Boximator: Generating Rich and Controllable Motions for Video Synthesis

Jiawei Wang*, Yuchen Zhang*, Jiaxin Zou,
Yan Zeng, Guoqiang Wei, Liping Yuan, Hang Li

Bytedance Research

Traditional video synthesis

- Text + Image (optional) → Video

“A low angle shot of a man walking down a street, illuminated by the neon signs of the bars around him.”

(video from Gen-2 website)

Limitations of text-based motion control (I)

- Imperfect model is not able to comply to all text prompts.

A cat is jumping to a table.

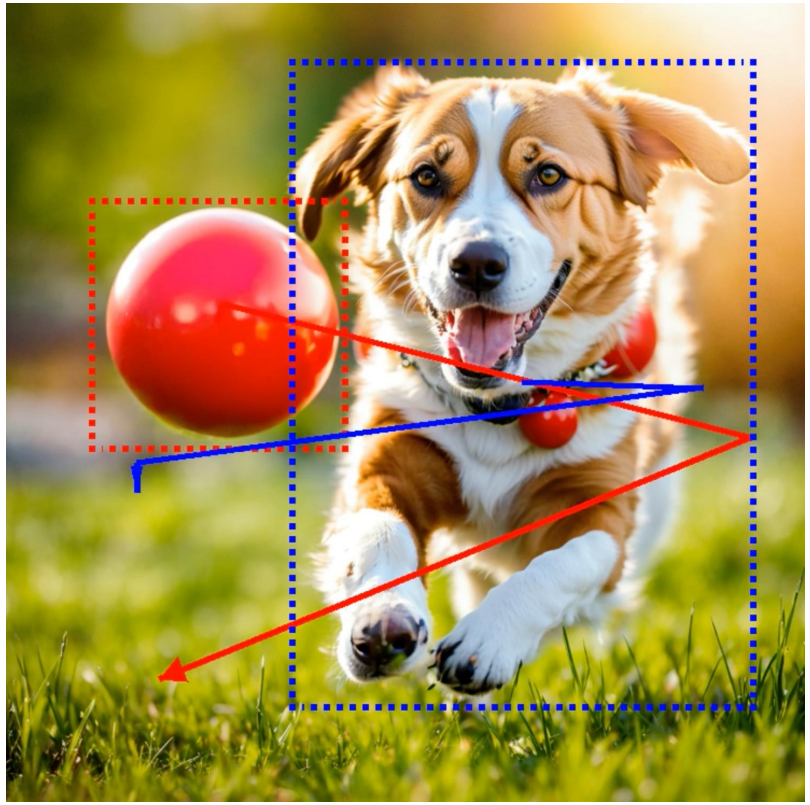
Why doesn't the water level change?

Adding wine to a glass.

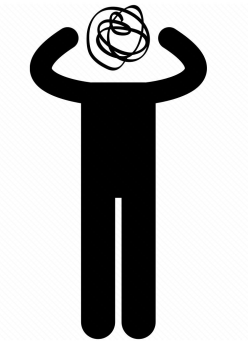
(videos generated by Gen-2

Limitations of text-based motion control (II)

- Position, shape, size, trajectory are not easy to express in text.



How to describe the desired motion?



The red ball flies to the lower right, reaching the middle of the right edge, then bounces back to fly to the lower left corner. The dog is chasing the ball and moves to the left direction, finally reaching the left edge?

Even if you can describe, the model will get confused!

Boximator can help!

(with a little bit user guidance)

A cat is jumping to a table.

Adding wine to a glass.

A dog is chasing a red ball.

(generated by PixelDance + Boximat

Two key questions in motion control

1. What to move?
2. How to move?

What to move?

- Traditional way: describe in text: “the man”, “the dog”, “the ball”.
- What about this?

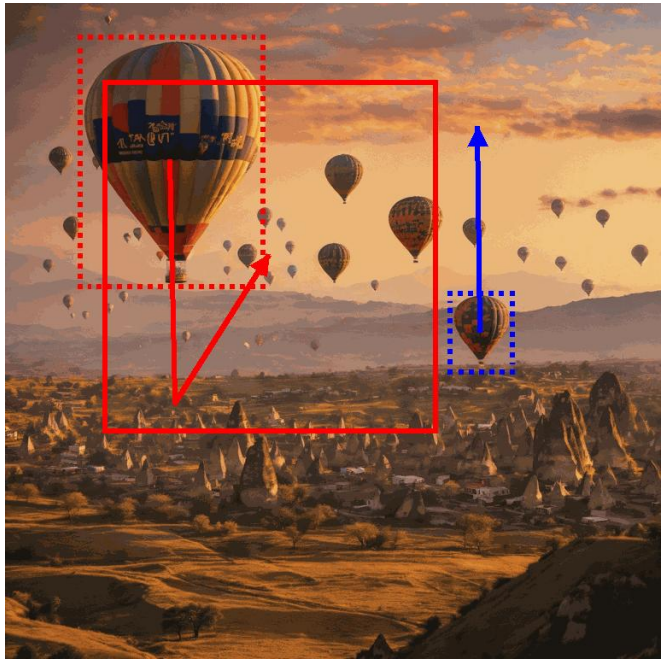
- Too hard for the model to understand complex descriptions, like “the left hand of the man in white shirt”.

