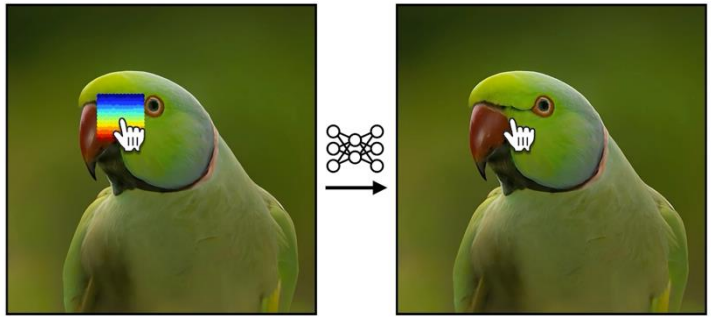
MotiMotion
Motion-Controlled Video Generation with Visual Reasoning

Lee Hsin-Ying[1] Hanwen Jiang[2] Yiqun Mei[2] Jing Shi[2] Ming-Hsuan Yang[1] Zhixin Shu[2]

[1]University of California, Merced [2]Adobe Research

Controlling Motion

User trajectories are often sparse and indicate only primary motion.



Motion Prompting

Tora

WanMove

→ What if the motion trigger an event or affect the environment?

Modeling Reactive and Secondary Motion

Motion should align with physics and common-sense knowledge.



Collision



Constraint Change



Airflow



Tool Mechanism



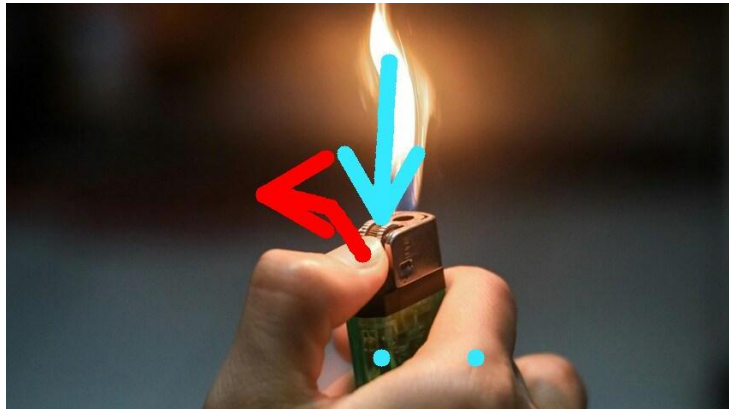
Common Objects

→ Current video generators can still fail to model such dynamics.

Aligning with Physics and Common Sense

A VLM predicts what may happen and how the objects move

User Trajectories



Predicted Trajectories

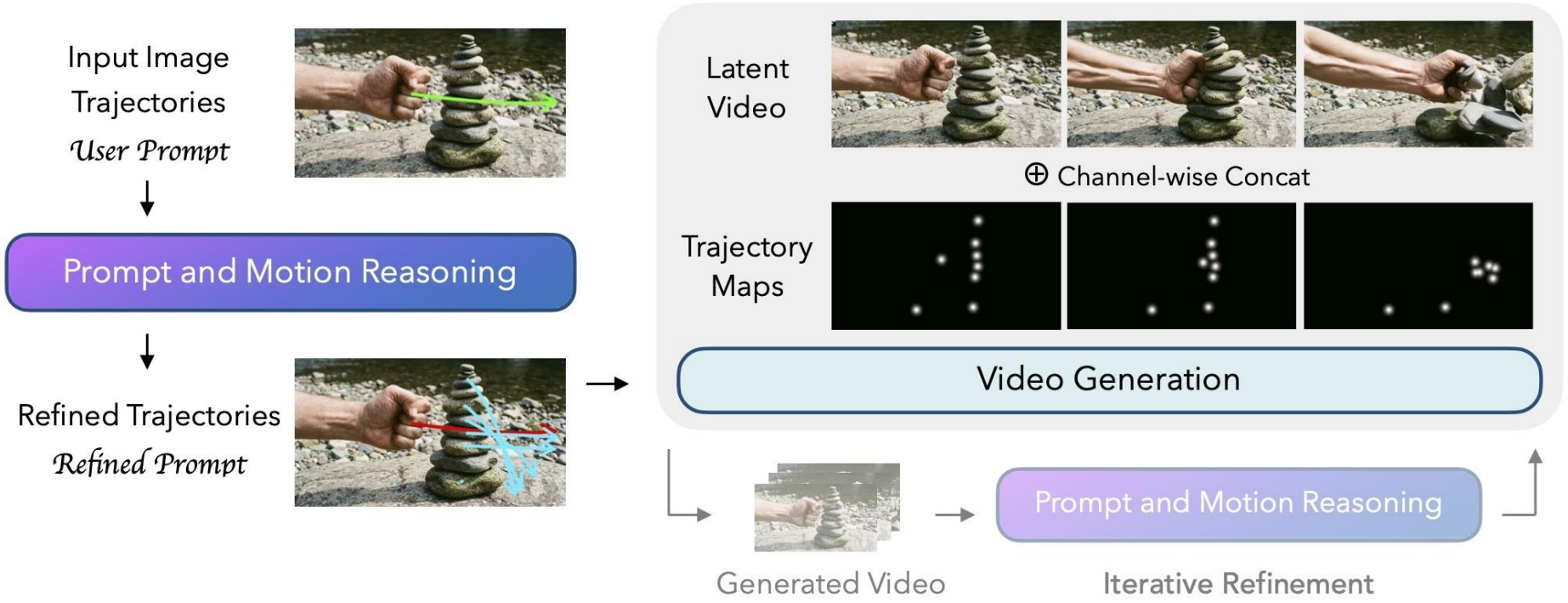
Red: refined user trajectories
Blue: proposed new trajectories

