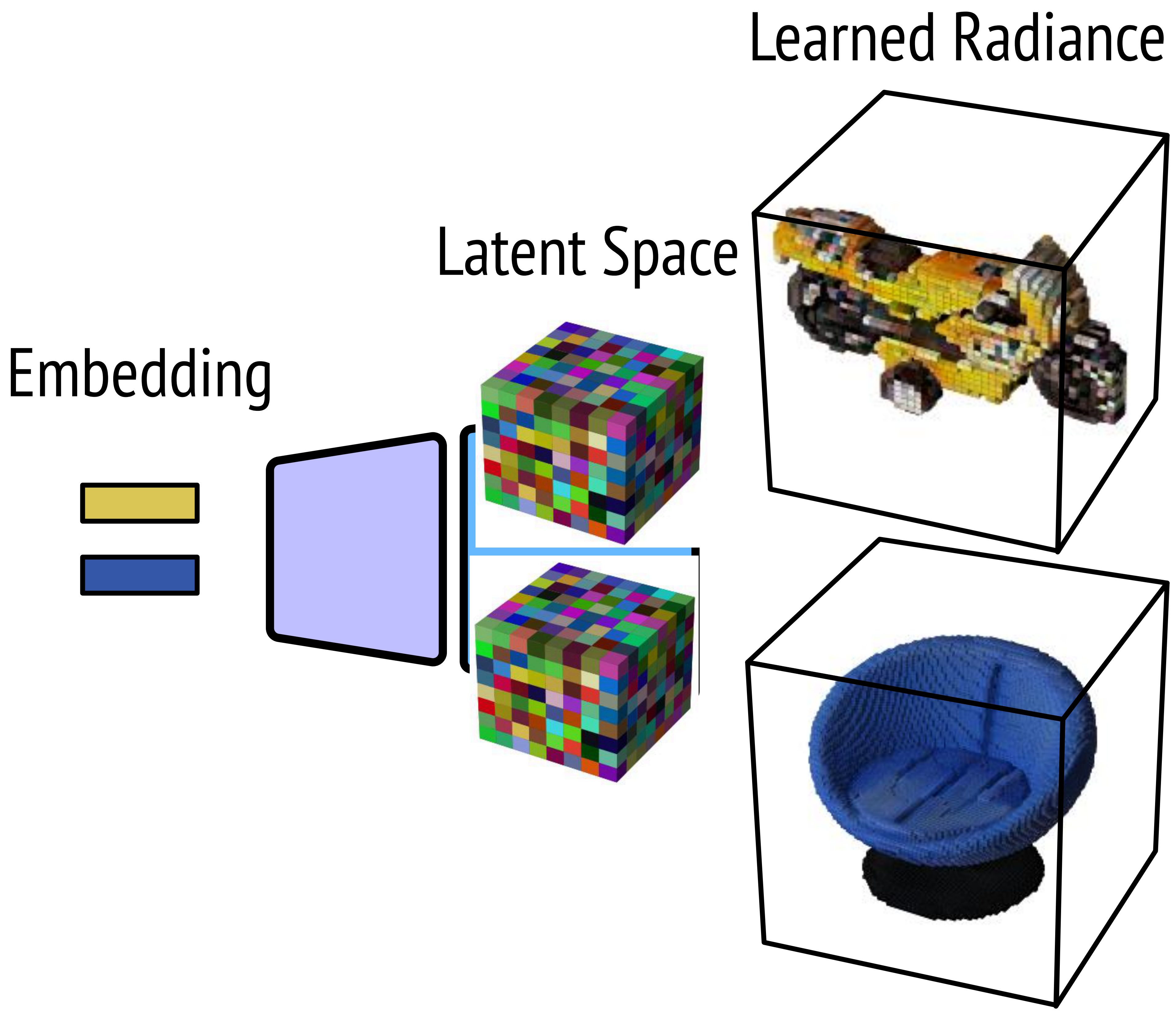
Generated, novel identities

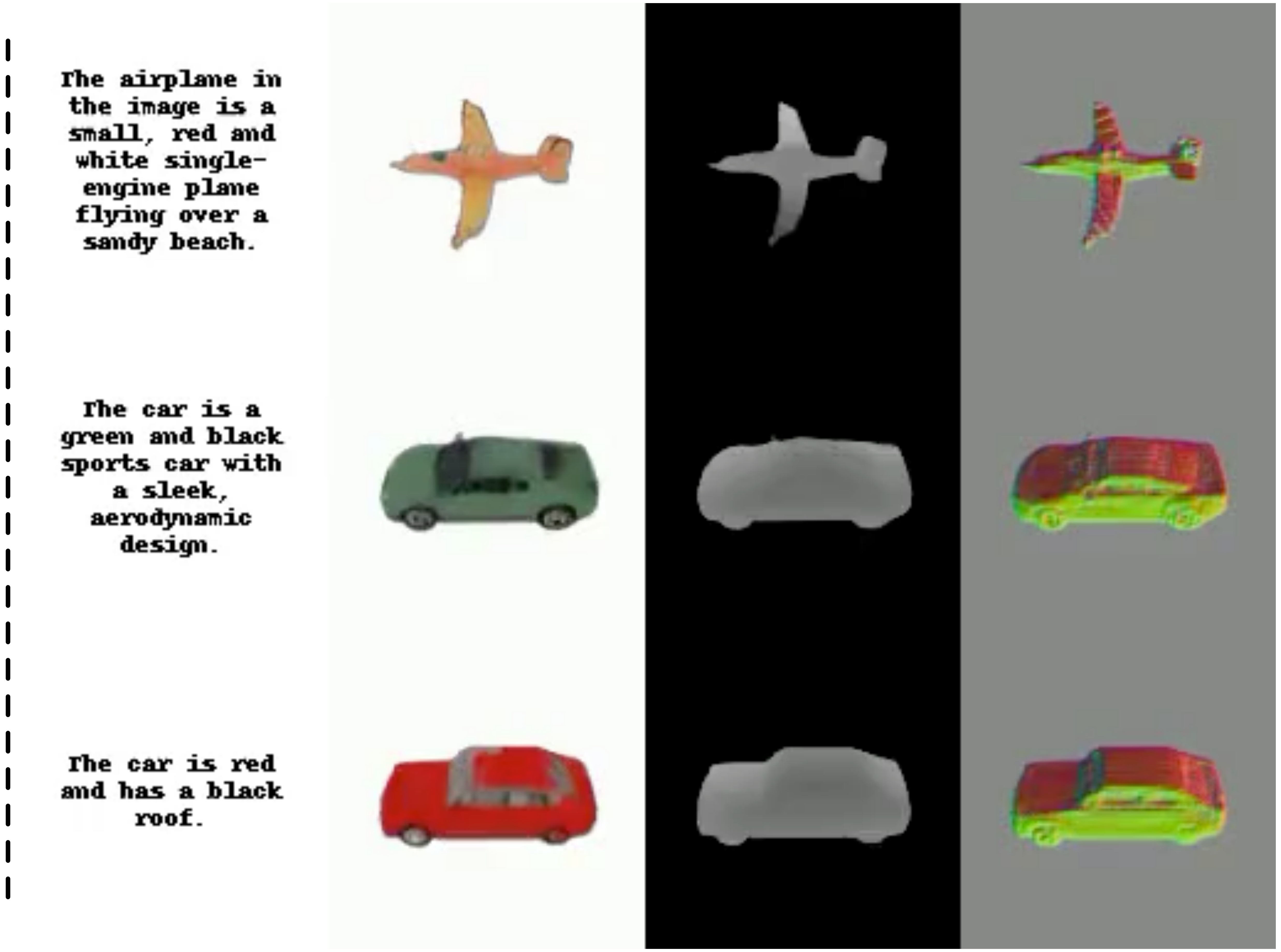


With UVA, we can learn to generate 3D objects without 3D data!

UVA learns a latent space. We can sample from it

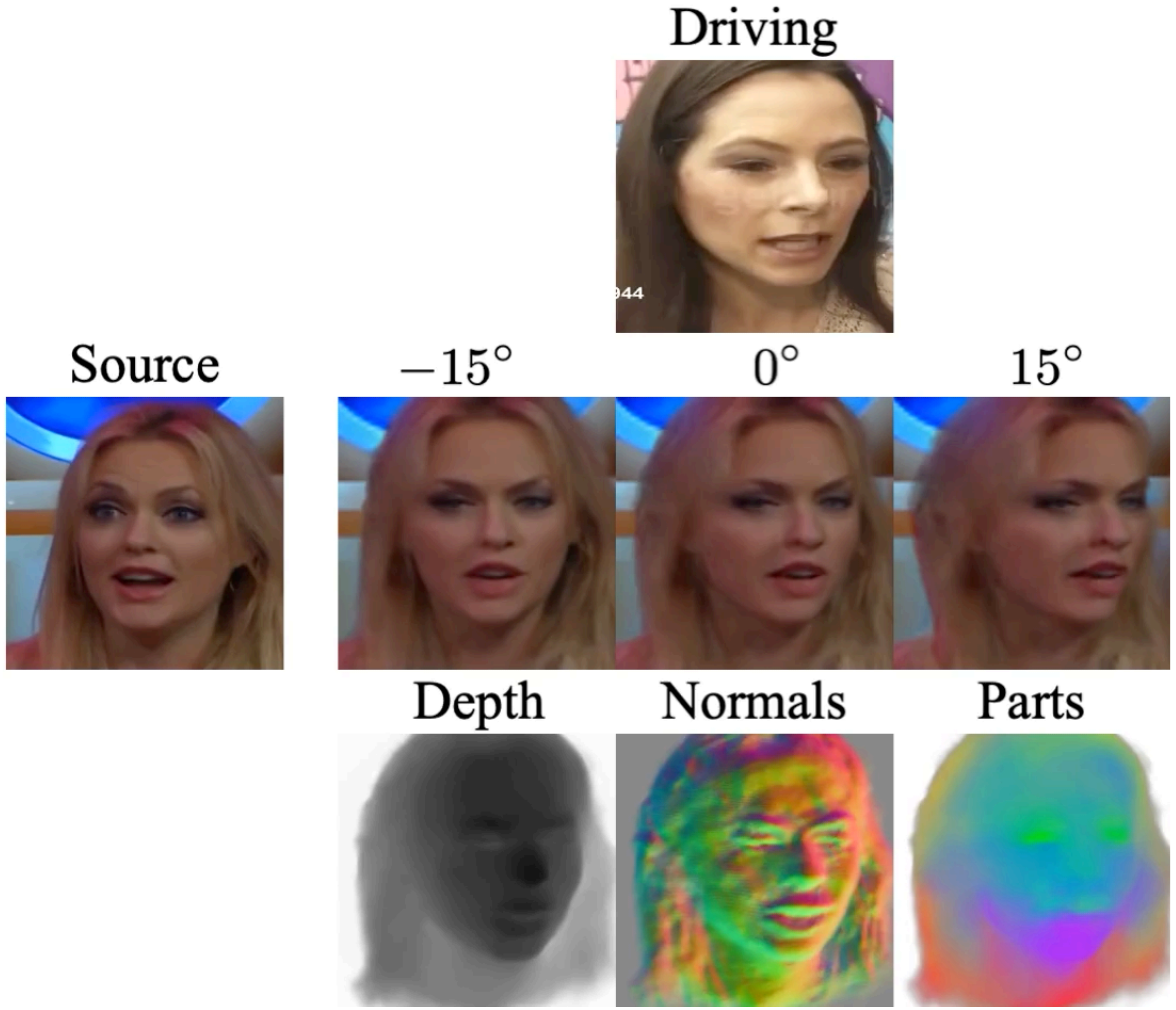
Synthesizing 3D without 3D Data





Ntavelis et al. "**AutoDecoding Latent 3D Diffusion Models**" ArXiv Preprint

Animation: Do as I Do




国窖1573
GUOJIAO 1573
国窖1573
GUOJIAO 1573
Emirates
FLY BETTER
Emirates
FLY BETTER
MELBOURNE