Our way

- Select from image!

How to move?

- Draw boxes and/or motion paths!



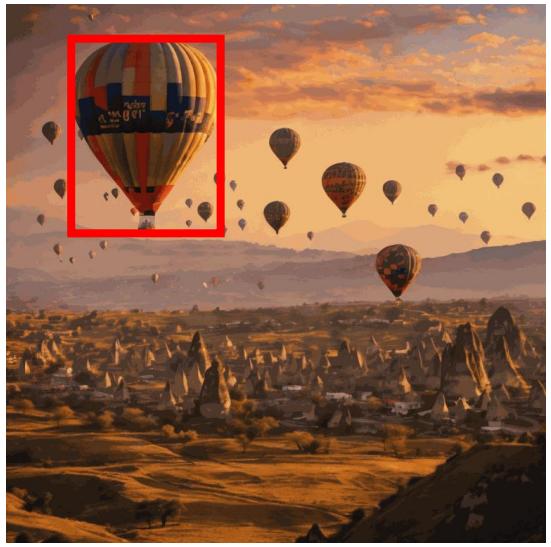
Object selection

Ending state

Motion path

A unified solution

- Boximator introduces two constraint types: hard box and soft box.



Hard box = object bounding box

Used for:

- Object selection
- Rigorously define ending state

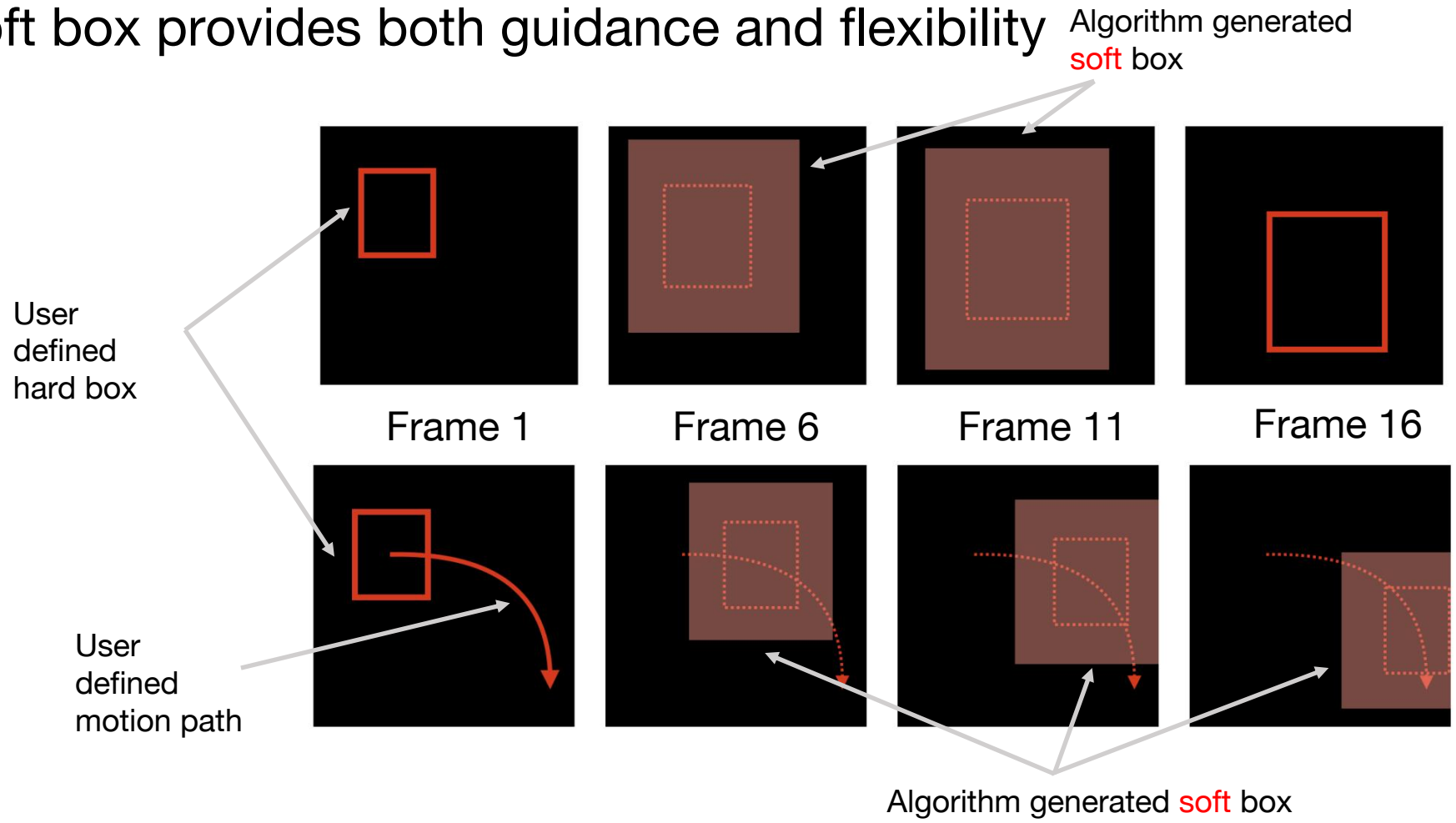
Soft box = a region where the object belongs to

Used for:

