CinemaTraj

Composing Atomic Camera Trajectories for 3D Scenes with LLM Agents

Qianru Li[1,2*] · Xuyang Chen[1,2*] · Erkin Türköz[2] · Lu Liu[2] · Xuqin Wang[1,2] · Liqiu Meng[1] · Tao Wu[2] · Yanfeng Zhang[2†]

[1] Technical University of Munich [2] Huawei Hilbert Research Center

* Equal contribution † Corresponding author

Main Track · Paper 2581 · DOI: 10.1145/3767308.3835366

01 Problem, Motivation & Challenges

TASK DEFINITION

Given RGB-D views of a static 3D scene and a text request, generate a **continuous 6-DoF camera trajectory** and a rendered cinematic video.

MOTIVATION

Virtual tours and real-estate videos need expert camera design. Text-driven generators can miss geometry; geometric planners lack shot semantics.

INPUT
RGB-D Views + Text Prompt → **CinemaTraj** → **OUTPUT**
Camera Trajectory + Video
Optional subtitles + voiceover

1 · Grounding Intent

Which objects, viewing order and viewpoints does the request imply?

2 · Preserving Shot Style

How can an orbit remain an orbit after geometric refinement?

3 · Respecting Geometry

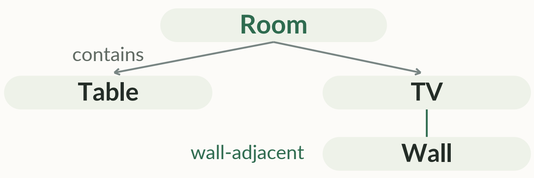
How can motion avoid obstacles and keep target objects visible?

02 Method: Ground, Compose, Refine

Core Idea: Ground the LLM in 3D and refine shot parameters while preserving cinematic structure.

A Build a Scene Graph

Lift RGB-D observations into object boxes, room containment and structural relations.



Feasibility: wall-adjacent targets rule out full orbits.

B Translate Intent

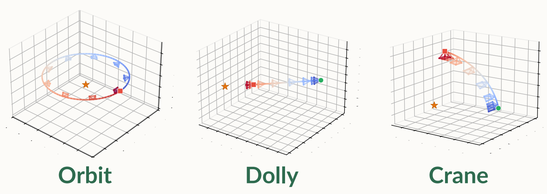
The LLM selects object IDs, shot types, order, elevation and viewing distance.

- anchor(table)
- orbit(table)
- anchor(sofa)
- arc(table → sofa)
- crane(sofa)

Plan: alternate object shots with arcs between target viewpoints.

C Compose Atomic Shots

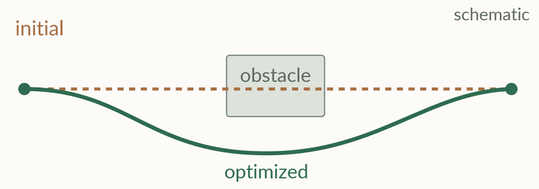
Select an anchor viewpoint by visibility and framing; instantiate a parametric shot.



Structure: tune radius, angles or distance within each shot type.

D Refine Motion

A signed distance field (SDF) guides camera clearance and target visibility.



Two passes: refine object shots, then reconnect arcs through doorways.

03 Qualitative Results

INPUT · OPEN-ENDED PROMPT

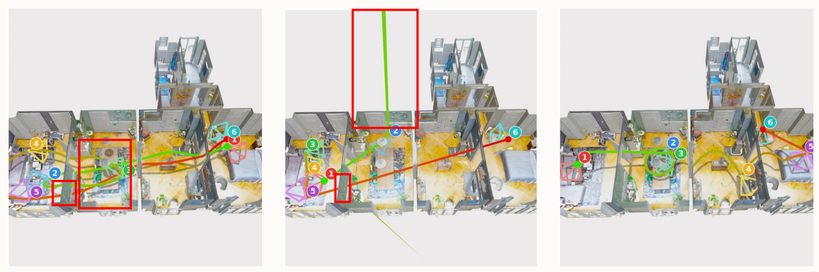
“Can you give an incoming Airbnb guest a detailed walkthrough of this house?”

OUTPUT · CAMERA TRAJECTORY + SIX RENDERED KEYFRAMES



BASELINE COMPARISON

Fixed prompt: dolly to sofa, orbit table, zoom PC, crane bed, then tilt to chair.



Observation: baseline paths cut through geometry; ours follows the requested object sequence with structured motion. Red boxes mark reported failures.

04 Quantitative Results

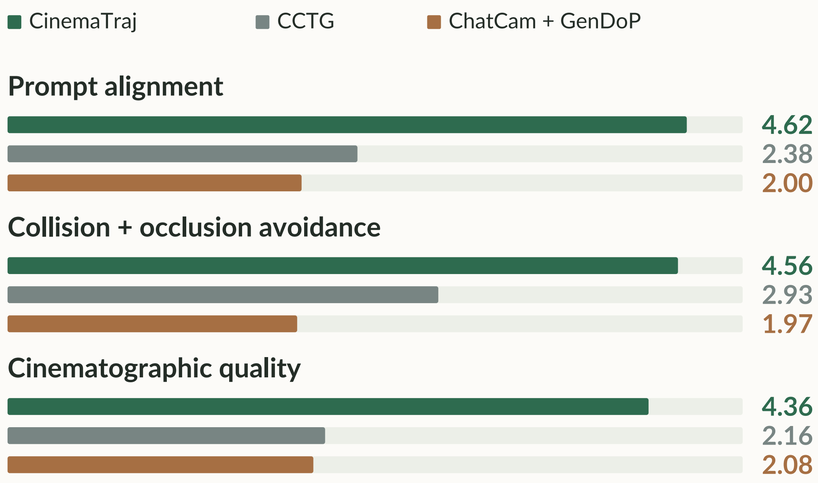
Evaluated on ScanNet++ Scenes

Fully-specified prompts · Comparison and ablation

Method	Motion MSE ↓	CLaTr score ↑	Collision rate ↓	Occlusion rate ↓	Object coverage ↑
ChatCam + GenDoP	7.741	19.461	0.209	0.580	0.661
CCTG	9.143	24.031	0.035	0.516	0.700
w/o Anchor Selector	7.055	28.538	**0.023**	0.579	0.882
w/o Parametric Traj.	2.843	26.243	0.081	**0.460**	**1.000**
w/o SDF-based Opt.	1.841	25.900	0.557	0.566	**1.000**
CinemaTraj (Ours)	**1.741**	**28.982**	0.056	0.503	**1.000**

User Study · Ratings / 5

People prefer CinemaTraj



Key Results: 1.741 Motion MSE · **100% planned-object coverage** · 4.36 / 5 cinematic quality

05 Conclusion

Scene graphs ground intent; parametric shots preserve style; SDF refinement improves spatial feasibility. Together they enable prompt-faithful cinematic walkthroughs with strong motion fidelity and user-rated quality.

Current scope: static scenes and continuous takes. **Next:** dynamic actors and edited multi-shot storytelling.

VIDEOS + CODE

cinematraj.github.io qianru.li@tum.de