Video

Method

A Reasoning-then-Generation Pipeline

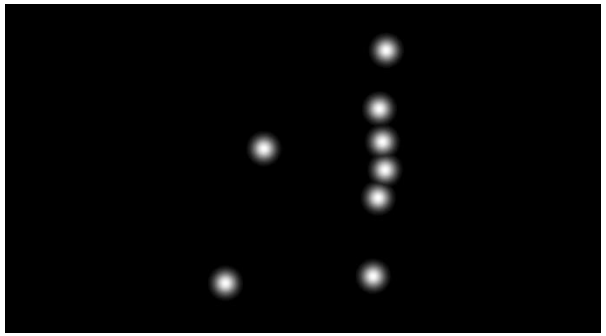
A VLM proposes a plan. Then a video generator executes the plan.



Motion-Controlled Video Generator

- First implement a motion-controlled generator with Gaussian maps
- Test with other motion-control approaches later

Trajectories



Control

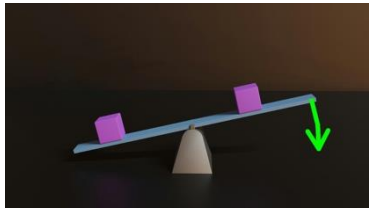
Video

Merged Video

Reasoning

Predict subsequent event and motion

Image with trajectory visualization



Trajectories

user_trajectories: (N, T, 2)

Prompt (Optional)

Press down on one side of the seesaw

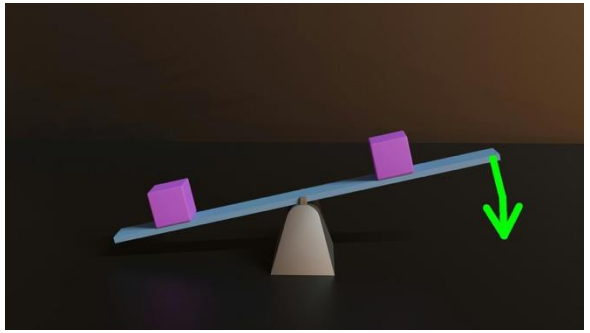
Prompt and Motion Reasoning

{
"refined_user_trajectories": (N, T, 2)
"proposed_new_trajectories": (N, T, 2)
}

An unseen force applies firm downward pressure to the elevated right side of the light blue seesaw plank …

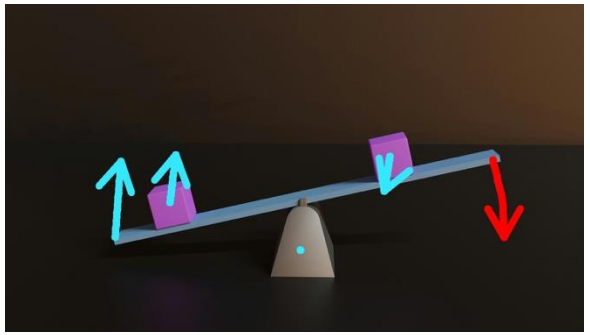
Motion Reasoning Example

Input and output are pure text of relative coordinates



```
[
  [
    [0.843, 475], [0.844, 488], ..., [0.862, 0.711]
  ]
]
```

