C-Drag-Official-Repo

C-Drag: Chain-of-Thought Driven Motion Controller for Video Generation

Yuhao Li, Mirana Claire Angel, Salman Khan, Yu Zhu, Jinqiu Sun, Yanning Zhang and Fahad Khan

Northwestern Polytechnical University, Mohamed bin Zayed University of AI, Australian National University, Linköping University

arXiv Paper

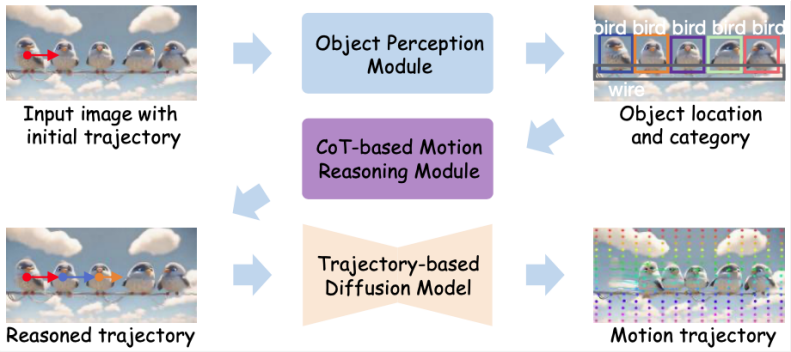
Latest

- 2025/02/12: We released our code and benchmark.
- 2025/02/12: We released our technical report on arxiv. Our code and models are coming soon!