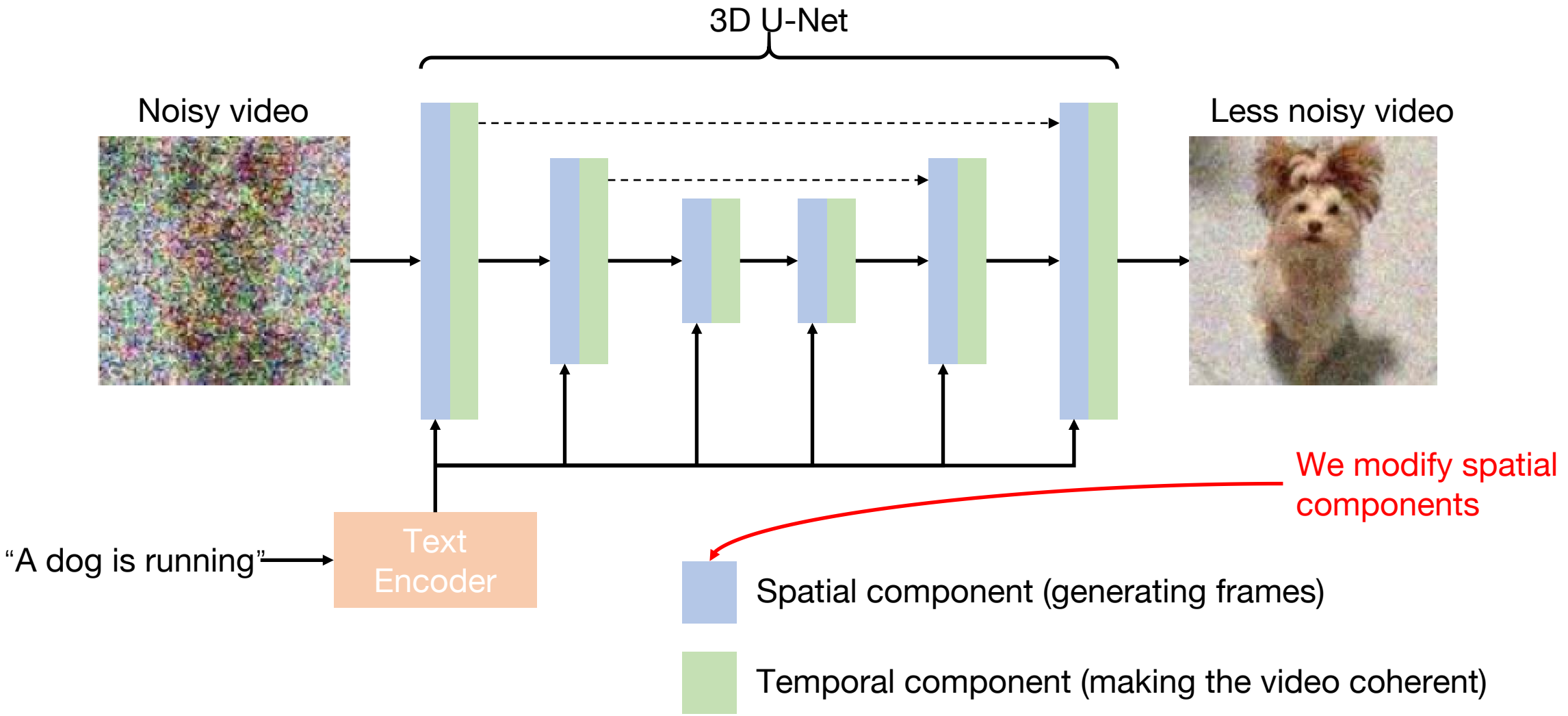
- Roughly define ending state
- Roughly define moving trajectory