Making the Player Win

Standard approach:

- Reconstruct the scene → Model 3D environment, its style and physics
- Devise the winning strategy → Support game AI and game logic
- Animate players → Provide a sequence of commands
- Render results → Decide on viewing direction and render

Game Engine

Model 3D environment, its style and physics
Support game AI and game logic
Provide a sequence of commands
Decide on viewing direction and render

Is there a simpler way? → **Neural Game Engines**

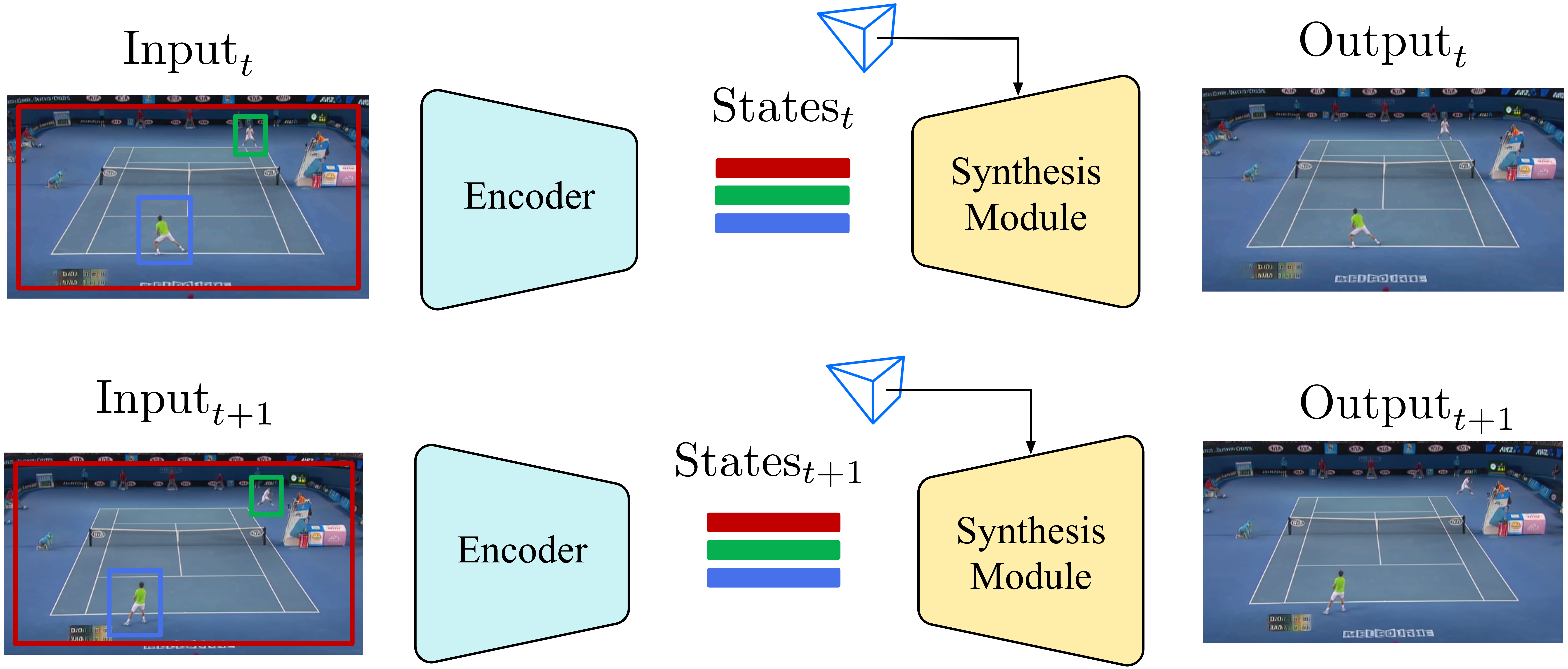
Prompt: "The top player is not able to catch the ball"


Prompt: "The top player is not able to catch the ball"
MELBOURNE
Critical moment: What should we do?

國窖1573
國窖1573
KIA
KIA
KIA
Emirates
FLY BETTER
Emirates
FLY BETTER
MELBOURNE


KIA
KIA
MELBOURNE

Generic Framework for Video Manipulation



Menapace et al. "**Playable Environments: Video Manipulation in Space and Time**" CVPR'2022

Menapace et al. "**Plotting Behind the Scenes: Towards Learnable Game Engines**" ArXiv Preprint