▸ Abstract

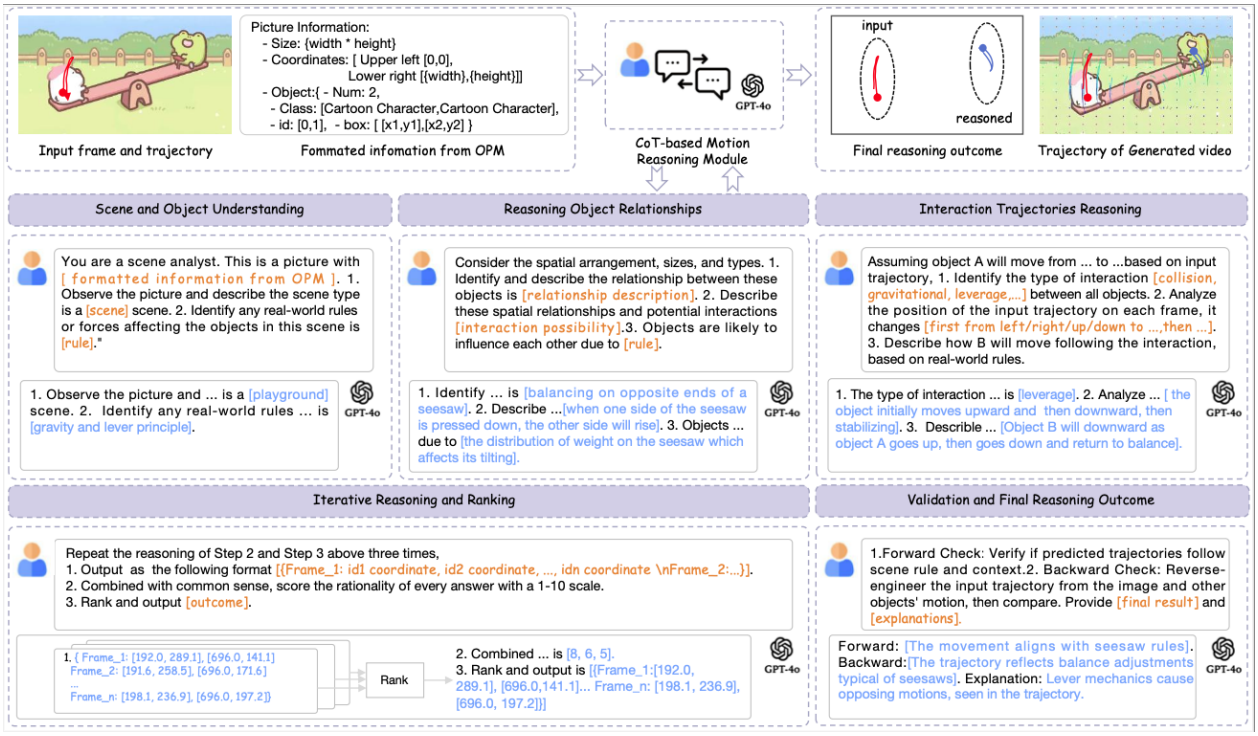
Intro

- **C-Drag** first takes a single RGB image and one or more drag motion trajectories as input. We employ an object perception module to obtain information about all related objects in the image. Chain-of-Thought (CoT)-based reasoning module introduces a reasoning strategy to precisely reason motion trajectories of all objects according to the detected position and category information. With the generated object trajectories, we use a pre-trained trajectory-based generation model to generate the videos with multiple-object interactions.



- **CoT-based Motion Reasoning Module** An illustrative view of CoT-based Motion Reasoning Module which undergoes a five-stage reasoning process. **Scene and Object Understanding**, where a pre-trained visual language model (VLM) interprets the scene and establishes motion rules using formated information from Object Perspection Module. In **Reasoning Object Relationship**, the VLM identifies spatial relationships and potential interactions among objects to inform trajectory predictions. **Interaction Trajectories Reasoning** follows, categorizing interactions (e.g., collisions, forces) and predicting affected object paths. During **Iterative**

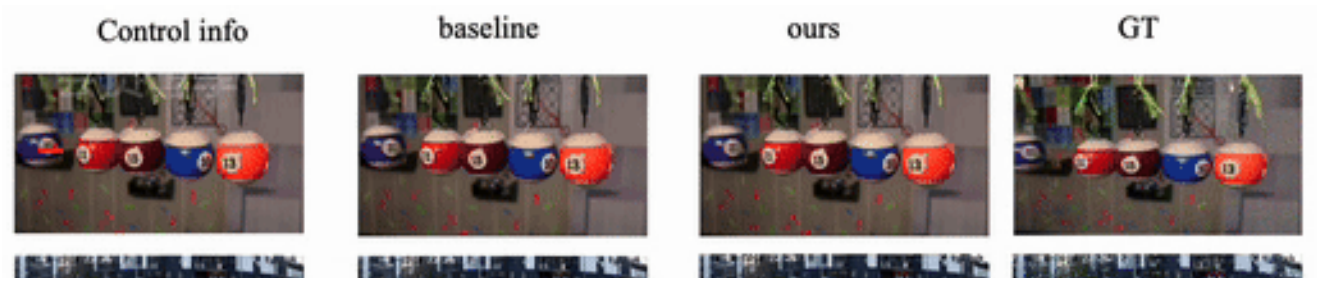
Reasoning and Ranking, initial predictions are iteratively optimized, with the VLM selecting the most consistent motion sequences. Finally, in **Validation and Final Reasoning Outcome**, forward and backward validation ensures predicted trajectories align with scene rules, iterating until accuracy is achieved.