Frame-level box constraints

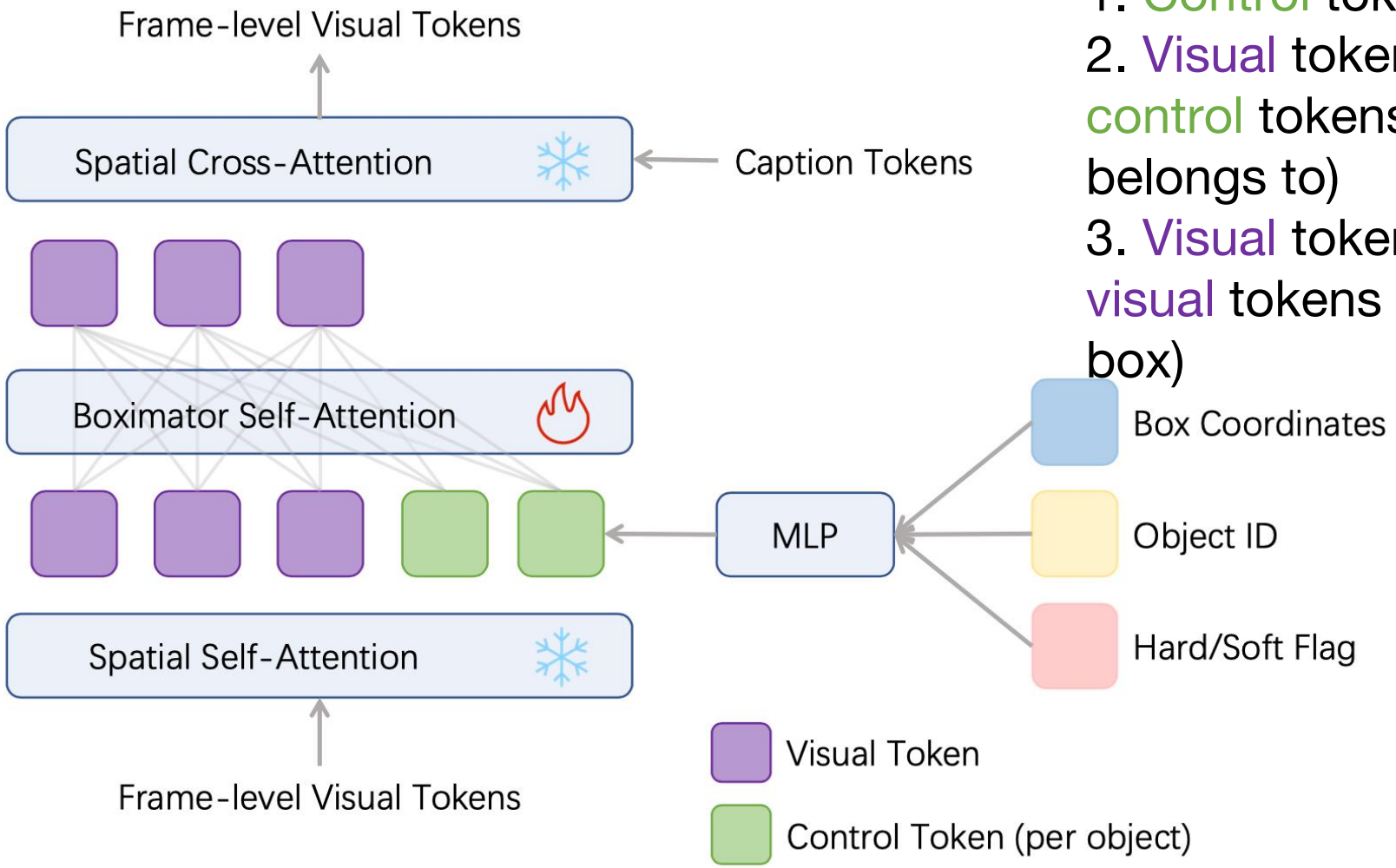
- Every object has a box constraint in every frame (hard or soft).
- Soft box provides both guidance and flexibility