Increasing y



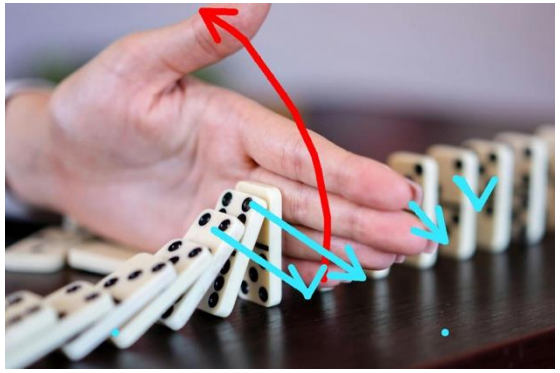
```
{
  "refined_user_trajectories": [
    [[0.842, 0.475], [0.846, 0.488], ..., [0.862, 0.711]]
  ],
  "proposed_new_trajectories": [
    [[0.190, 0.710], [0.191, 0.698], ..., [0.200, 0.480]],
    [[0.510, 0.750], [0.510, 0.750], ..., [0.510, 0.750]],
    ...
  ]
}
```

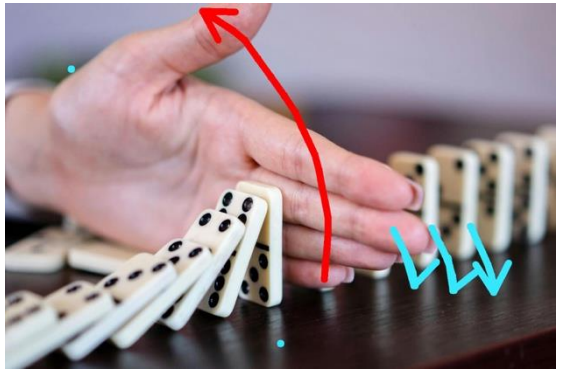
Decreasing y

Fixed point

Imperfect Trajectories

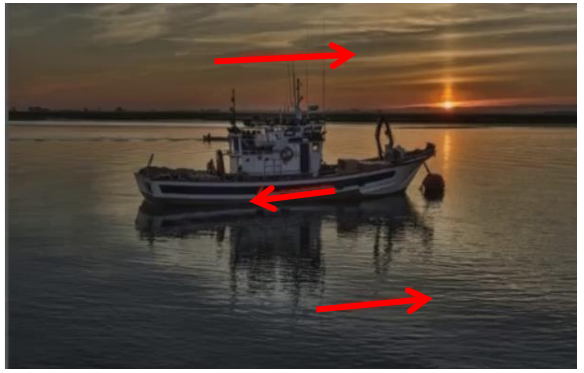
VLM predictions sometimes don’t align perfectly.





Shifted

Human trajectories are often smooth and simple.



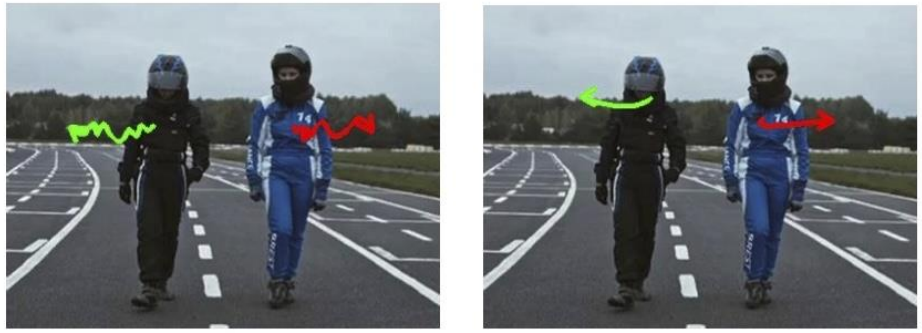
Human Annotation
(MotionPro)