Animation Model to Learn Game AI

Player state

pose

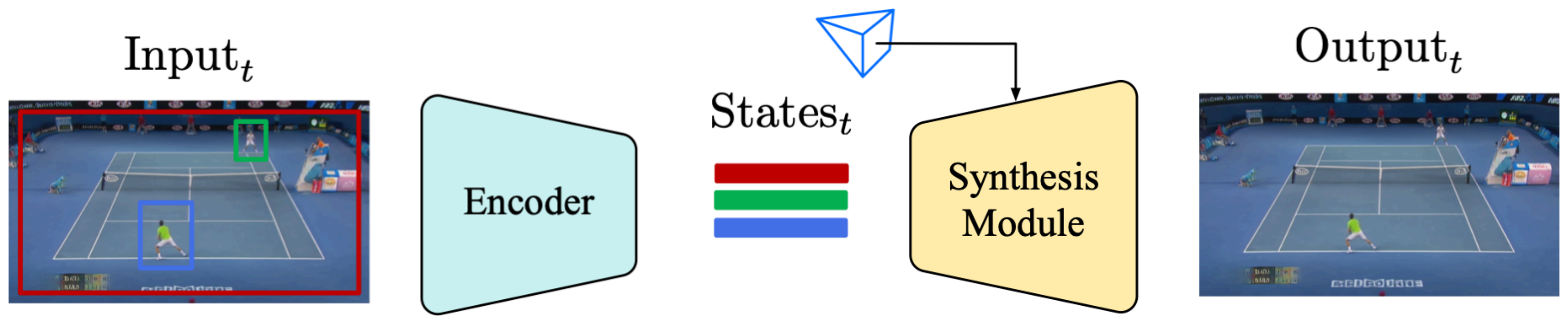
location

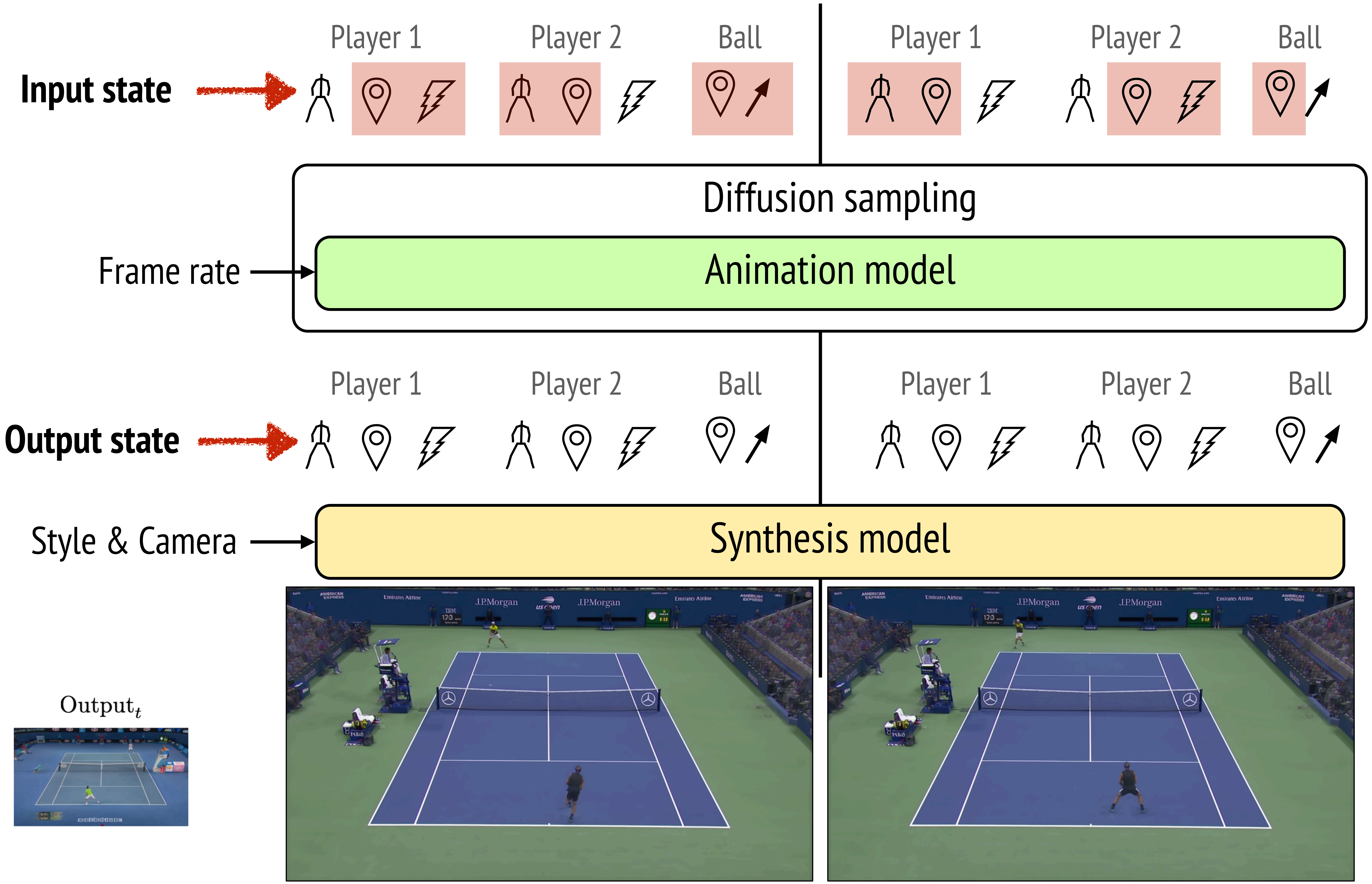
prompt or action

Tennis ball state

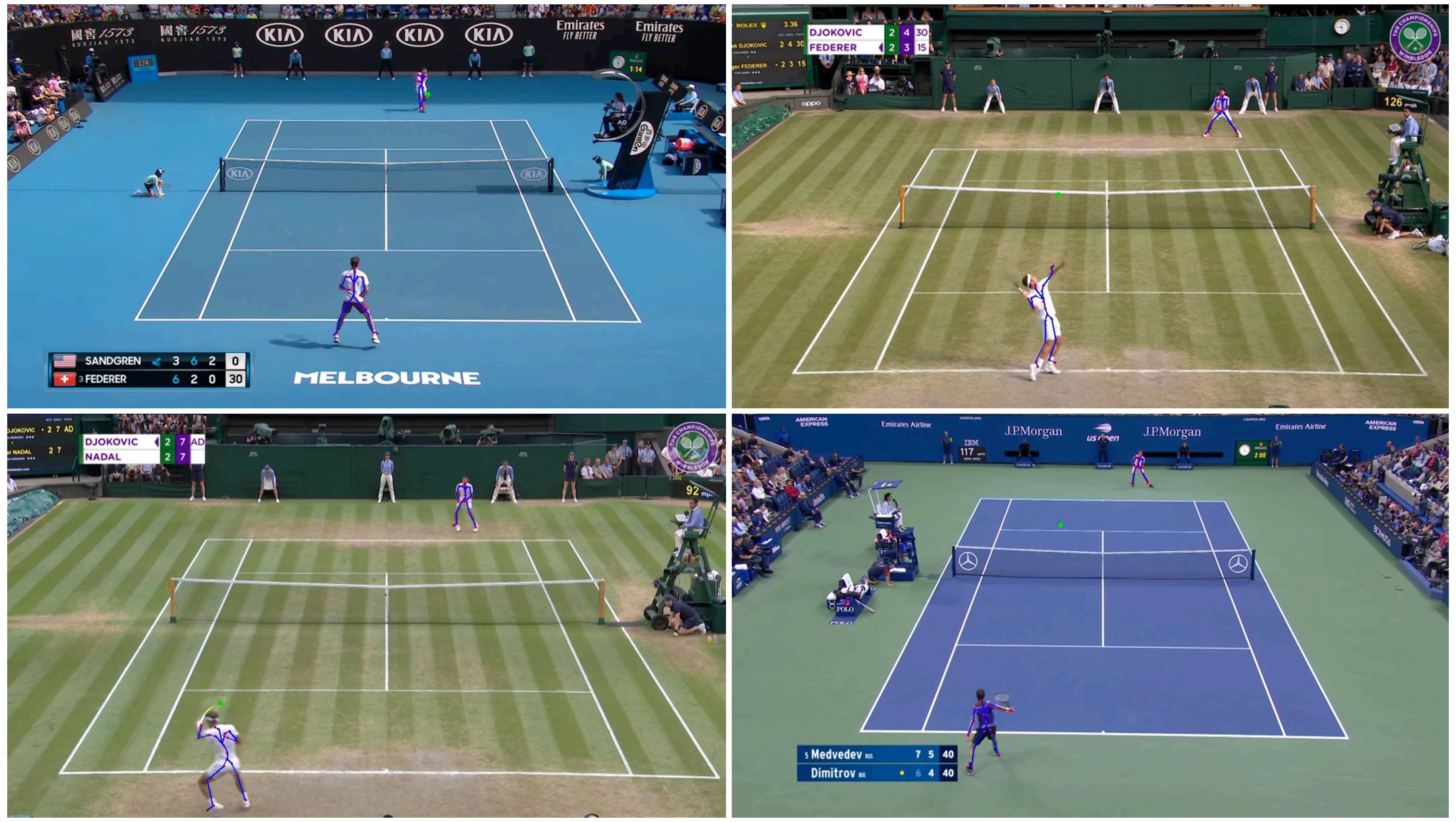
location

velocity



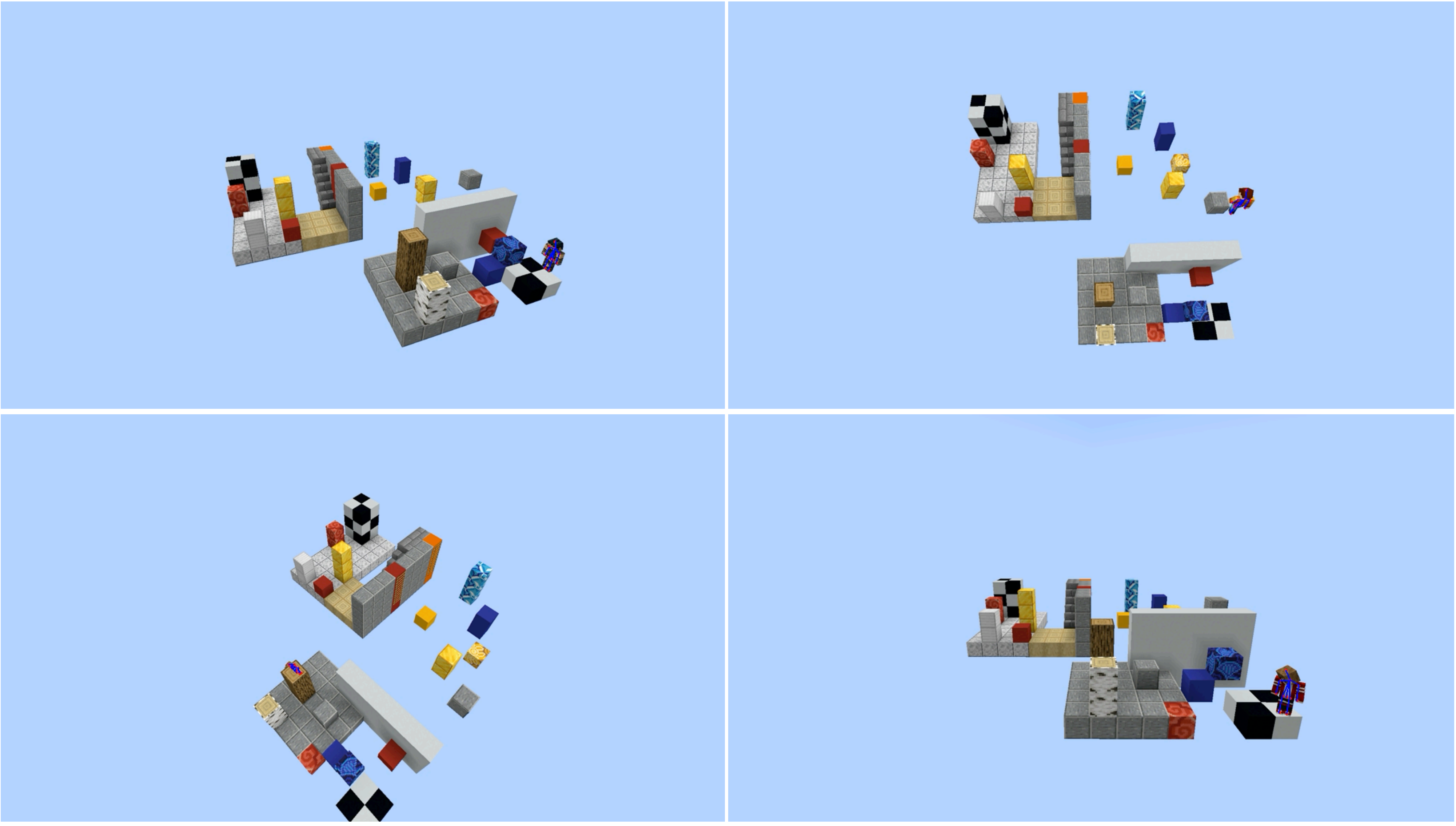


Datasets



Tennis

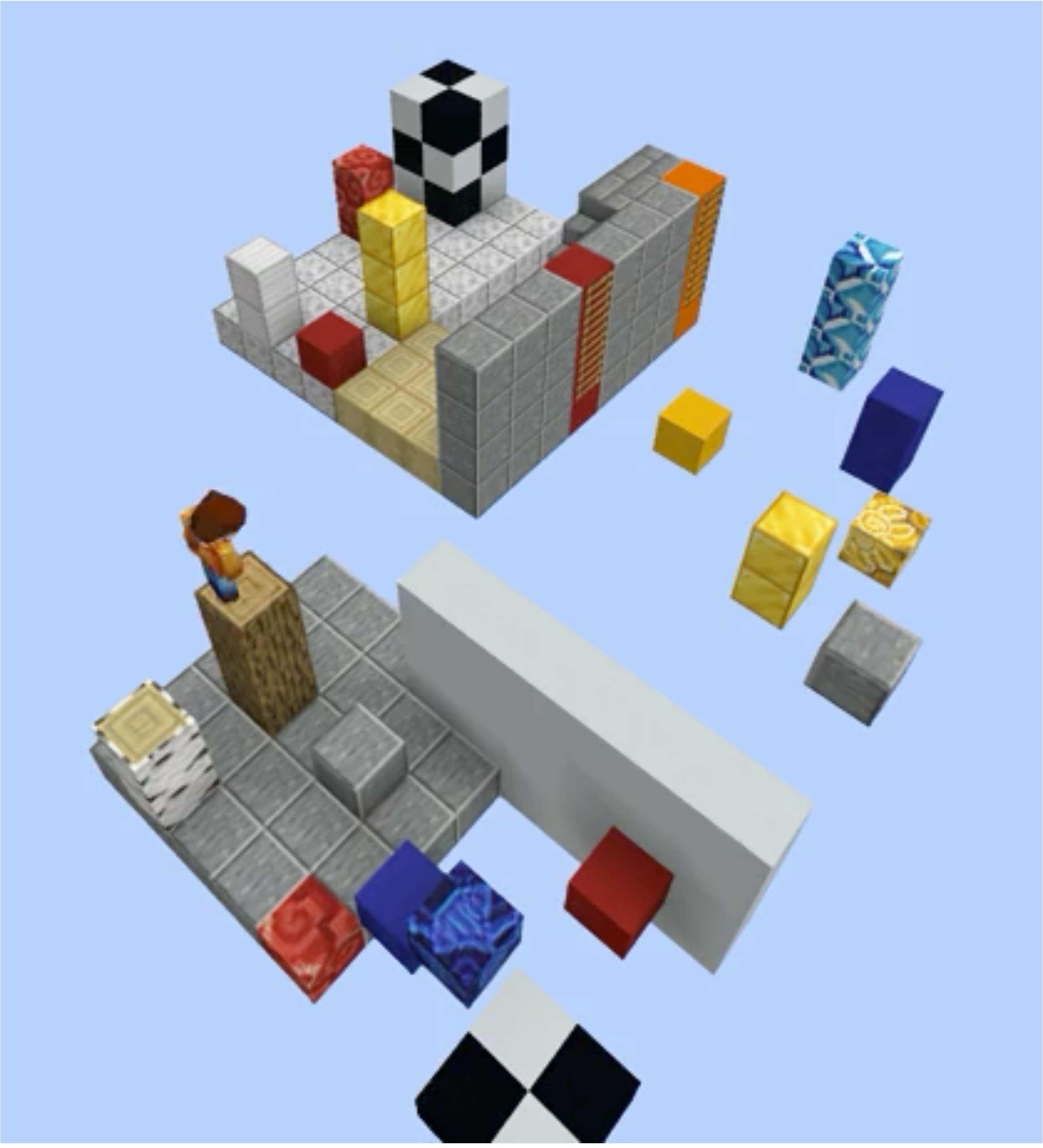
- 7112 video sequences at 1920x1080@25fps
- 15.5 hours of videos
- 1.12M fully annotated frames
- 25.5k unique captions



