MOTIONHARNESS: AGENTIC SPATIOTEMPORAL COORDINATION FOR HUMAN MOTION GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretrained motion models provide expressive control over individual motions, yet this control alone does not ensure spatiotemporal coordination among multiple characters and objects. We propose MotionHarness, an agentic framework that jointly plans behavior, trajectories, and timing without task-specific retraining of the motion backbone. Event specifications make the required spatial and temporal relations explicit and guide verification of execution outcomes. History-guided search explores alternative plans and selectively recombines locally successful decisions from prior executions. To adapt transferred decisions to revised execution contexts, the framework invalidates affected evidence, rebinds transferred targets, and performs dependency-aware motion regeneration. This separates reuse of coordination decisions from reuse of their original trajectories and verification outcomes. We also introduce MotionCoordBench, a benchmark of 300 tasks spanning ten coordination families and three difficulty levels, with offline and online generation settings. Evaluation separately measures task satisfaction, motion validity, and computational cost. Evaluation on offline-generation tasks shows that MotionHarness improves both task success and validity-constrained task success over direct generation across three language-model backbones.

1 INTRODUCTION

Language-model agents coordinate generative processes through planning, tool use, and iterative refinement (Yao et al., 2023b; Shinn et al., 2023; Xu et al., 2025c). In parallel, pretrained motion models provide language-conditioned synthesis and kinematic control (Rempe et al., 2026; Zhao et al., 2025). Together, these capabilities motivate *agentic motion generation* (Figure 1), in which a high-level agent coordinates multiple characters and objects. We study *spatiotemporal coordination*: synthesizing plausible motions that jointly satisfy spatial constraints and temporal dependencies while maintaining consistent character and object states across events. Beyond animation, such sequences could provide motion references for embodied learning (Rempe et al., 2026) or 3D previsualizations for controllable video generation (Zhang et al., 2025b). Existing work advances motion composition, interaction synthesis, and multi-character planning (Petrovich et al., 2024; Li et al., 2024; Zhang et al., 2025a; Chen et al., 2025). Yet coordinating these capabilities across events requires more than composing locally plausible actions: revising one action can alter shared object state and invalidate the prerequisites of later events.

Achieving this coordination presents three interrelated challenges. First, relational goals—such as yielding, synchronized arrival, and object transfer—must be grounded in mutually compatible motion controls. This requires joint reasoning over behavior, geometry, and timing rather than specifying each character independently. Second, these controls do not by themselves establish event completion: a character may reach the prescribed location without clearing a passage or establishing the required contact. Completion must therefore be determined from generated motion rather than inferred from the intended plan. Third, refining that plan requires accounting for coupled event dependencies. Changing the event order, timing, or shared object state can invalidate a previously successful plan component. Refinement must therefore determine which decisions remain applicable and which dependent motions must change, rather than treating each attempt as wholly reusable or wholly disposable.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

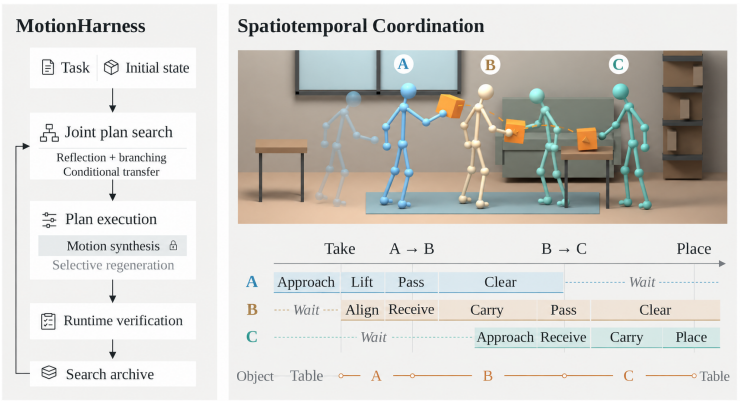

Figure 1: **Spatiotemporal coordination with MotionHarness.** A three-actor handover illustrates overlapping actions and object custody.

Our key insight is that an unsuccessful attempt can still contain useful coordination decisions, but their applicability depends on the execution context. History can therefore supply reusable decisions rather than serve only to diagnose failures. Such reuse must distinguish a decision from the trajectory that realized it and the evidence that supported it: changing shared state or event timing can invalidate both. This distinction allows coordination to be refined independently of the motion model, provided that each revised plan is verified again.

Building on this insight, we propose MotionHarness (Figure 2). An event specification defines the relations that a joint plan of behaviors, trajectories, and timing must satisfy. A compiler maps this plan to model-specific controls, after which a verifier extracts event-level observations and localized violations from the generated sequence. These observations, together with the corresponding plans, form a per-task search archive that supports reflective revision and conditional decision transfer. When a decision is transferred, affected evidence is invalidated, targets are rebound using the revised plan's state, and motion is regenerated according to execution dependencies. The revised sequence is then verified against the unchanged task requirements. To evaluate this coordination process, we also introduce MotionCoordBench, which measures task completion, motion validity, and computational cost separately.