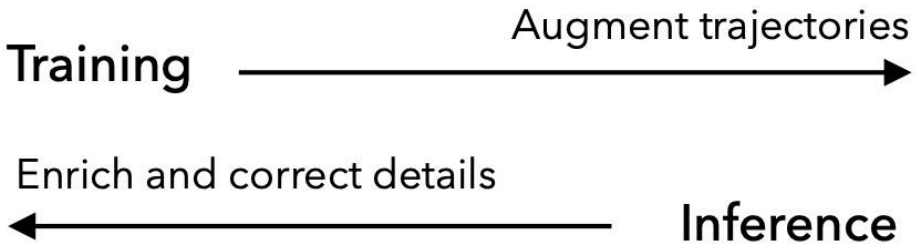
Tracking
(MoveBench)

Confidence-Aware Control

Learning to follow trajectories elastically according to confidence

- High confidence: follow strictly
- Low confidence: allow the video generator to improvise → scale Gaussian maps in trajectory control





Experiments and Results

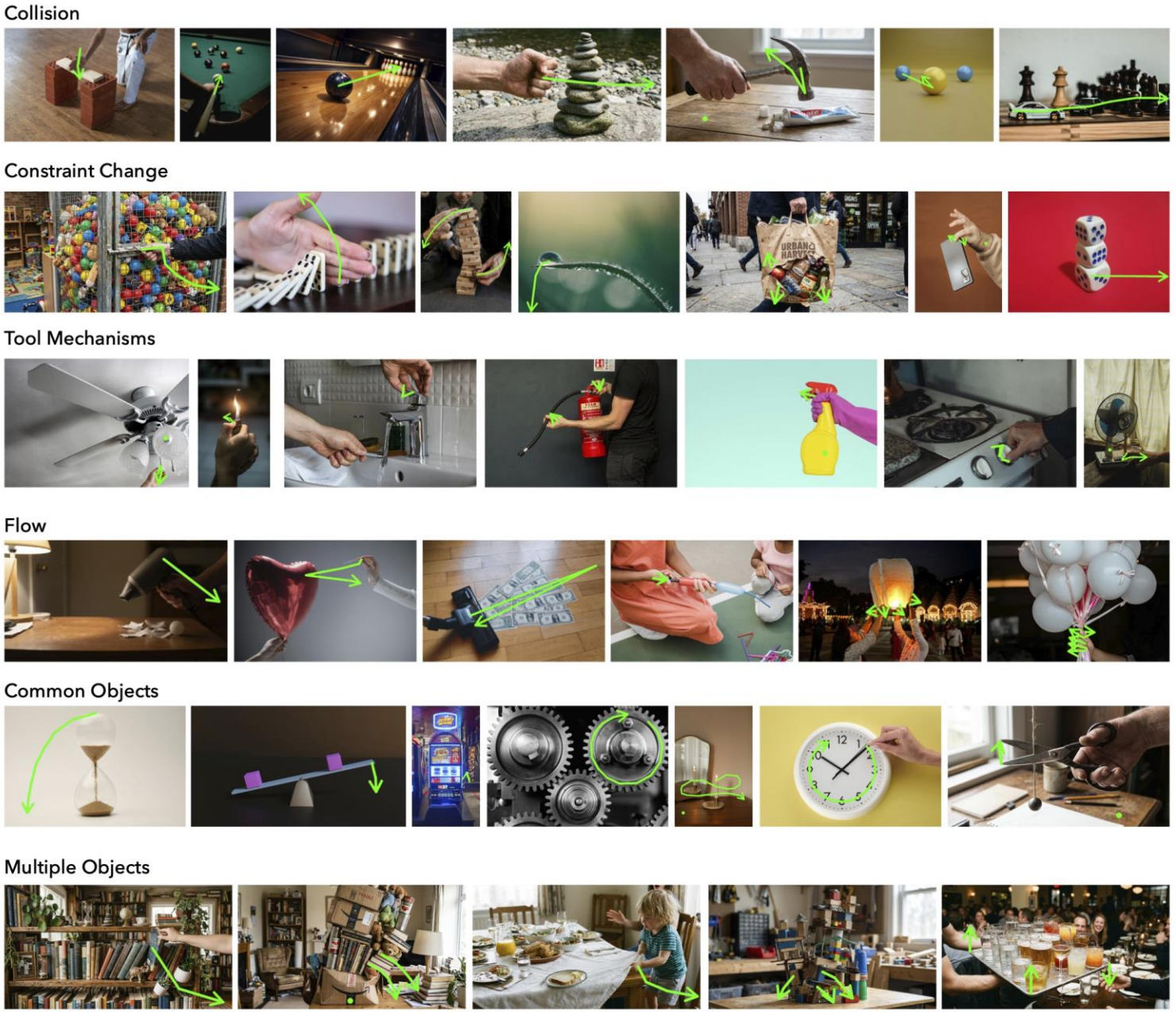
MotiBench

Pre-event scenes

Images, trajectories, and prompts