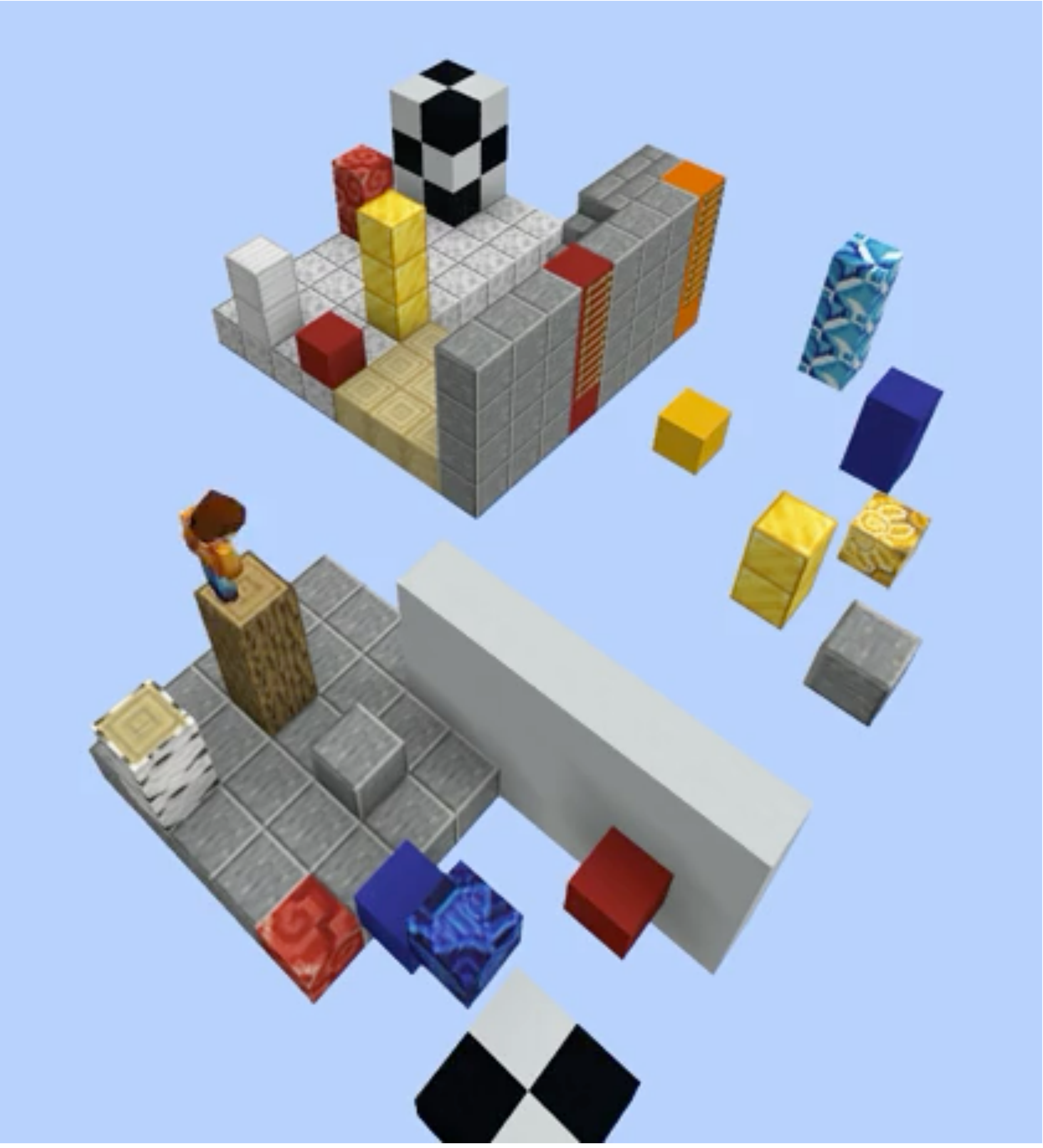
Minecraft

- 61 video sequences at 1024x567@20fps
- 1.2 hours of videos
- 68.5k fully annotated frames
- 1.24k unique captions

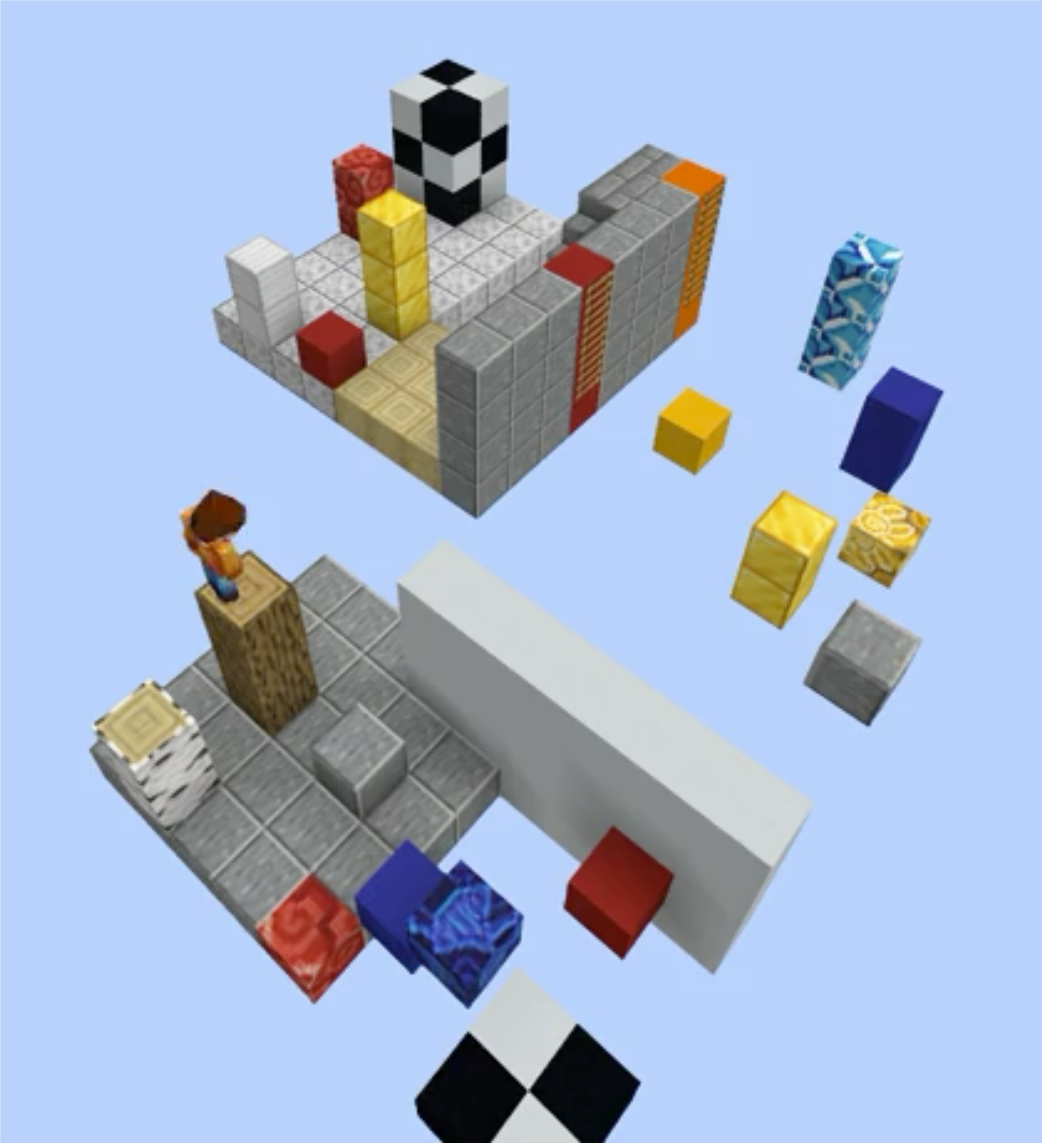
Controlling Generation with Prompts



"Falls on the stone platform"



"Jumps on a birch wood pillar"

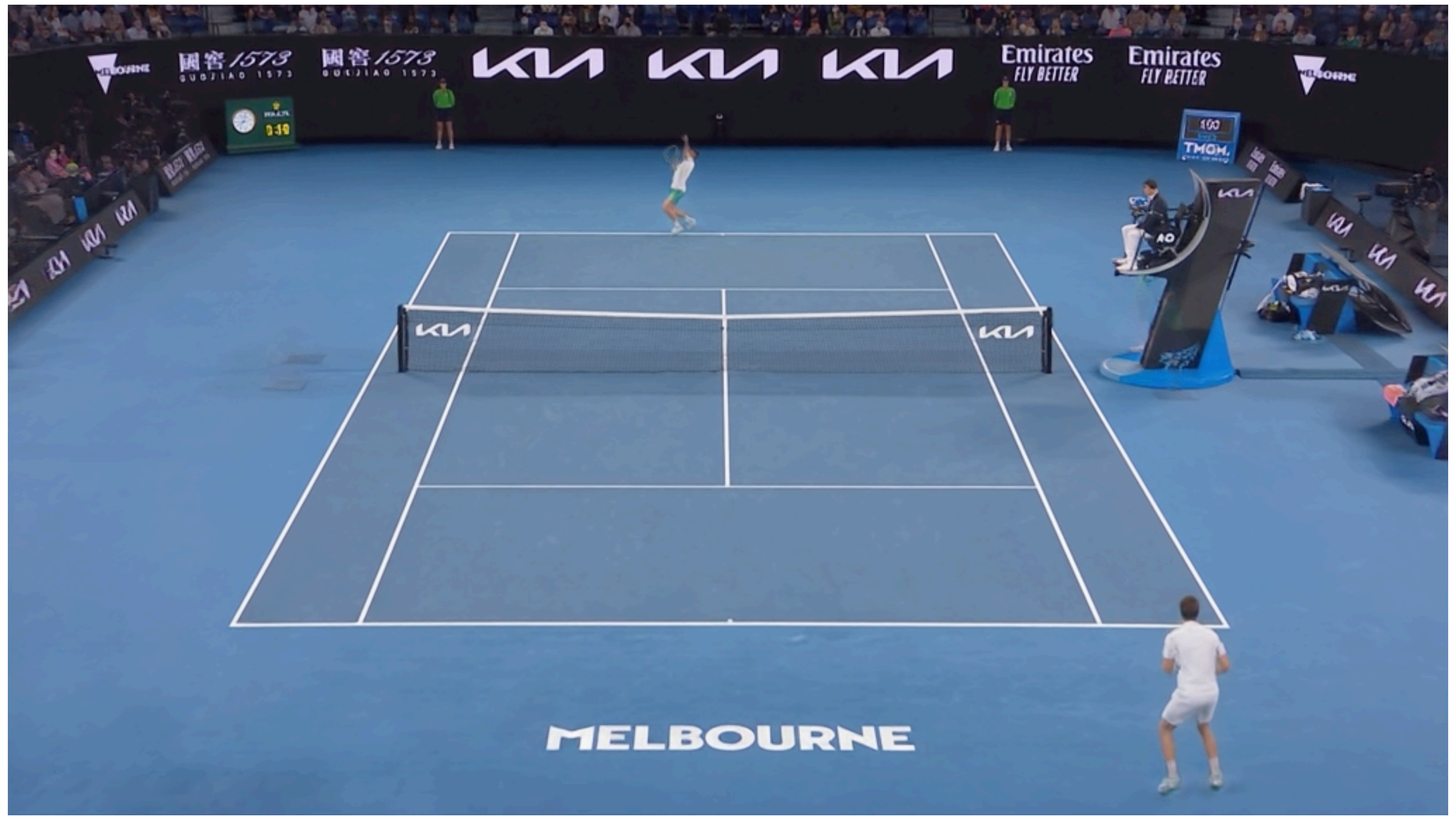


"Sprints and jumps on a white wall"