Video diffusion model



Boximator control module

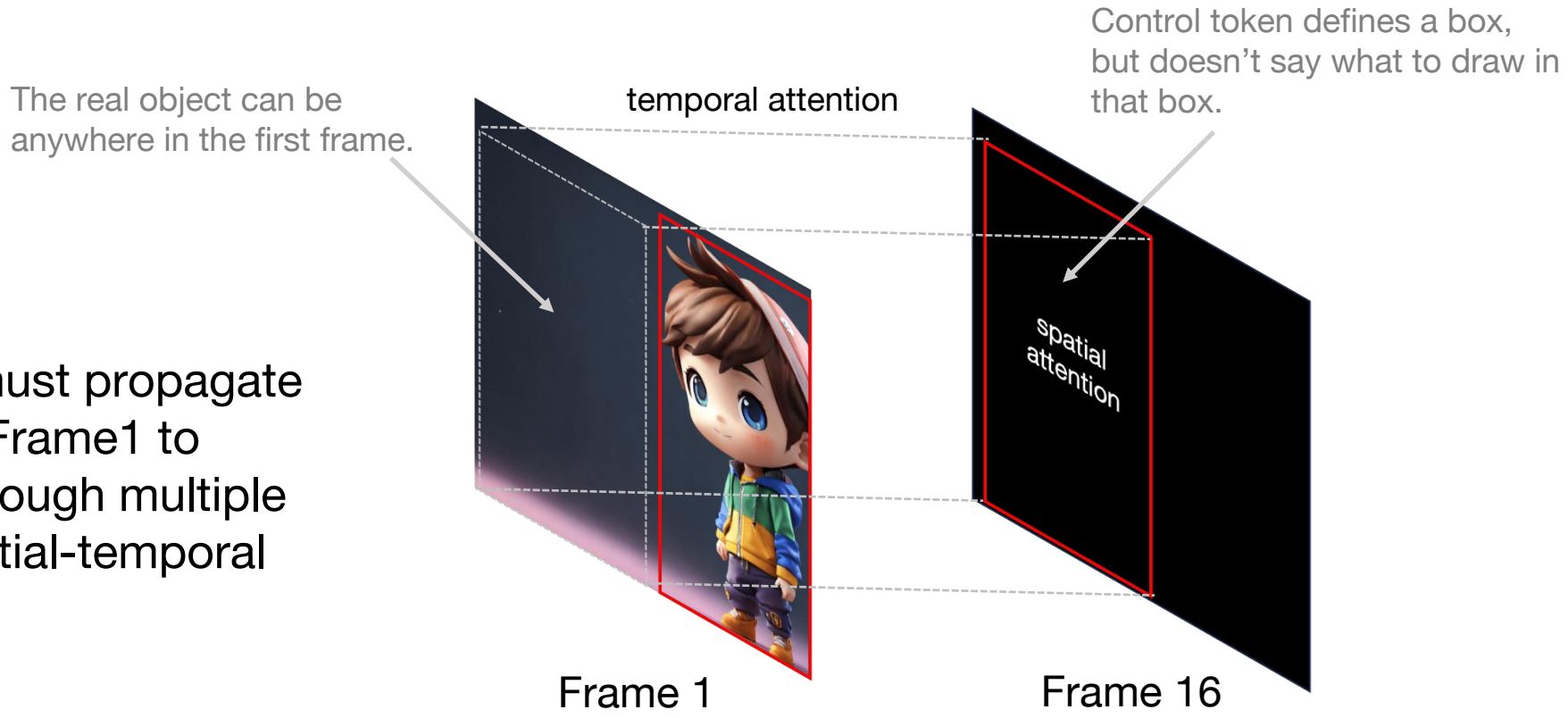


1. Control token contains box info
2. Visual token attends to related control tokens (of the boxes it belongs to)
3. Visual token attends to related visual tokens (that share the same box)

Training challenge

- Initially, we found Boximator very difficult to train – the model doesn't learn to put objects into the correct box.
- Why?

The model must propagate object from Frame1 to Frame 16 through multiple layers of spatial-temporal attention!



Self-tracking (I)

- **Simple idea**: train the model to generate colored bounding box for every object in every frame.

- (colored bounding box, object) are tied together
- Alignment chain:

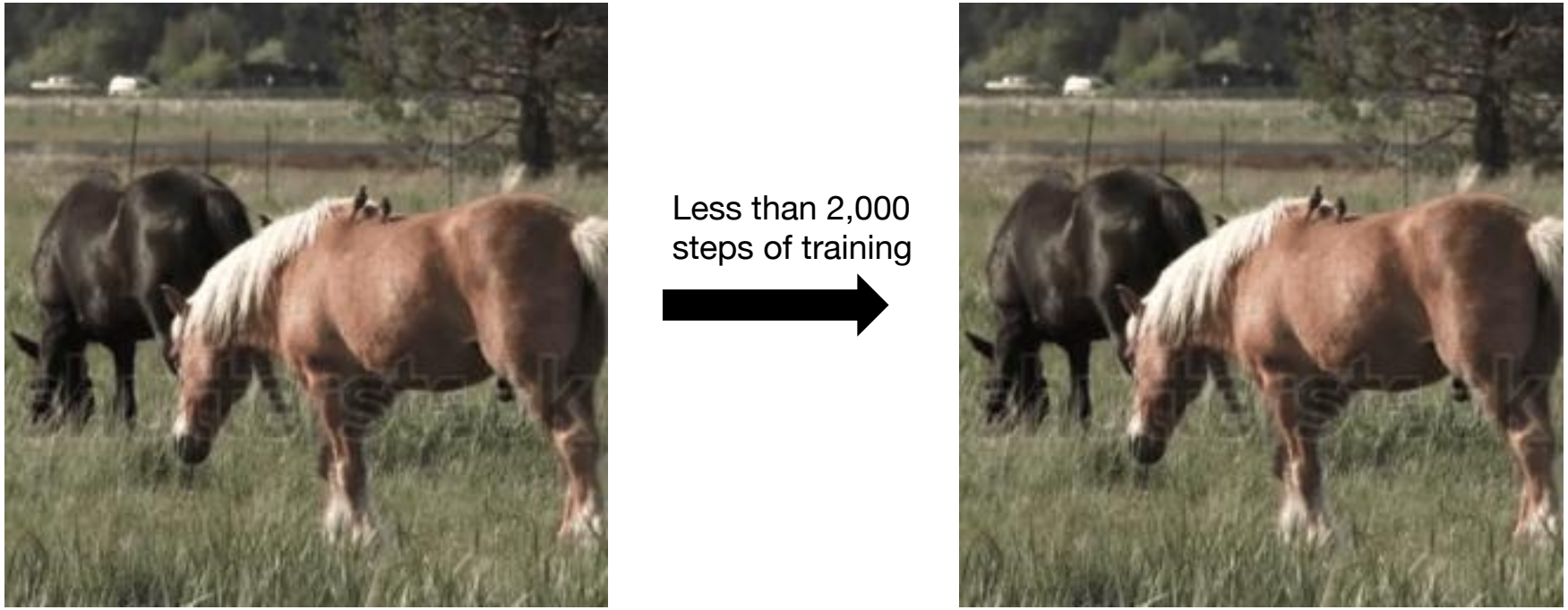
Spatial consistency: in each frame, align bounding box's position and color to the control token.

object → bounding box → Boximator constraint

Temporal consistency: the same object is always bound by the same color.

Self-tracking (II)

- After the model mastered motion control skills, we train it to stop generating visible bounding boxes.