Category	Num of Samples	Percentage (%)
Collision	9	15
Constraint Change	17	27
Tool Mechanisms	8	13
Flow	9	14
Common Objects	19	31

Statistics



Evaluation

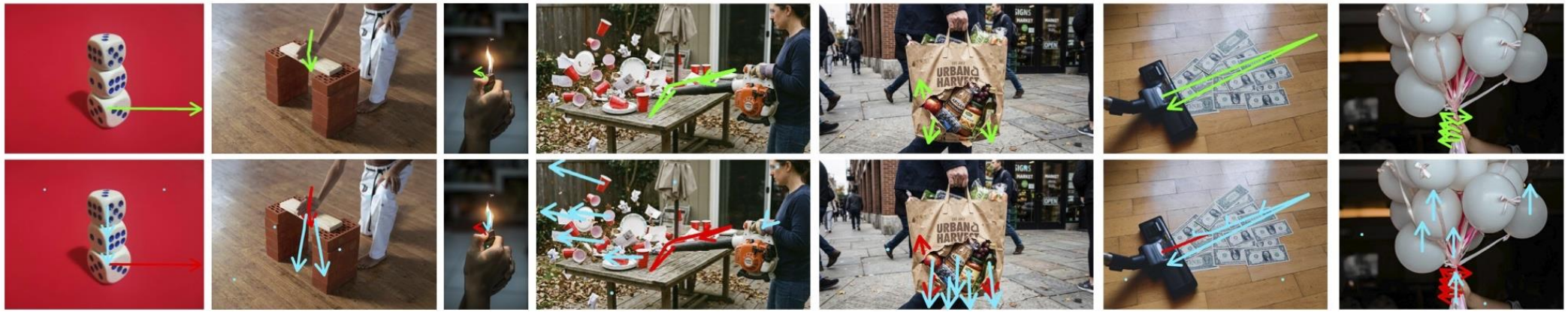
- Objective Metrics: VLM auto-evaluation following previous work
  - Physical Realism, Photorealism, and Semantic Consistency
- Preference Rate: VLM and user study
  - Object properties and interaction