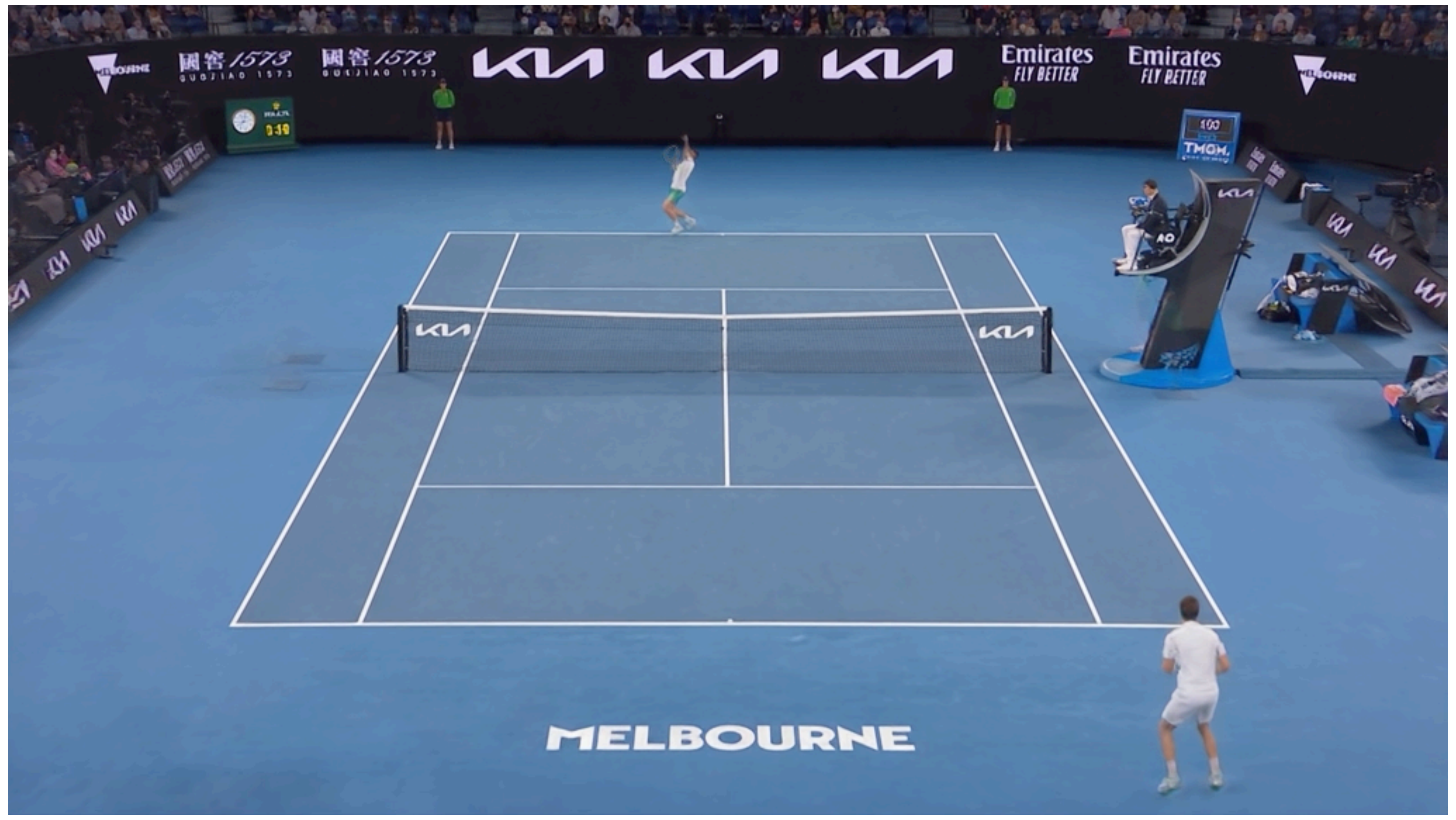
Player 1

Mask everything but the prompt

Controlling Generation with Prompts



"The player hits with a backhand"



"The player hits with a forehand"