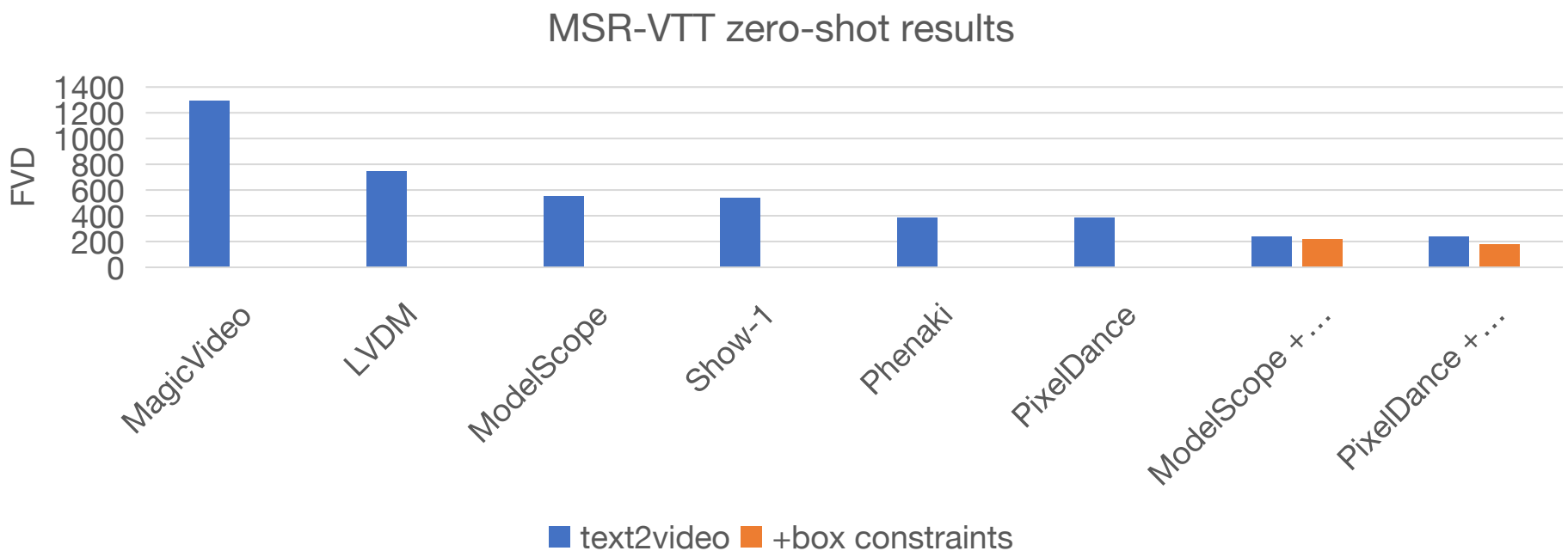


Data challenge

- Most videos in existing datasets (e.g. WebVid-10M) don't contain significant motion, and no bounding box is annotated.
- We created a large-scale, highly dynamic dataset where every video clip contains motion events, and multi-grained objects are annotated in every frame.

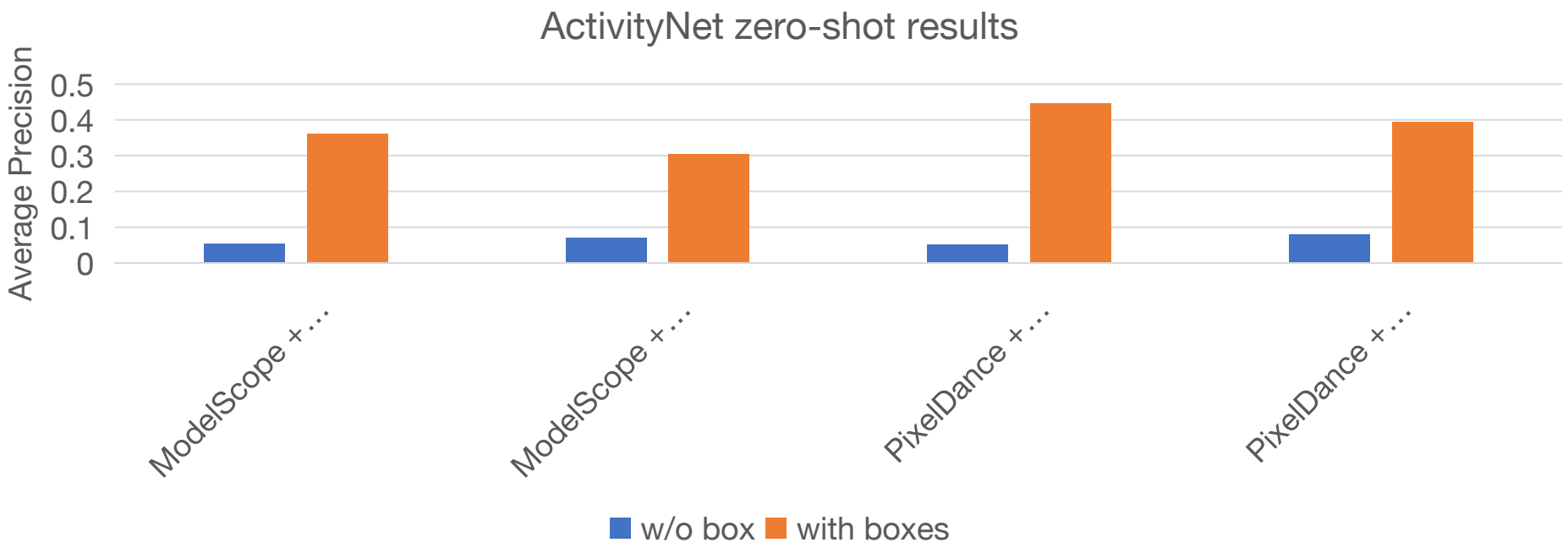
Empirical results: video quality

- Boximator achieves state-of-the-art video quality (lower the better). Adding box constraints further improves the score.



