Compositional 3D-aware Video Generation with LLM Director

Hanxin Zhu[1]*, Tianyu He[2], Anni Tang[3], Junliang Guo[2], Zhibo Chen[1], Jiang Bian[2]

[1]University of Science and Technology of China
[2]Microsoft Research Asia
[3]Shanghai Jiao Tong University

hanxinzhu@mail.ustc.edu.cn, tianyuhe@microsoft.com, memory97@sjtu.edu.cn,
junliangguo@microsoft.com, chenzhibo@ustc.edu.cn, jiang.bian@microsoft.com

➢ Motivation:

In nature, our understanding of the world is compositional, and the interaction with the world takes place in a 3D. Motivated by this, in contrast to the prior endeavors implicitly learn different concepts in 2D space, we are interested in exploring an alternative solution that explicitly composes concepts in 3D space for video generation.

➢ Insight:

We view the process of 3D-aware video generation as a composition of static environment and dynamic objects.

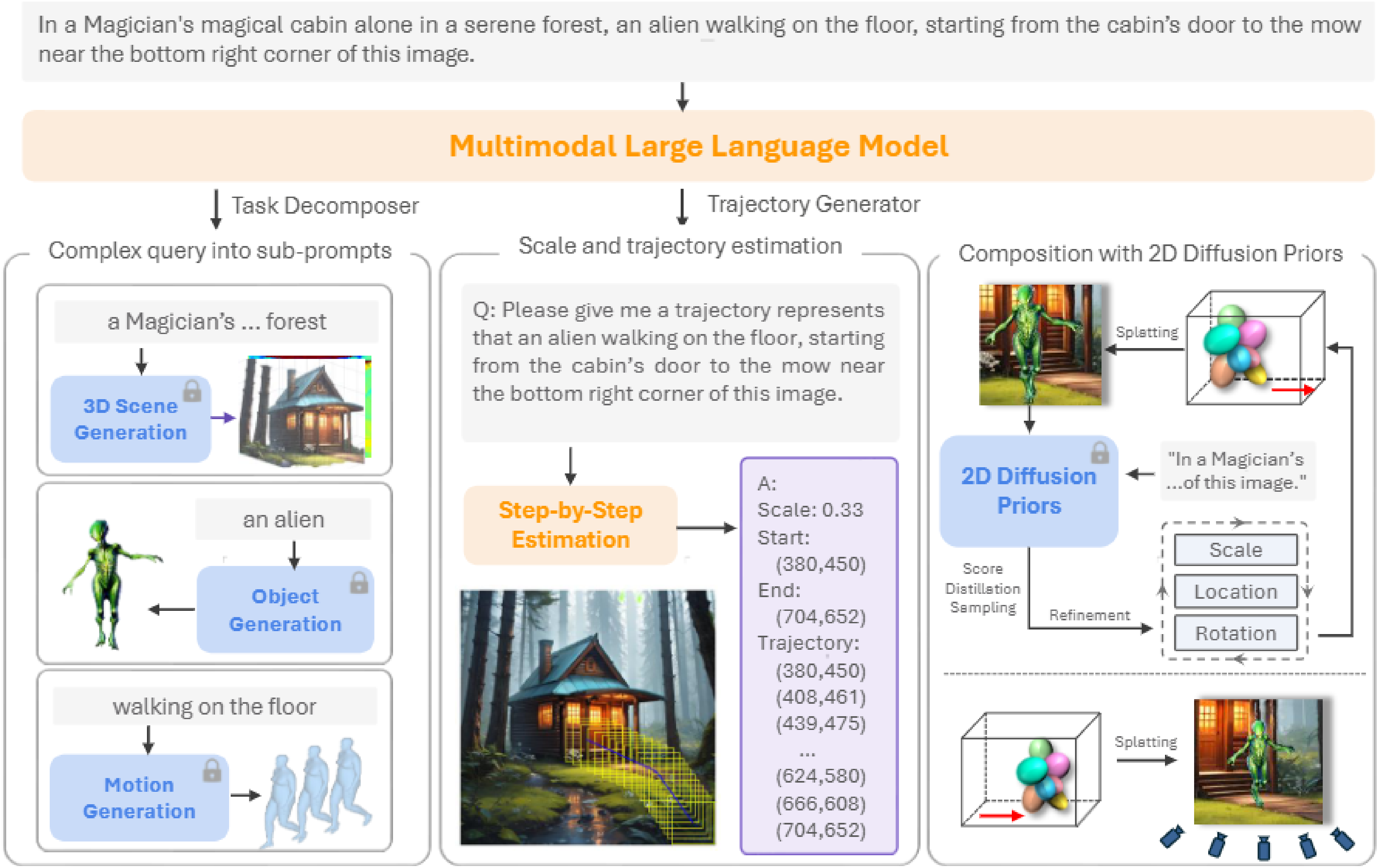
➢ Advantages:

- Because each concept is represented by individual 3D representations, it naturally supports flexible control and interaction of each concept.
- It inherently excels at synthesizing complex and long videos such as drama, etc.
- The viewpoint is controllable.