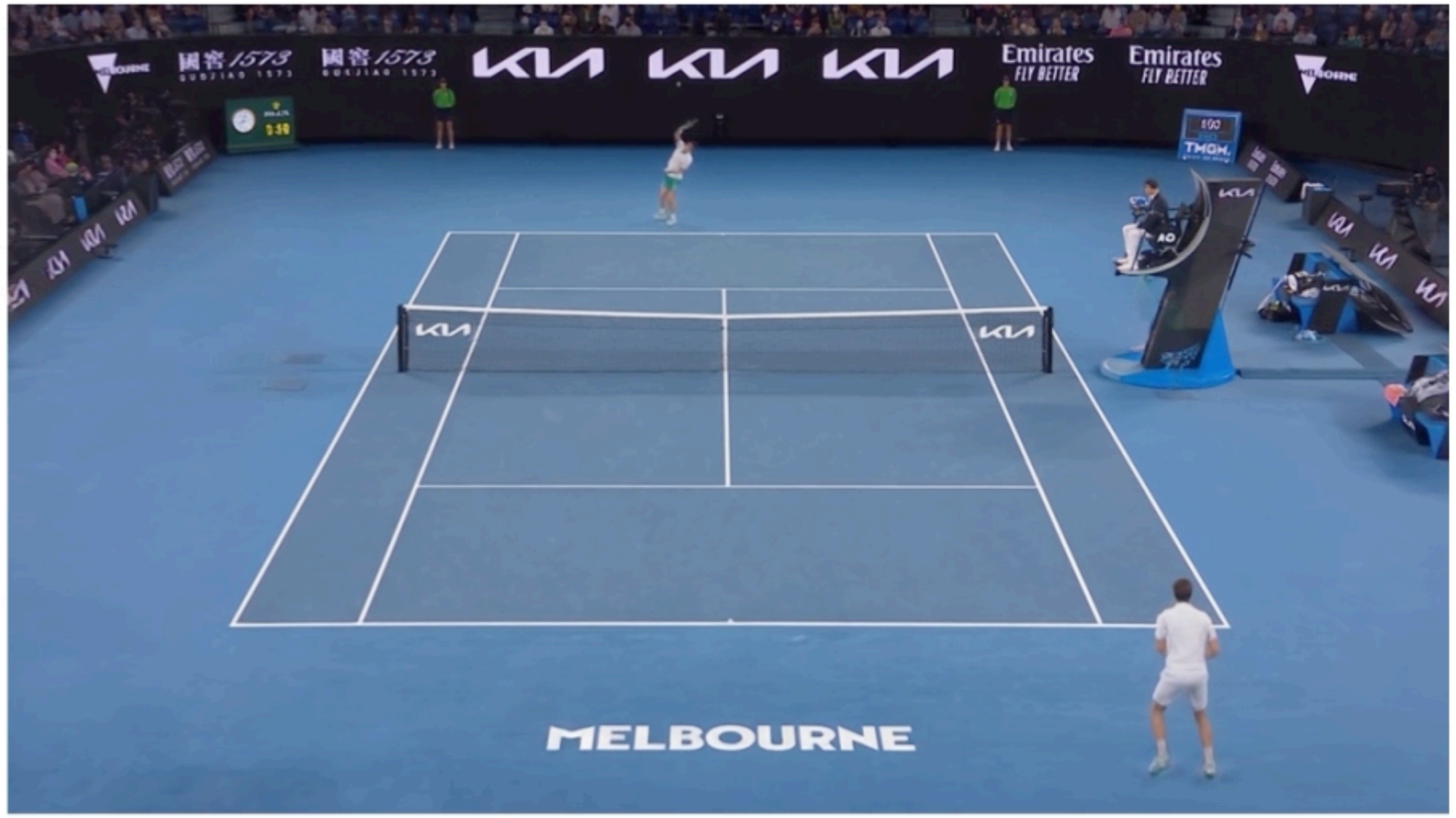
Player 1

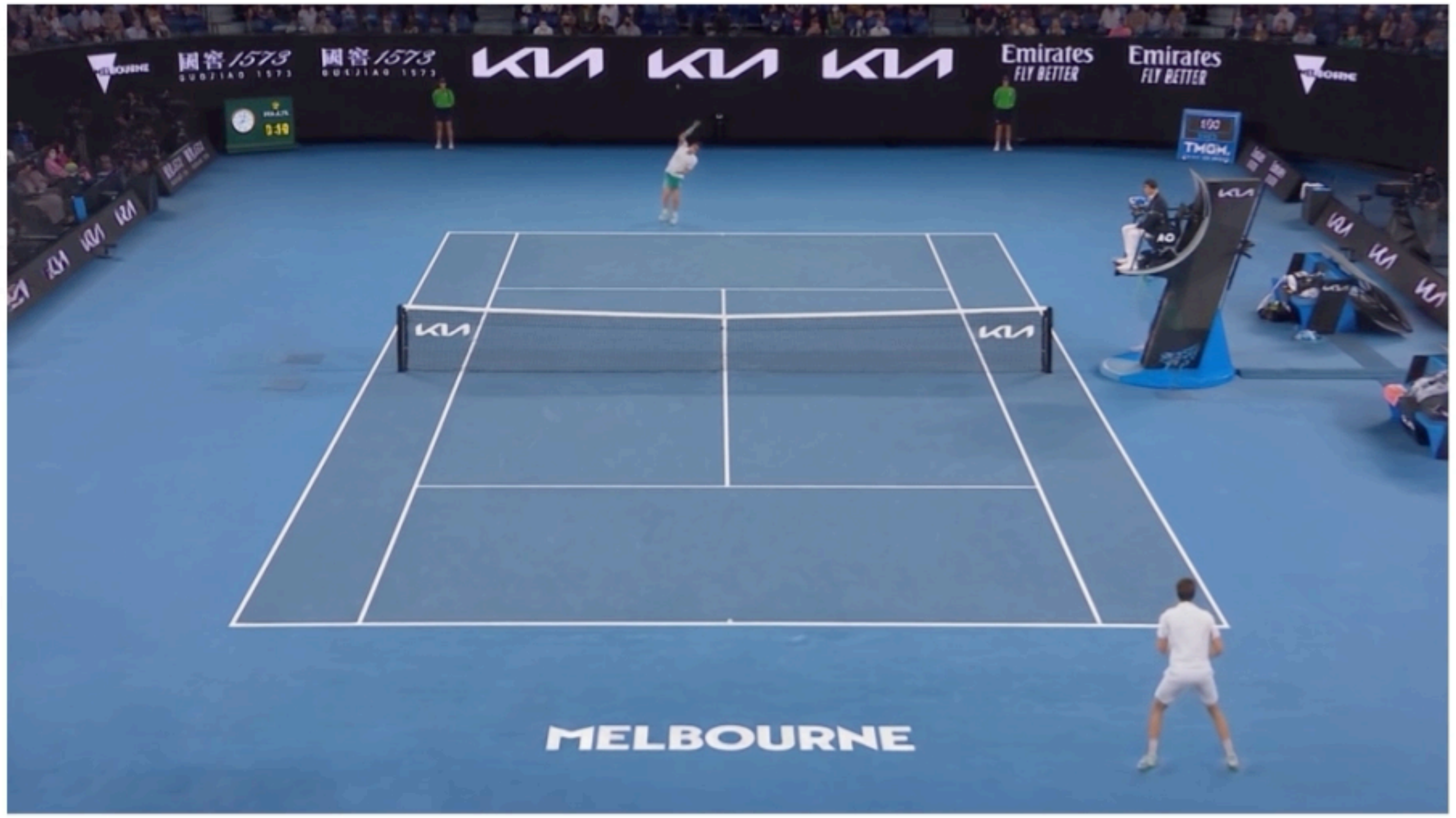
Mask everything but the prompt

Playing Against Learn Game AI

No action

The player serves and sends the ball to the service box





The player stands still waiting for a serve

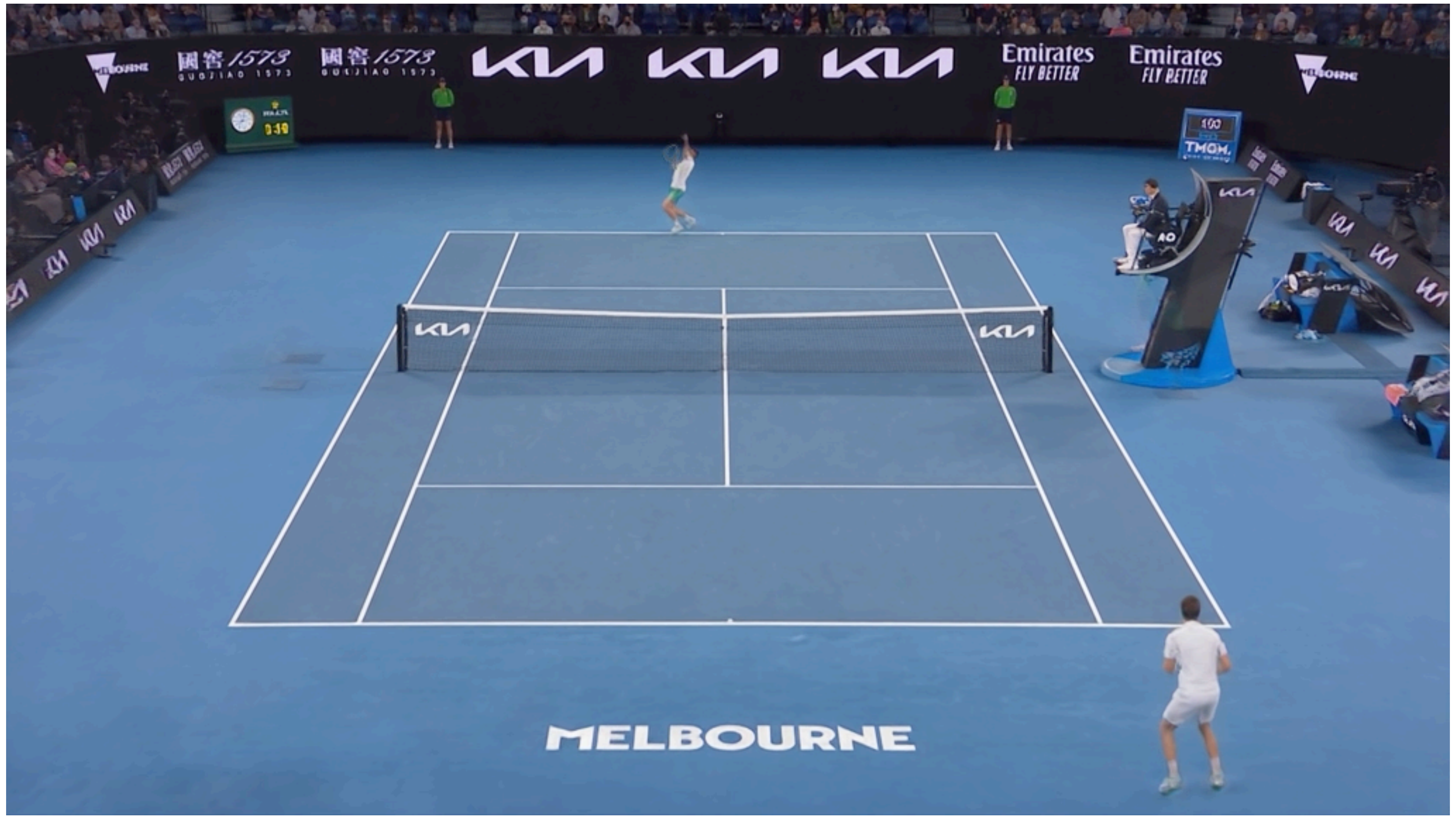
No action