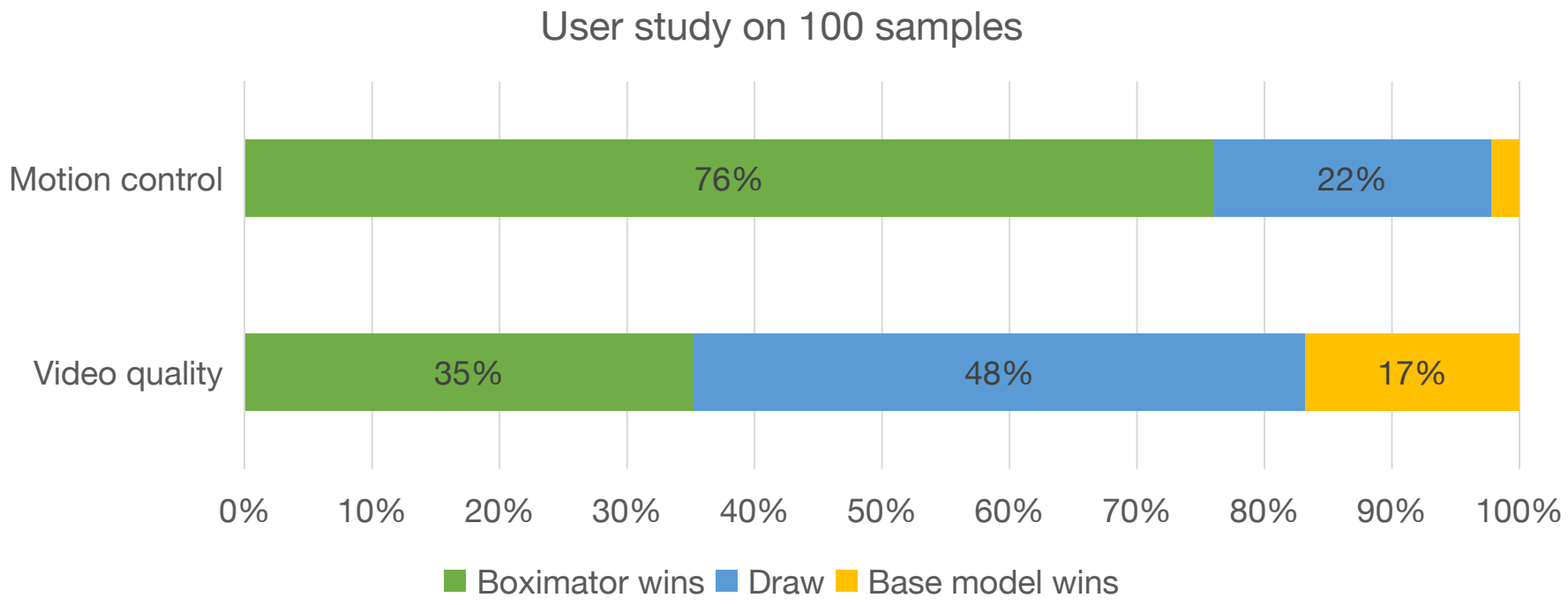
Empirical results: motion control

- In all settings, the motion control precision metric (higher the better) saw drastic increases when box constraints are applied.