➢ Challenges:

- Since a textual prompt contains multiple concepts, how to coordinate the generation of various concepts?
- Given the generated concepts, how to compose them to follow common sense in the real world?

➢ Illustration of our method.:



In a Magician's magical cabin alone in a serene forest, an alien walking on the floor, starting from the cabin's door to the mow near the bottom right corner of this image.

Multimodal Large Language Model

Task Decomposer

Trajectory Generator

Complex query into sub-prompts

a Magician's ... forest

3D Scene Generation

an alien

Object Generation

walking on the floor

Motion Generation

Scale and trajectory estimation

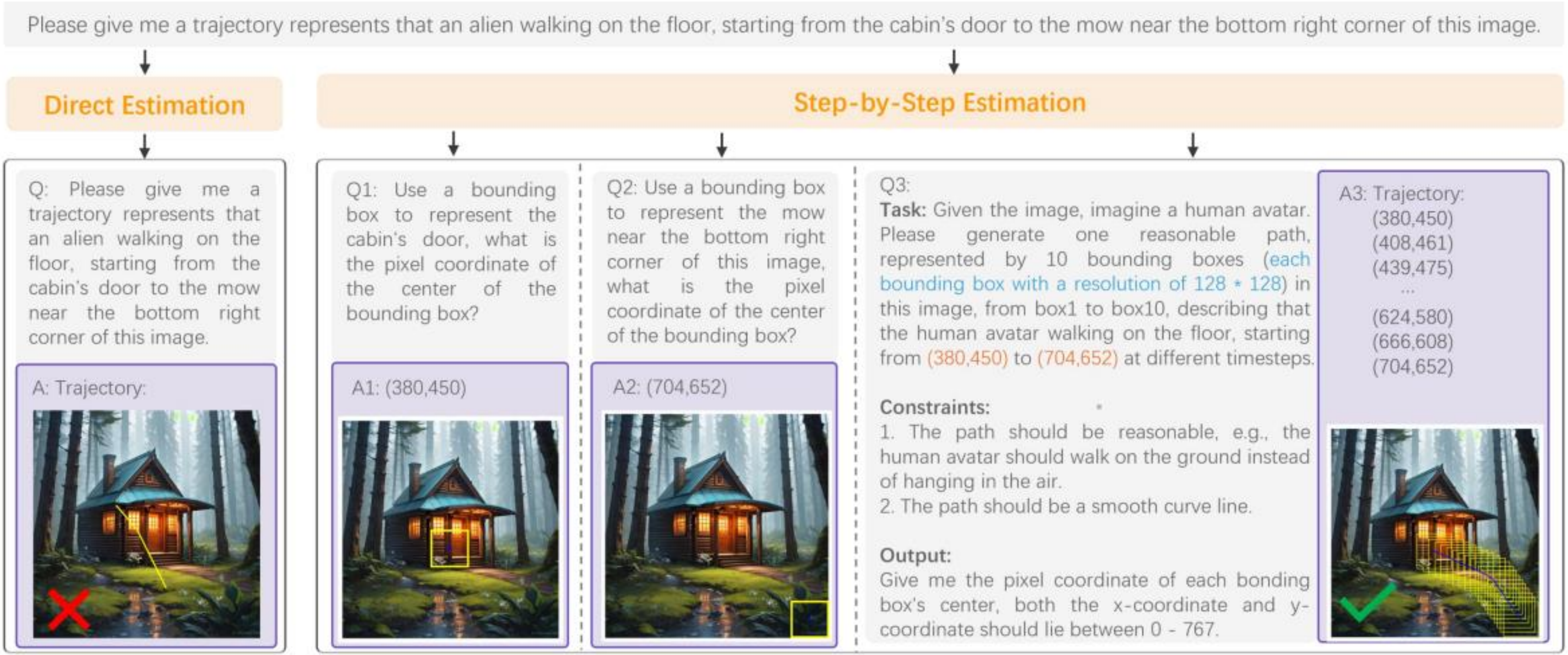
Q: Please give me a trajectory represents that an alien walking on the floor, starting from the cabin's door to the mow near the bottom right corner of this image.