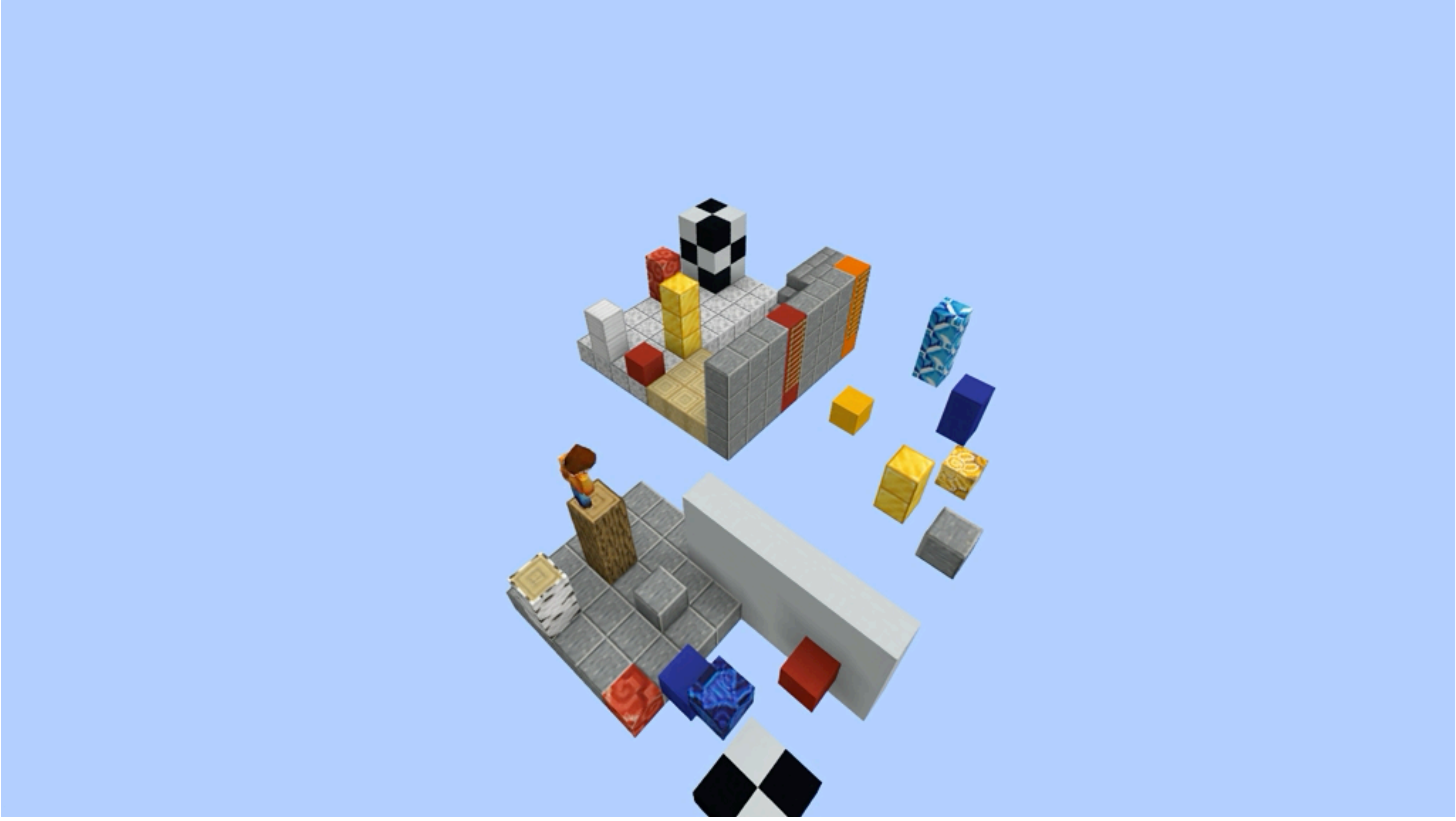
Player 1

Player 2

Completely mask the opponent

Game AI vs Game AI





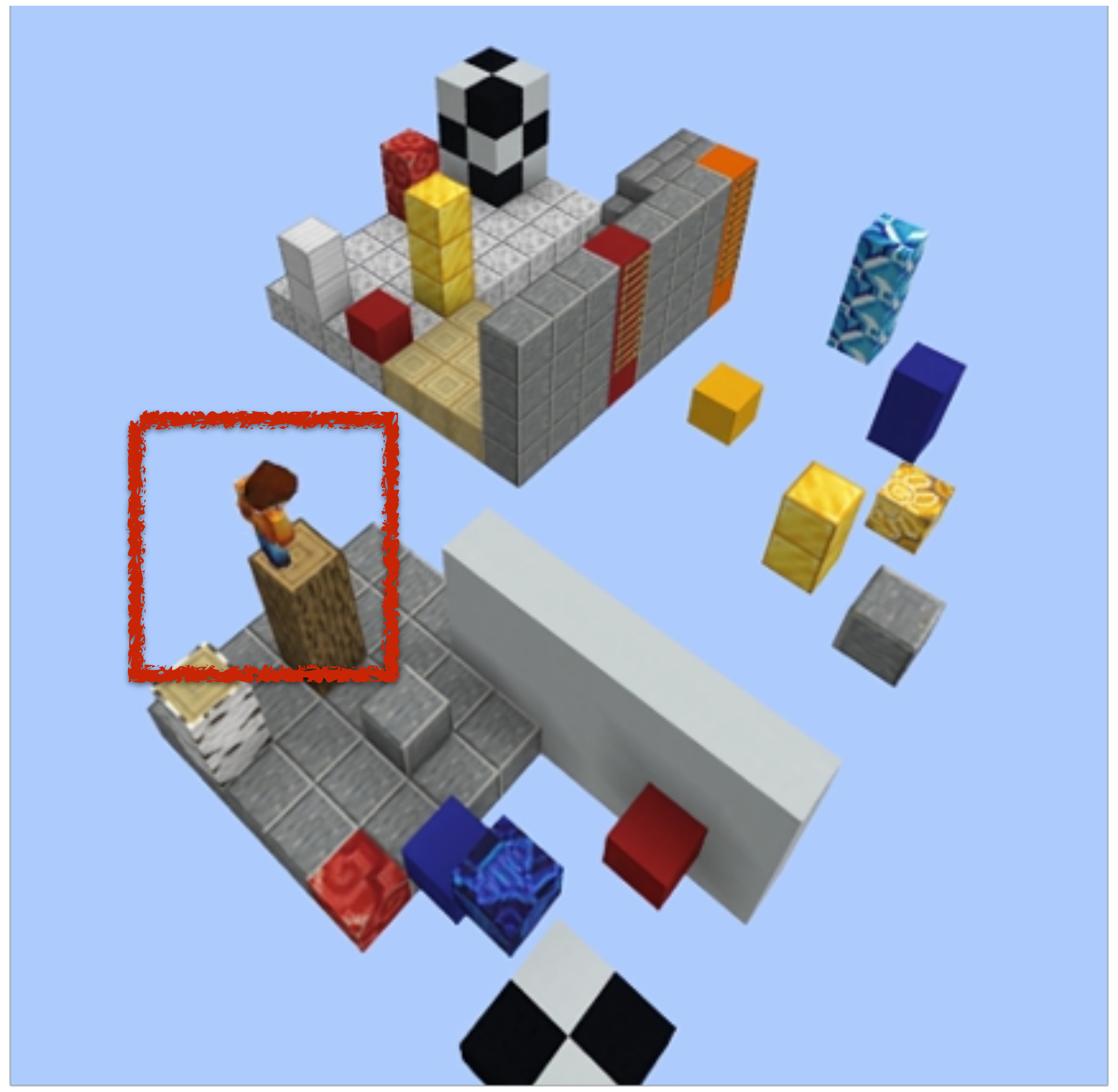
Mask both players

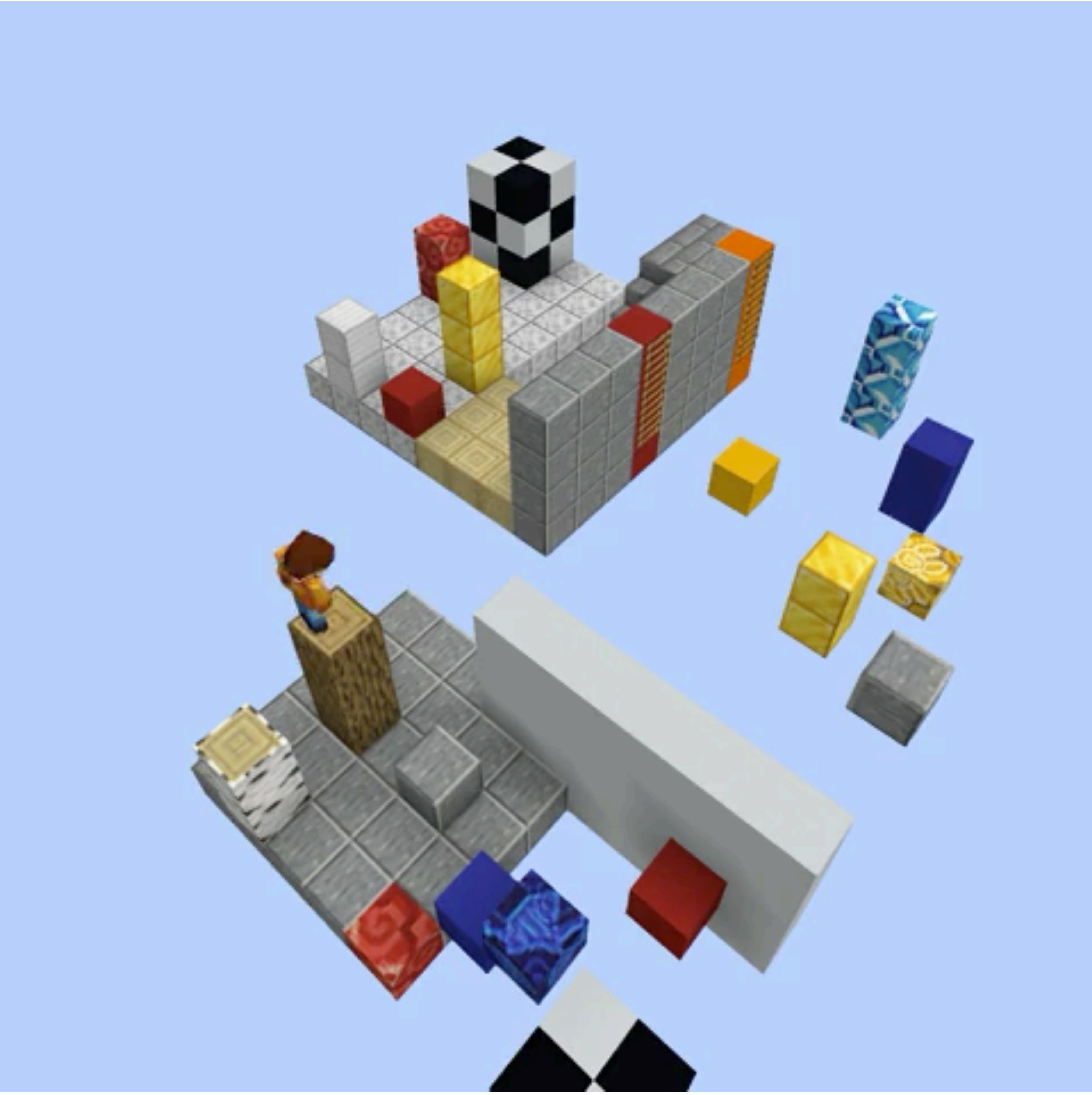
Player 1

Player 2

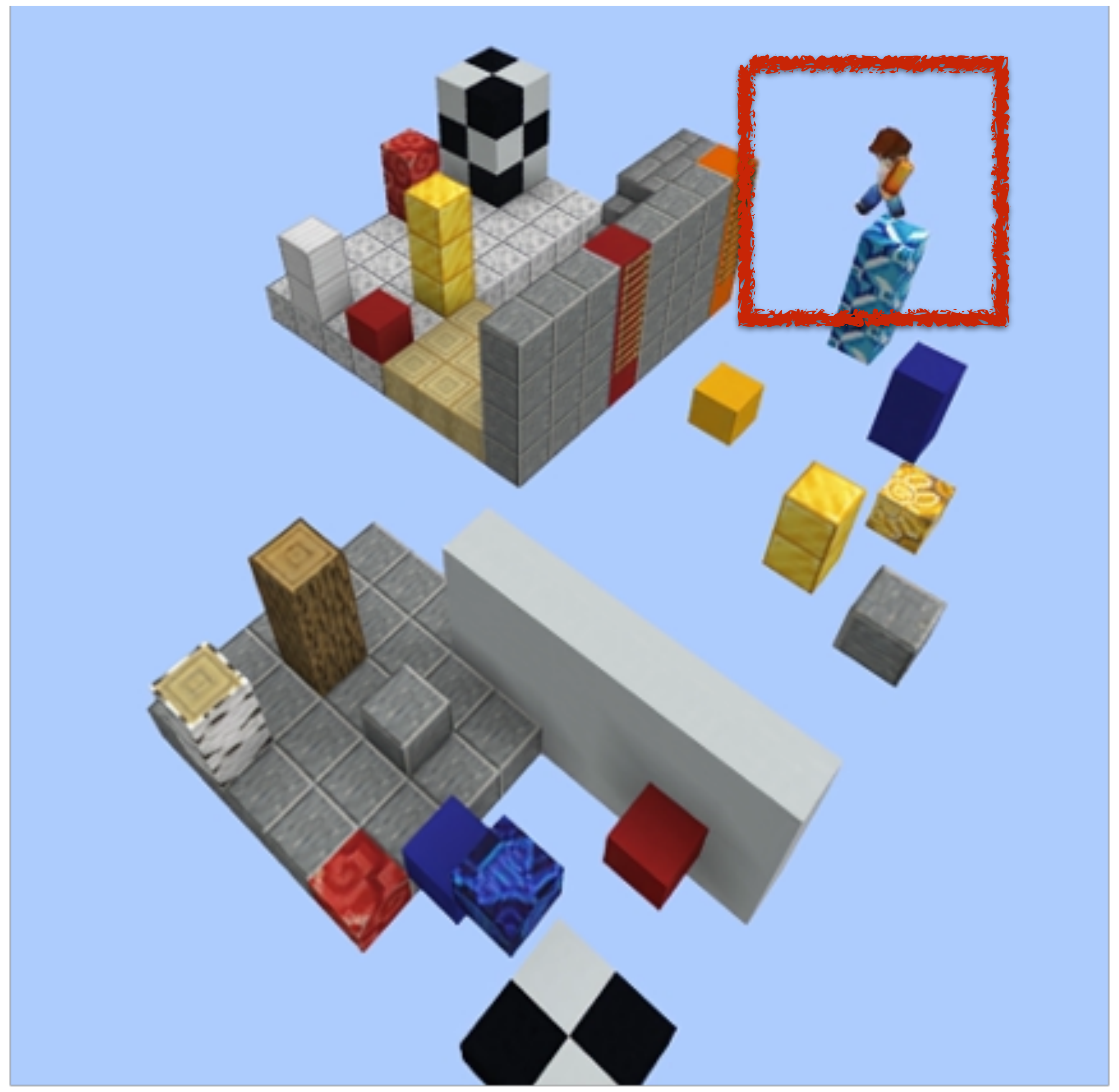
Director's Mode

Initial state

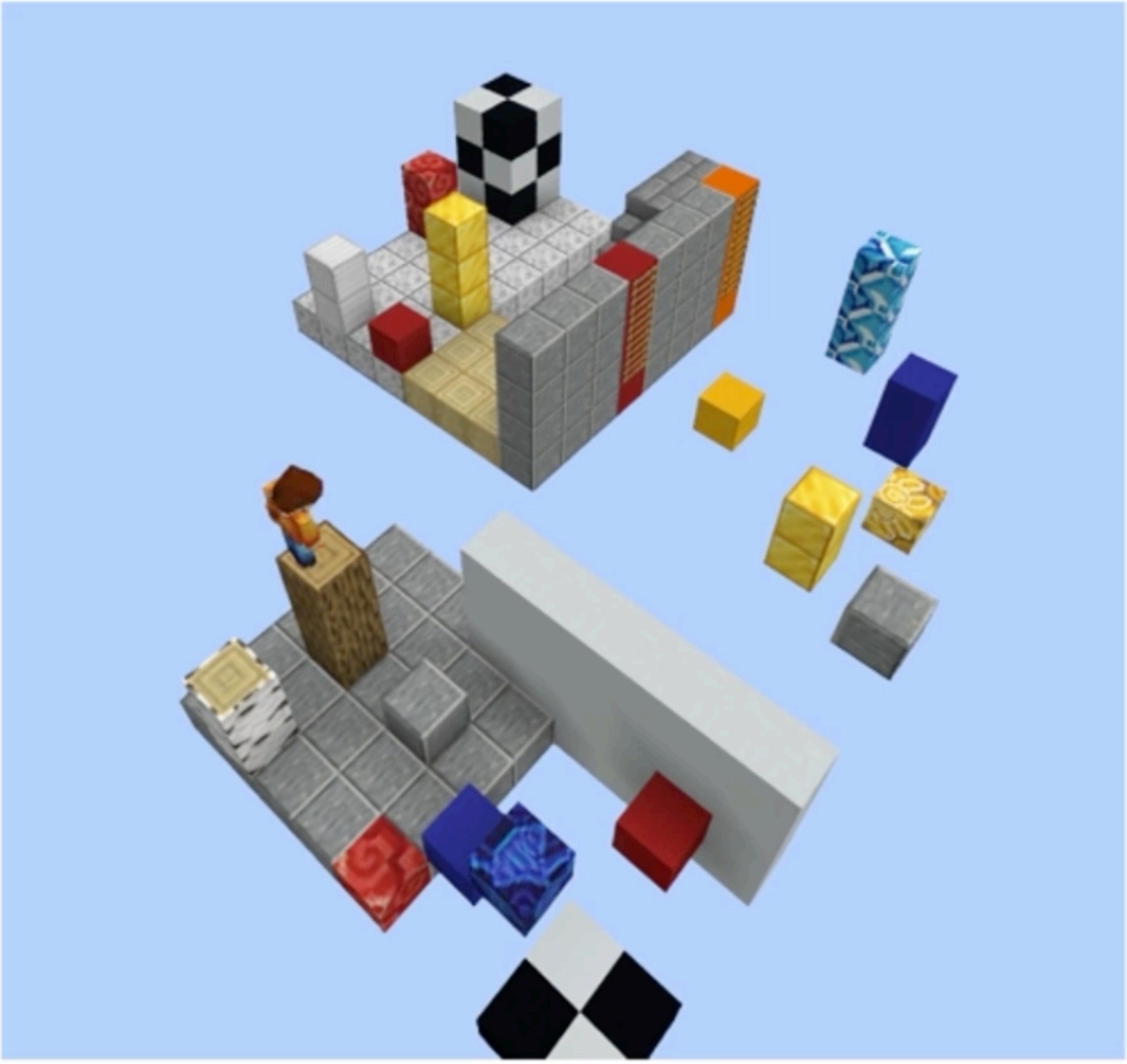