VLM Prediction

User

VLM

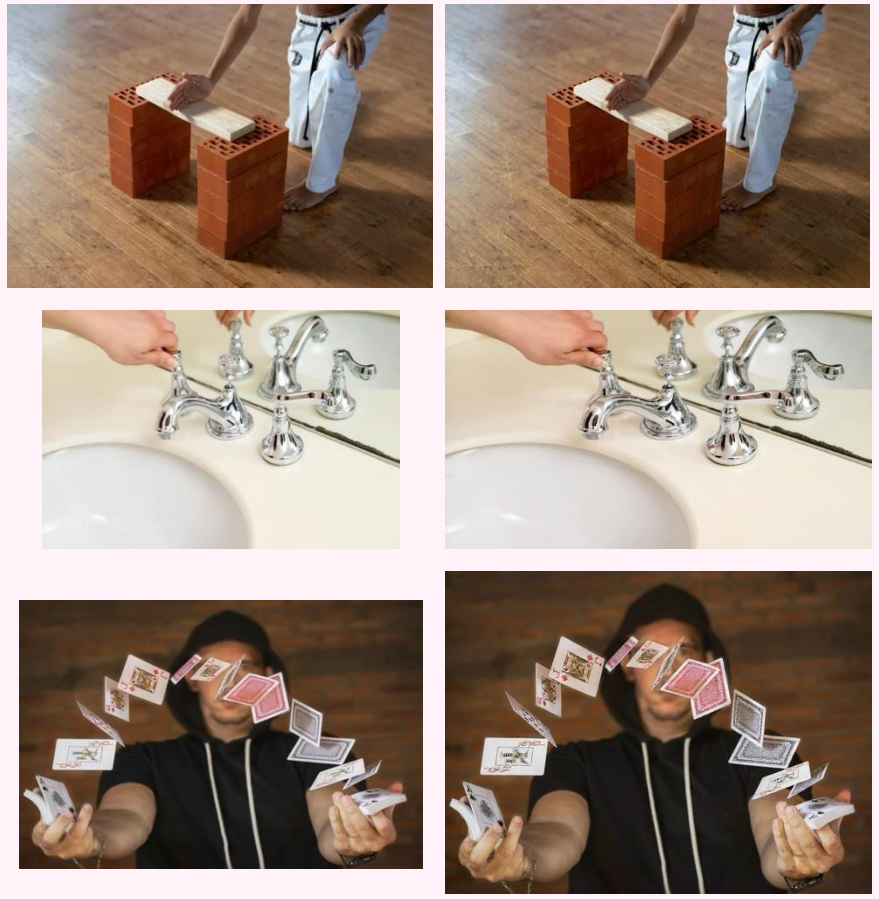
User



VLM

Comparison

Input Prediction



MagicMotion WanMove

MotiMotion (Ours)

More results in https://motimotion.github.io

Comparison

Method	Physical↑	Photo↑	Semantic↑
MagicMotion (Li et al., 2025a)	0.157	0.550	0.343
Wan-Move (Chu et al., 2025)	0.218	0.483	0.511
MotiMotion	**0.302**	**0.520**	**0.665**

VLM Auto-Evaluation

	Win Rate of MotiMotion (%)			
against Baselines	Obj.	Int.	Overall	Human
MagicMotion (Li et al., 2025a)	72.9	80.8	78.0	97.9
Wan-Move (Chu et al., 2025)	71.5	75.0	73.8	81.4

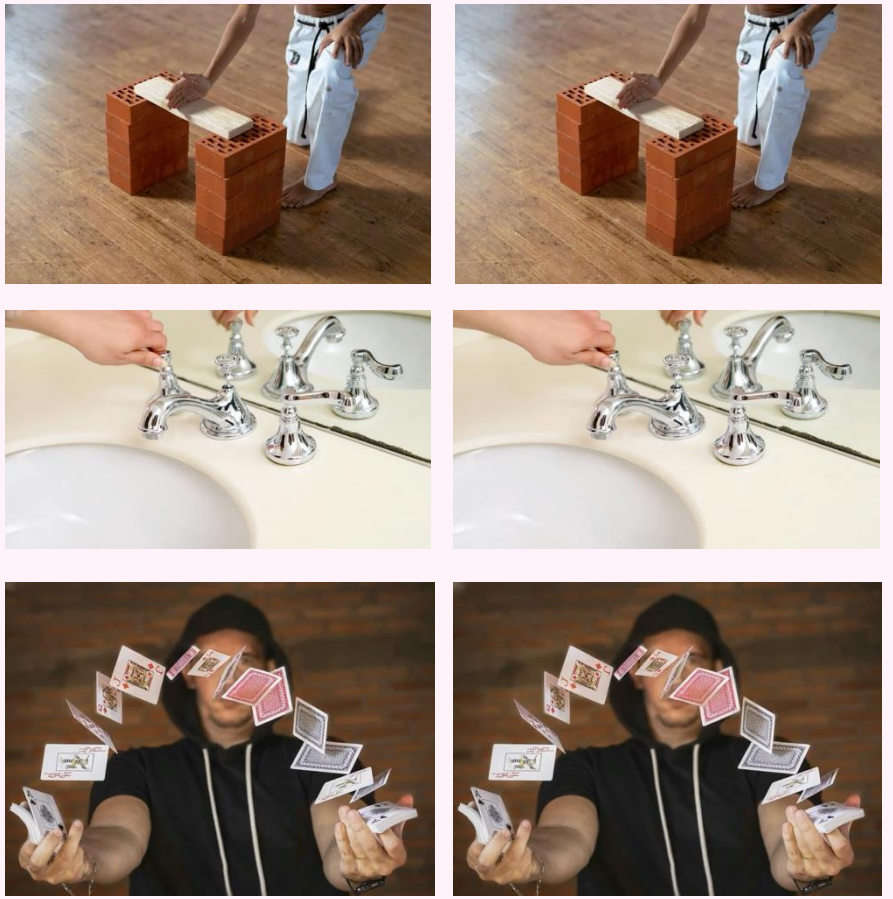
Preference Rate (Ours vs Baselines)