Step-by-Step Estimation

A:
Scale: 0.33
Start:
(380,450)
End:
(704,652)
Trajectory:
(380,450)
(408,461)
(439,475)
...
(624,580)
(666,608)
(704,652)

Composition with 2D Diffusion Priors

➢ Illustration of coarse-grained trajectory generation with LLM



Please give me a trajectory represents that an alien walking on the floor, starting from the cabin's door to the mow near the bottom right corner of this image.

Q: Please give me a trajectory represents that an alien walking on the floor, starting from the cabin's door to the mow near the bottom right corner of this image.

A: Trajectory:

Q1: Use a bounding box to represent the cabin's door, what is the pixel coordinate of the center of the bounding box?

A1: (380,450)

Q2: Use a bounding box to represent the mow near the bottom right corner of this image, what is the pixel coordinate of the center of the bounding box?

A2: (704,652)

Q3:

Task: Given the image, imagine a human avatar. Please generate one reasonable path, represented by 10 bounding boxes (each bounding box with a resolution of 128 * 128) in this image, from box1 to box10, describing that the human avatar walking on the floor, starting from (380,450) to (704,652) at different timesteps.

Constraints:

1. The path should be reasonable, e.g., the human avatar should walk on the ground instead of hanging in the air.
2. The path should be a smooth curve line.

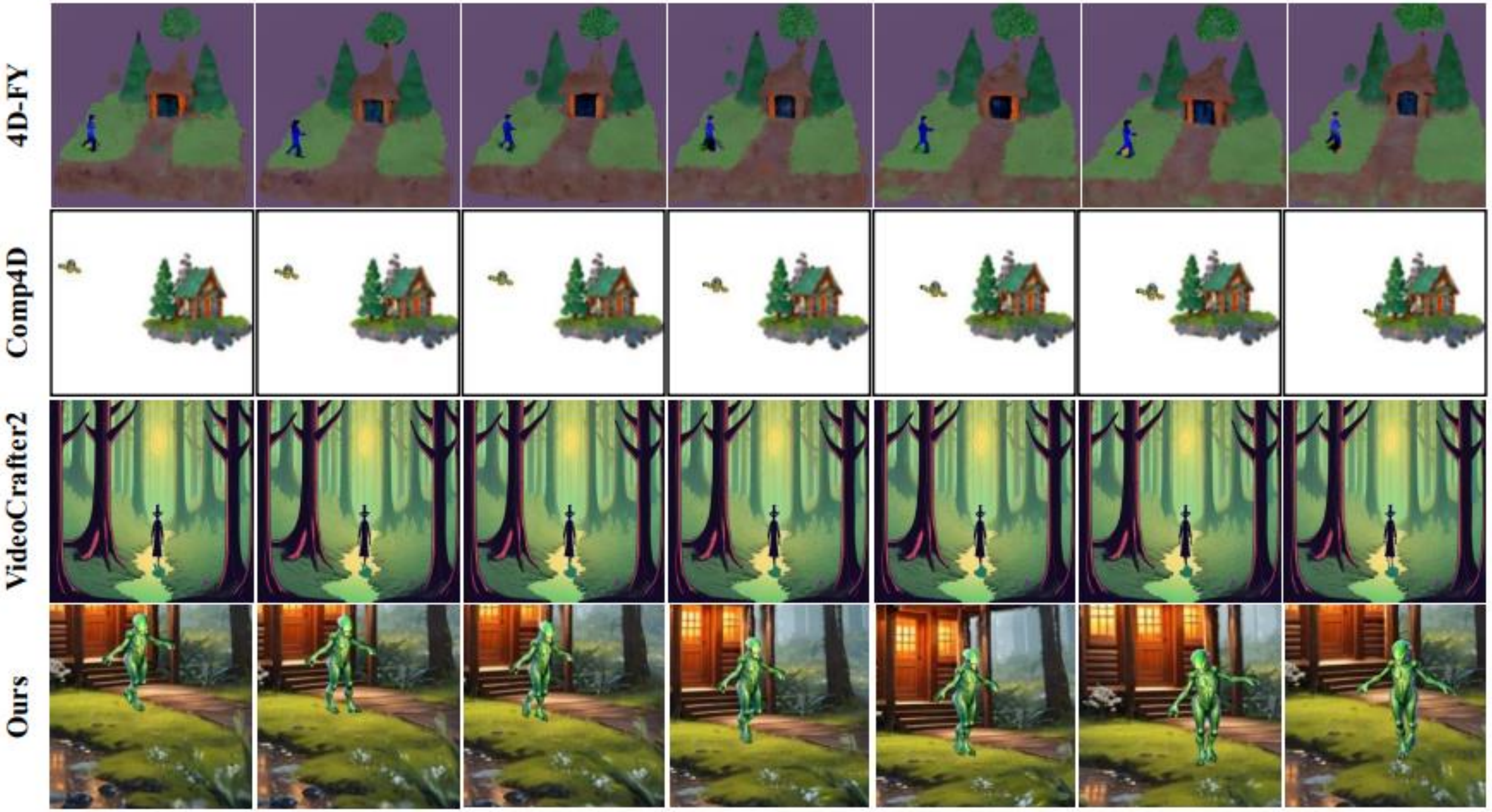
Output:

Give me the pixel coordinate of each bonding box's center, both the x-coordinate and y-coordinate should lie between 0 - 767.

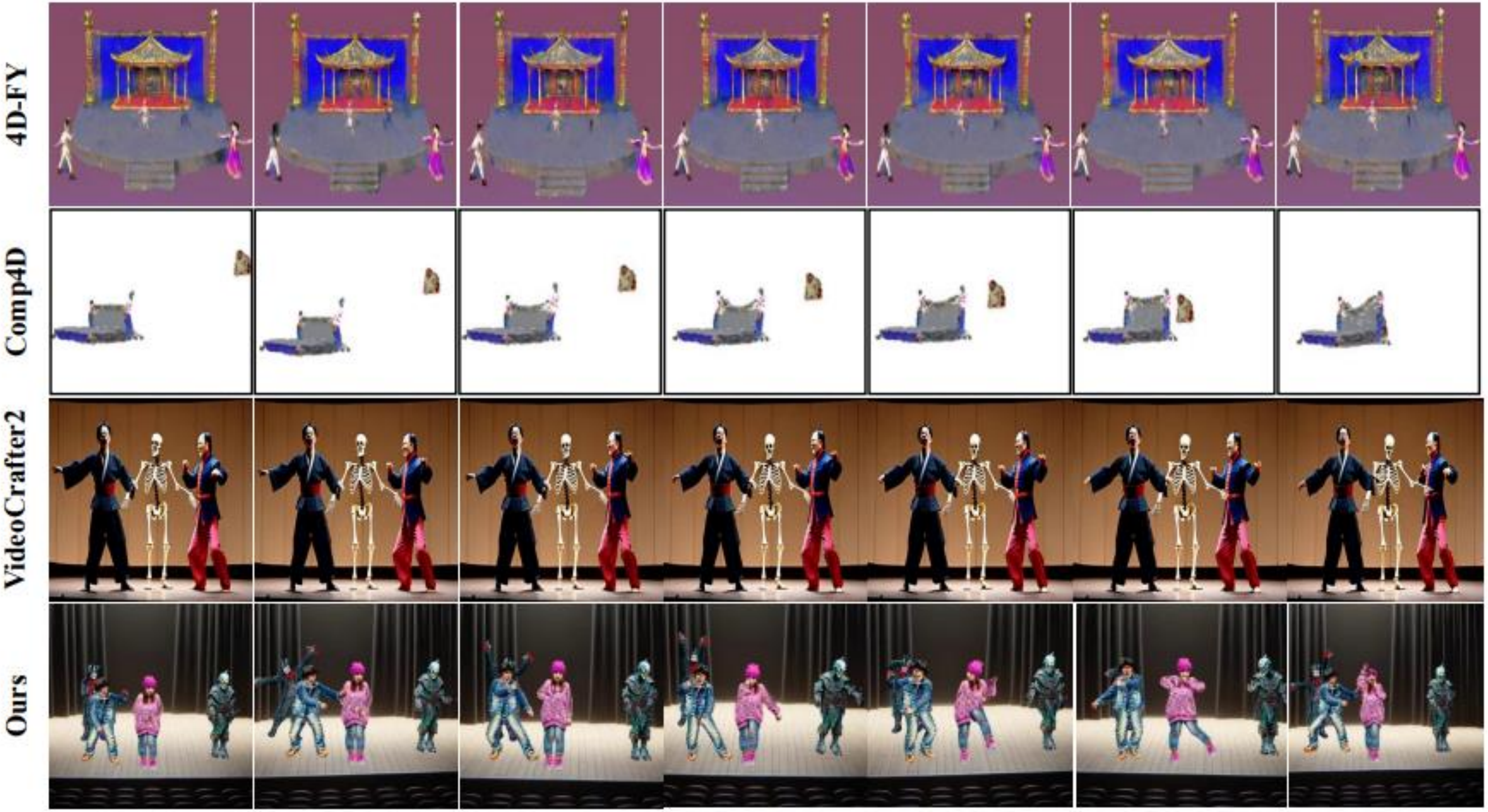
A3: Trajectory:

(380,450)
(408,461)
(439,475)
...
(624,580)
(666,608)
(704,652)

➢ Experiments:

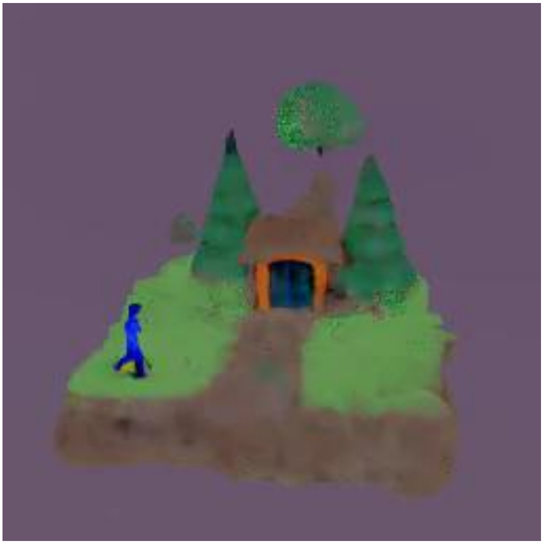


(a) Text prompt: "*In a Magician's magical cabin alone in a serene forest, an alien walking on the floor, starting from the cabin's door to the mow near the bottom right corner of this image*".



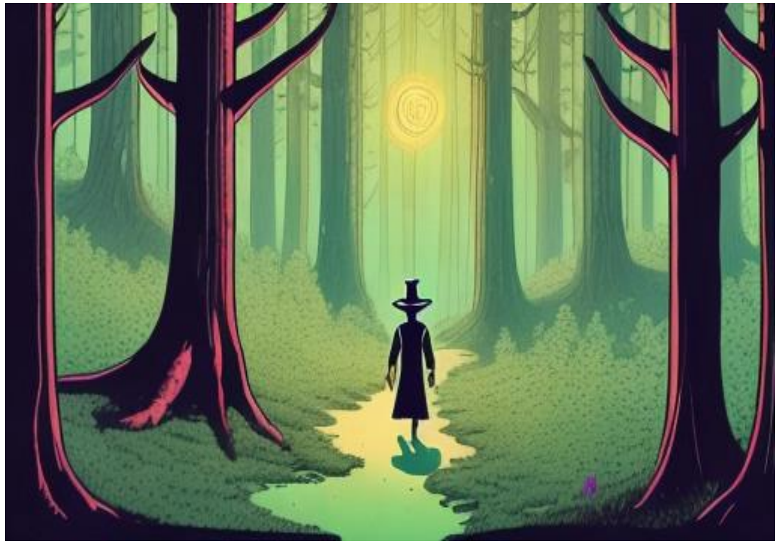
(b) Text prompt: "*Four characters stood on the stage. In front of the stage, a man and a woman are performing Kung Fu and dancing respectively. On the right side of the stage, a skeleton man is dancing, and behind them, a clown is performing*".