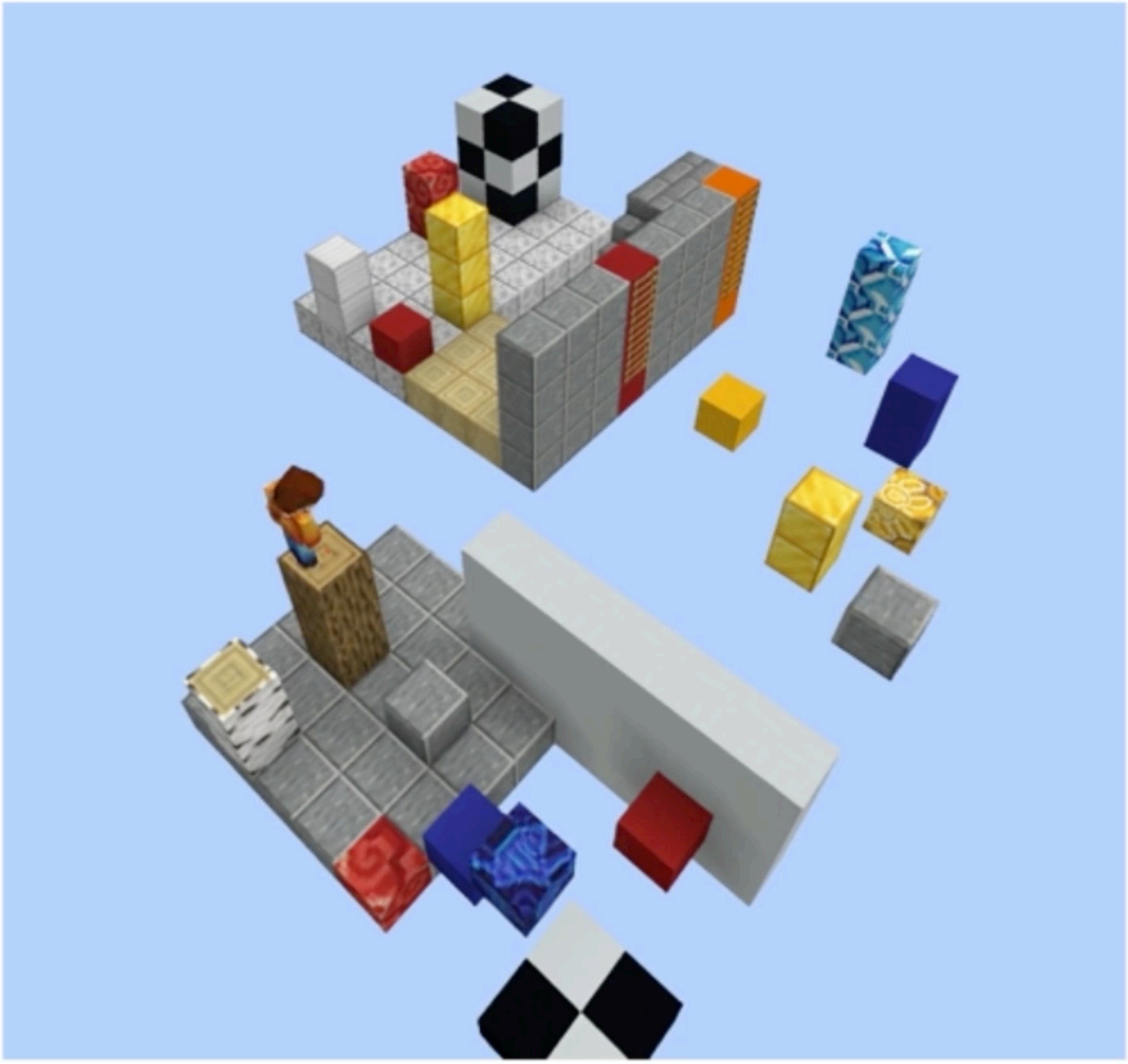


Final state



Director’s Mode

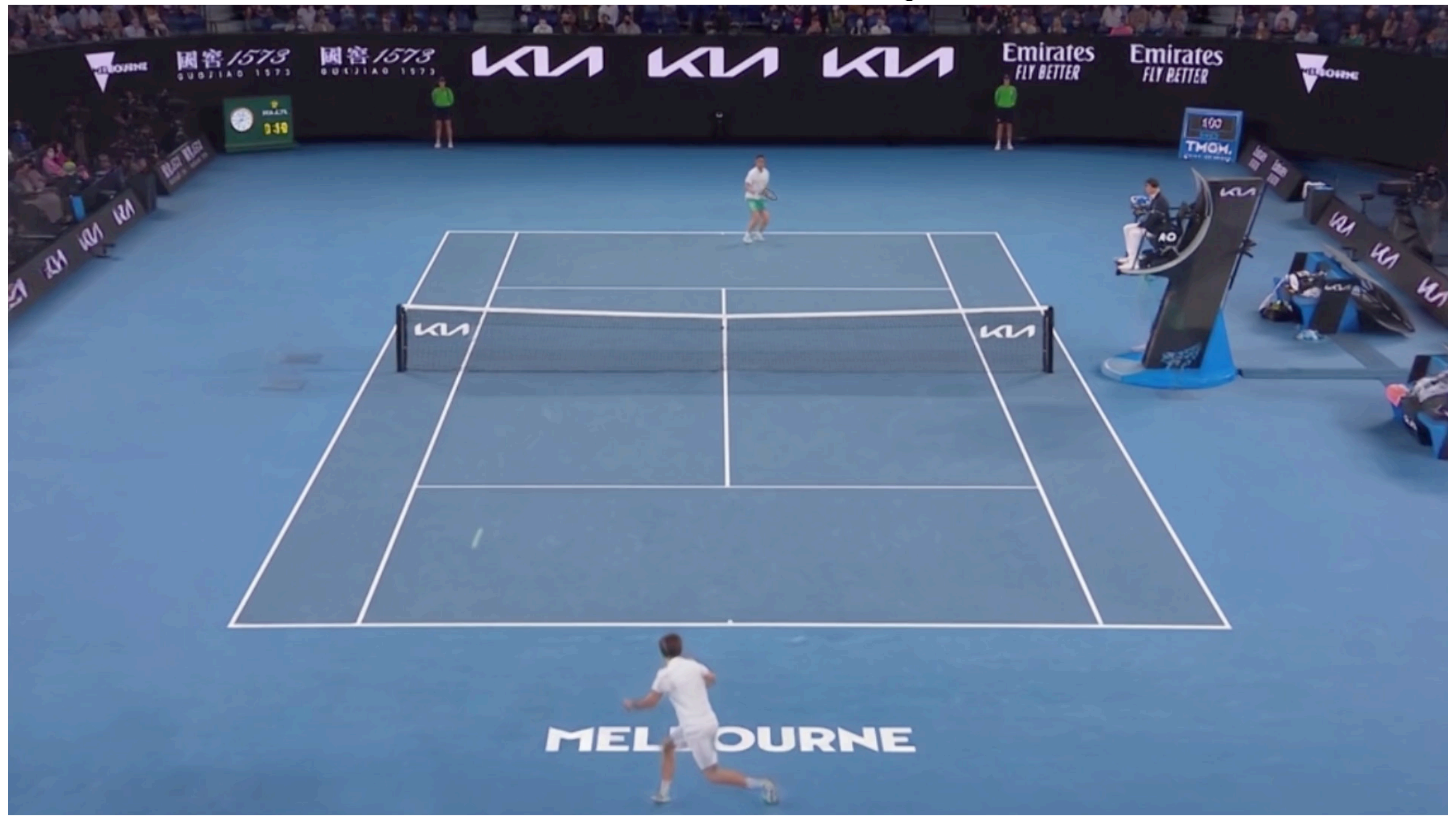




Shoutout to our Team, Interns, and Collaborators

‘Do as I Do’

‘Do as I Say’



Elisa Ricci

Stéphane Lathuilière

Willi Menapace

Kyle Olszewski

Aliaksandr Siarohin

Hsin-Ying Lee

Vladislav Golyanik

Ivan Skorokhodov