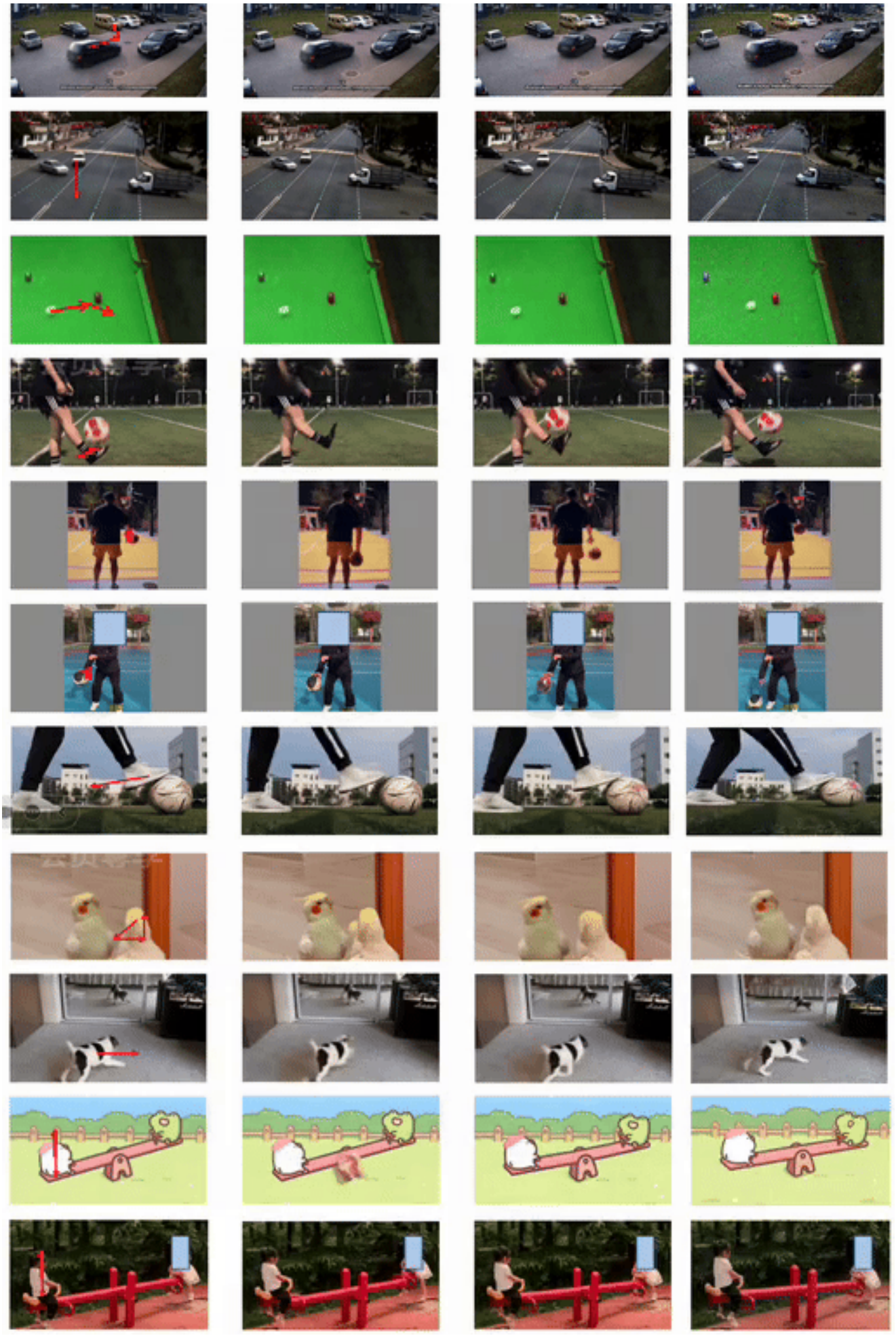


- We propose a new VOI dataset. This dataset has 72 videos and contains three typical types of object interactions, including *collision and chain reaction*, *gravity and force*, and *levers and mirrors*. We counted the number of videos, annotated boxes, and the objects trajectories.

Category	Sub-category	Video	Anno Boxes	Anno Trajectories
Collision and Chain Reaction	Billiard	16	2160	198
	NewtonCradle	7	300	79
	Traffic	10	180	90
Gravity and Force	Basketball	6	1080	34
	FootBall	7	960	76
Levers and Mirrors	Seesaw	15	840	145
	Mirror	11	1800	89
Total	-	72	7320	711

Visualization





Getting Start

Setting Environment

```
git clone https://github.com/WesLee88524/C-Drag-Official-Repo.git
cd C-Drag-Official-Repo

conda create -n C-Drag python=3.8
conda activate C-Drag
pip install -r environment.txt
```

Download Pretrained Weights

Download the Pretrained Weights to `models/` directory or directly run `bash models/Download.sh`.

Drag and Animate!

```
python demo.py
```

It will launch a gradio demo, and you can drag an image and animate it!

Citation

if you use our work, please consider citing us:

```
@misc{li2024multigranularity,
      title={Multi-Granularity Language-Guided Multi-Object Tracking},
      author={Yuhao Li and Muzammal Naseer and Jiale Cao and Yu Zhu and
Jinqiu Sun and Yanning Zhang and Fahad Shahbaz Khan},
      year={2024},
      eprint={2406.04844},
      archivePrefix={arXiv},
      primaryClass={cs.CV}
}
```

License

This project is released under the Apache license. See LICENSE for additional details.

Acknowledgement

We appreciate the open source of the following projects:
DragNUWA, DragAnything;