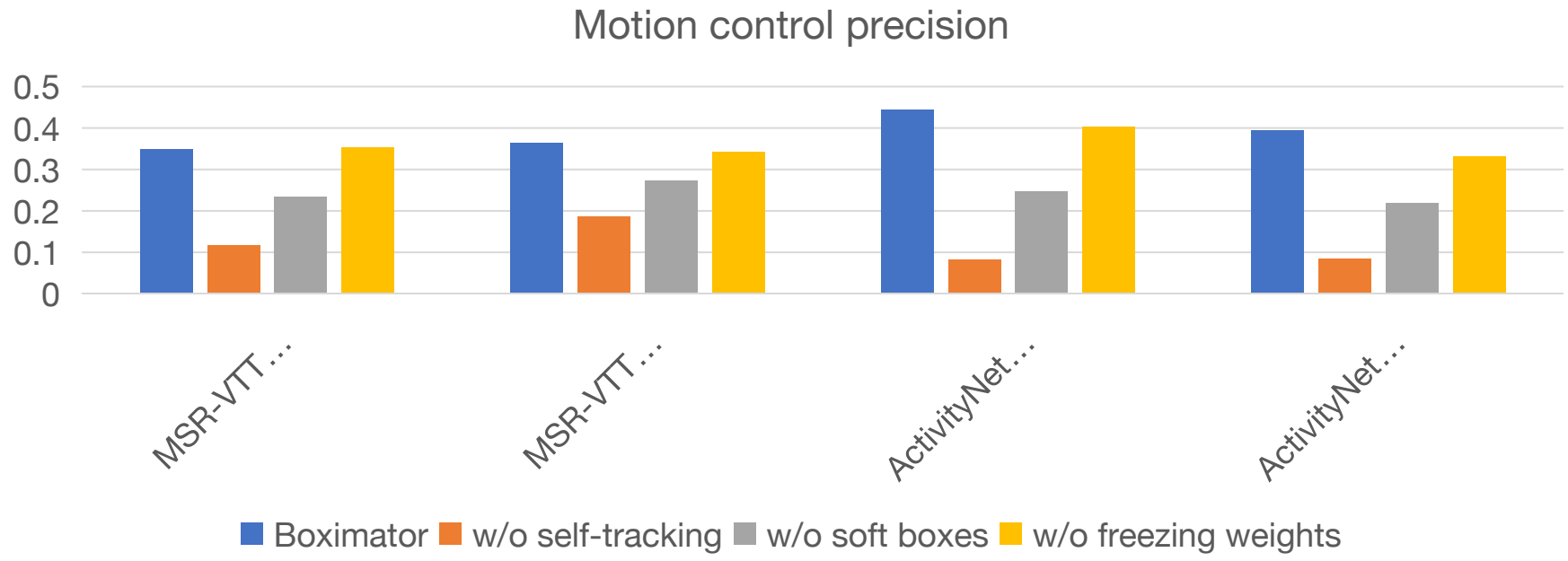
Empirical results: human evaluation

- Users favor the video quality and motion control of Boximator over the base model.



Empirical results: ablation study

- In all settings, self-tracking and soft boxes are essential to motion control; Freezing the base model doesn't hurt performance.



More examples

- Fine-grained control of character movements.

A boy and a girl are kissing

Two boxers are fighting

More examples

- Fine-grained control of human pose.

A man is drinking a cup of coffee

Anime girl playing piano.

More examples

- Control objects to move closer or farther away.

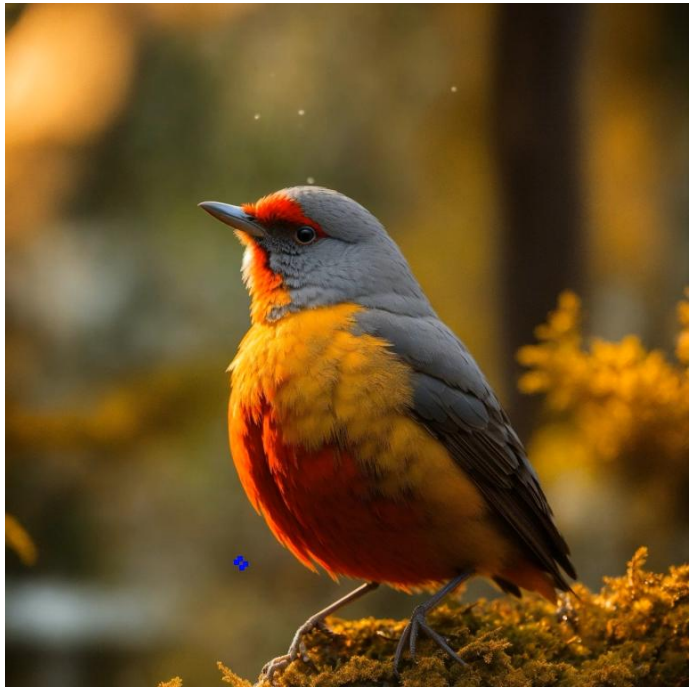
Spiderman swings towards the camera

A cute fluffy hamster pilot walking towards a fighter aircraft.

More examples

- Control objects to enter or leave the scene.

A girl is covering her face with hands.



A bird with grey, red and yellow feathers flies away.

More examples

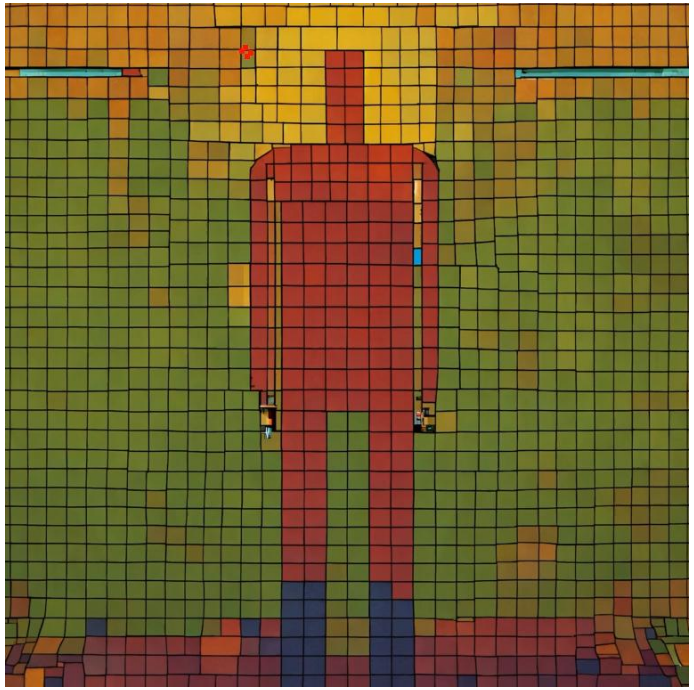
- Control the camera to move forward or rotate.

Drone flying over New Zealand beach

Camera rotate in a bedroom, showing a big landscape painting on the wall.

More examples

- Motion control for different painting styles.



The character made of pixels is dancing.

A wolf howled at the moon and then jumped down from the stone and ran away.

For more information

- Website: https://boximator.github.io/
- Paper: https://arxiv.org/abs/2402.01566
- Thanks!