Method	Physical↑	Photo↑	Semantic↑
MagicMotion (Li et al., 2025a)	0.157	**0.550**	0.343
+ Reasoning	**0.199**	0.528	**0.427**
Wan-Move (Chu et al., 2025)	0.218	0.483	0.511
+ Reasoning	**0.283**	**0.488**	**0.588**

Reasoning for Other Approaches

Ablation Study

Input Prediction



w/o Reasoning w/o Motion Reasoning

MotiMotion (Ours)

More results in https://motimotion.github.io

Ablation Study

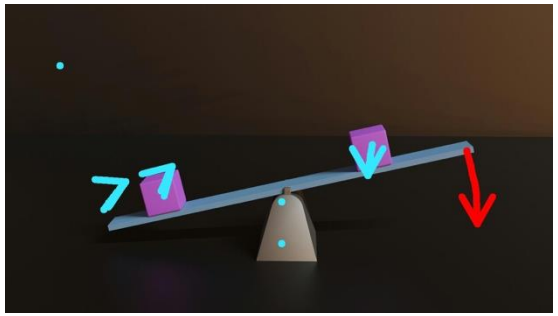
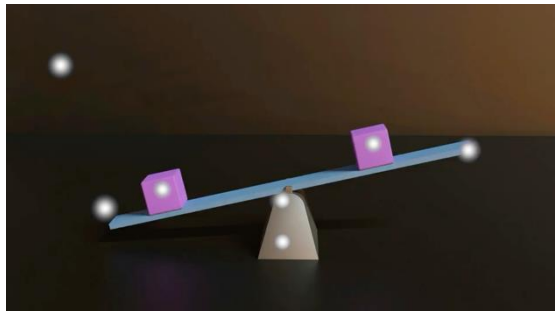
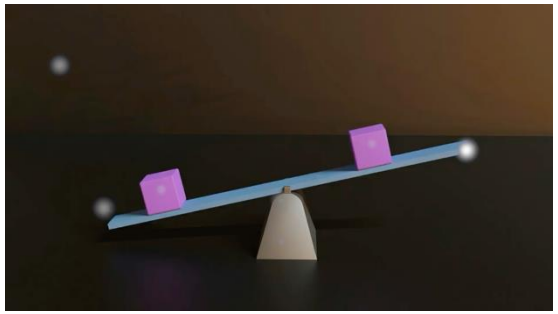
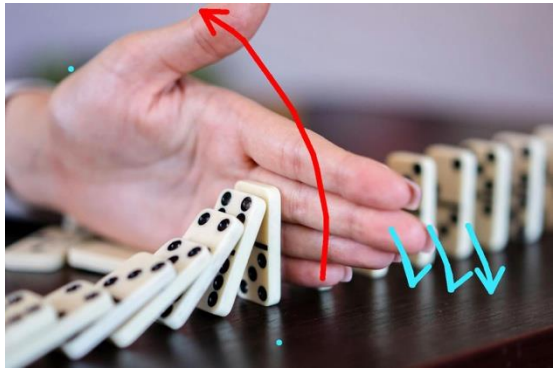
Method	Physical↑	Photo↑	Semantic↑
Motion-Controlled Generator	0.166	0.389	0.337
+ Prompt Reasoning	0.237	0.475	0.544
+ Motion Reasoning	0.285	0.493	0.641
+ Confidence-Aware Control	**0.302**	**0.520**	**0.665**

Main Components

Method	Physical↑	Photo↑	Semantic↑
Image + Trajectories	0.177	0.353	0.272
+ Prompt Reasoning	**0.229**	**0.452**	**0.473**

Prompt Reasoning w/o Input Prompts

Ablation Study

Prediction

High Confidence

Low Confidence

Summary

- Identify the gap in reasoning about visual context for motion-controlled video generation and address it with VLMs.
- Propose MotiMotion: VLM-driven semantic and motion control, with confidence modulation for natural motion.
- Curate MotiBench, a benchmark for physical and causal reasoning, where MotiMotion is preferred over prior methods.