- **Visualization comparisons:**

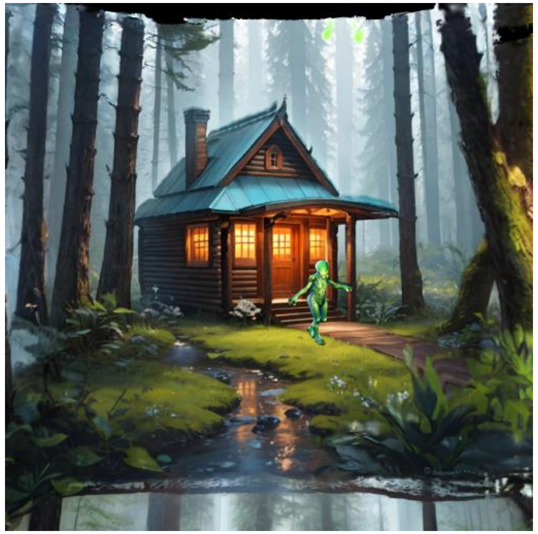


4D-FY

Comp4D

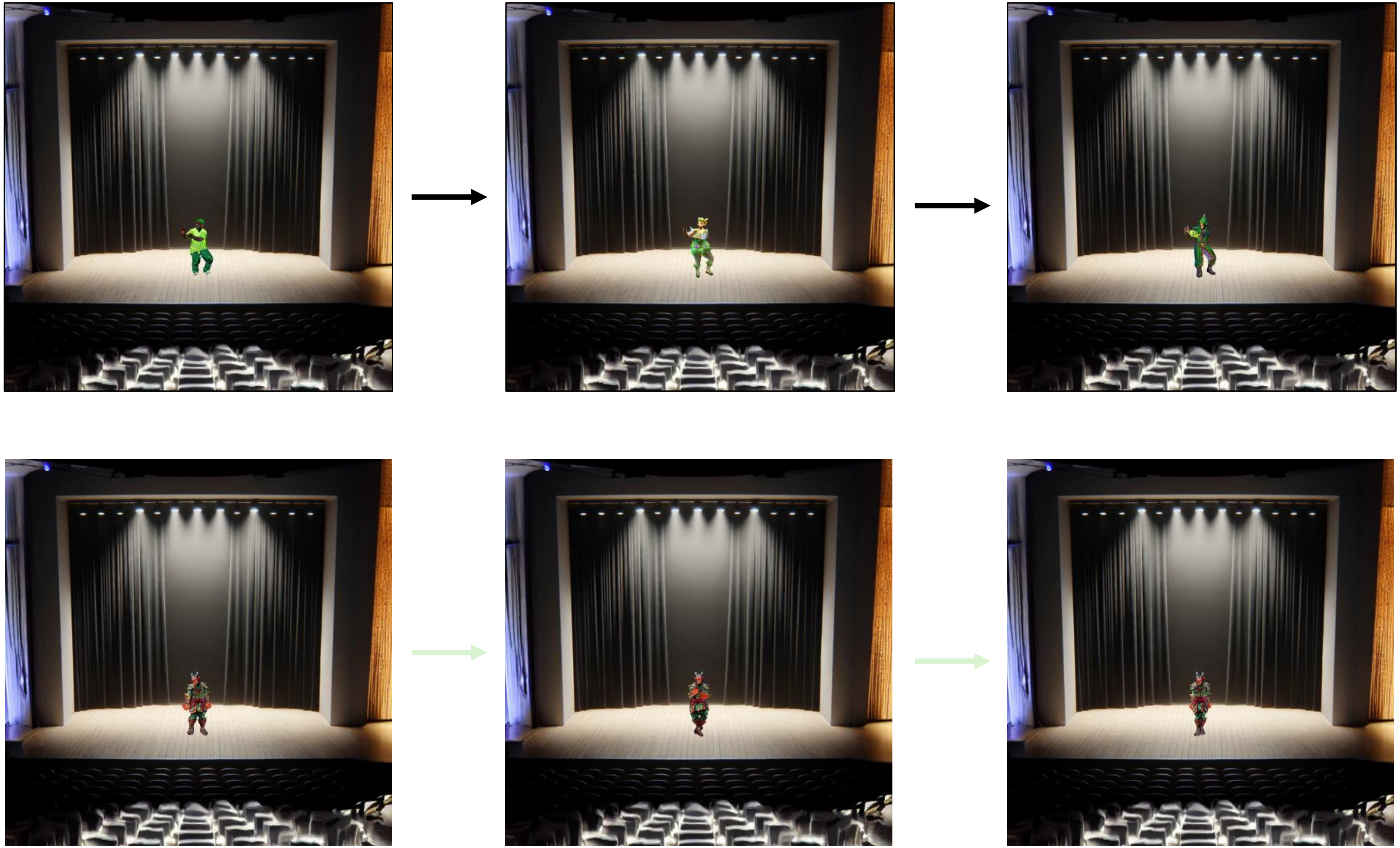


VideoCrafter2



Ours

➢ Visualization of editing results:



Thank You!