Our contributions are threefold:

- **Agentic spatiotemporal coordination.** We propose MotionHarness, a framework that coordinates actors and objects through joint behavior, trajectory, and timing decisions without task-specific retraining of the motion backbone.
- **History-guided search and plan recombination.** We propose a verification-guided search that branches over alternative plans and conditionally transfers local decisions, with target rebinding, evidence invalidation, and dependency-aware regeneration.
- **Coordination-centric evaluation.** We introduce MotionCoordBench, a benchmark of 300 tasks spanning ten coordination families and three difficulty levels, with offline and online settings and separate measures of task satisfaction, motion validity, and computational cost.

2 Related Work

Controllable motion generation. Motion control requires both semantic alignment and constraints on where and when actions occur. Text-conditioned priors use variational generation (Guo et al., 2022a), discrete motion tokens (Guo et al., 2022b; 2024), or diffusion in motion and latent spaces (Tevet et al., 2023; Zhang et al., 2024b; Chen et al., 2023). Motion–language models unify understanding, generation, and conversational control (Jiang et al., 2023; 2024), while unified editing supports modifying existing motion (Guo et al., 2025). Beyond text, speech semantics provide another control signal: Semantic Gesticulator combines semantic gesture retrieval with rhythm-aware synthesis (Zhang et al., 2024c). To specify where and when motion occurs, explicit controls

range from paths and keyframes (Rempe et al., 2026; Zhao et al., 2025) to programmable constraints (Liu et al., 2024) and body-part timelines (Petrovich et al., 2024; Li et al., 2026a). Long-text scheduling (Qing et al., 2023), continual scene-conditioned motion (Mir et al., 2024), and reusable interaction primitives (Wang et al., 2026) extend these interfaces to longer activities. Building on these controls, MotionHarness addresses a further requirement: coordinating actions whose prerequisites depend on previously generated outcomes.

Interaction-aware motion generation. Interaction models couple actors and objects through joint priors (Liang et al., 2024; Li et al., 2024), object-motion conditioning (Li et al., 2023), or explicit contact and geometric constraints (Diller & Dai, 2024; Xu et al., 2023). Multi-entity capture (Zhang et al., 2024a) provides data for studying these relations. Relational control and composition of pairwise priors extend generation to multiple actors (Wang et al., 2024b; Xu et al., 2025b), while physics-based imitation learns controllers that track motion under simulation dynamics (Peng et al., 2018; Xu et al., 2025a). At the behavioral level, motion–language modeling captures social interactions (Yu et al., 2026); Social Agent derives dyadic nonverbal interaction constraints through conversational agents (Zhang et al., 2025a). High-level instructions can also guide object manipulation (Wu et al., 2025), while activity planning and synchronization organize multi-character sequences (Cao et al., 2024; Chen et al., 2025). Across these settings, changing an interaction may alter shared state and invalidate later actions. MotionHarness focuses on maintaining consistency after such revisions, complementing interaction priors and controllers rather than replacing them.

Agentic planning and verification. Reasoning–action interleaving (Yao et al., 2023b), affordance-grounded skill selection (Ichter et al., 2023), and executable policy programs (Liang et al., 2023) connect language reasoning to action interfaces. Building on such interfaces, skill libraries support reuse across tasks (Wang et al., 2024a), while self-feedback, episodic reflection, and search over alternatives improve candidate solutions (Madaan et al., 2023; Shinn et al., 2023; Yao et al., 2023a). In generative and embodied systems, feedback supports film production (Xu et al., 2025c), streaming social interaction (Zhang et al., 2026), and embodied decision-making (Li et al., 2026b). Motion assessment and refinement (Li et al., 2026c), spatiotemporal predicate checks (Ma & Agrawala, 2025), and geometric feasibility checks (Maillard et al., 2026) provide task-specific diagnoses. Reward-guided policy training (Bu et al., 2026) and reflective prompt evolution (Agrawal et al., 2026) optimize reusable agent behavior; GEPA also studies inference-time search. Within this broader setting, MotionHarness examines conditional reuse within one motion task: transferring a decision requires rechecking its applicability, rebinding its targets, and updating dependent motion and evidence. For evaluation, MotionCoordBench complements relational evaluation of embodied collaboration (Chang et al., 2025) by distinguishing task satisfaction from kinematic validity.

3 METHOD

MotionHarness coordinates multiple actors and objects through joint planning, motion generation, and execution feedback (Figure 2). Its central operation is conditional reuse: a locally effective decision may be transferred between plans, but the motion and verification evidence supporting it must be reconsidered in the revised context.

3.1 PROBLEM FORMULATION

The input task is $G = (g, S_0, \mathcal{R}_{\text{pub}})$, where g is a language instruction, S_0 specifies the initial 3D scene and actor–object states, and $\mathcal{R}_{\text{pub}}$ contains public requirements and tolerances. Given a fixed motion model $\mathcal{M}_\theta$, its available control interface, and a computational budget $\mathbf{B}$, MotionHarness seeks coordinated motion $Y = (X, Q)$ over the task horizon $[0, T]$. Here X contains the actors' root and joint trajectories, while Q contains object poses and, where applicable, articulation coordinates. Object motion is prescribed kinematically rather than obtained from physical simulation.

A *candidate* is a solution to G, represented by an event specification and a joint plan. Executing the plan adds its motion sequence and verification record. Search revises candidates within $\mathbf{B}$, while the motion model and agent's system prompt remain fixed; each task starts with an empty search archive. In continuation settings, a supplied motion prefix must remain unchanged.

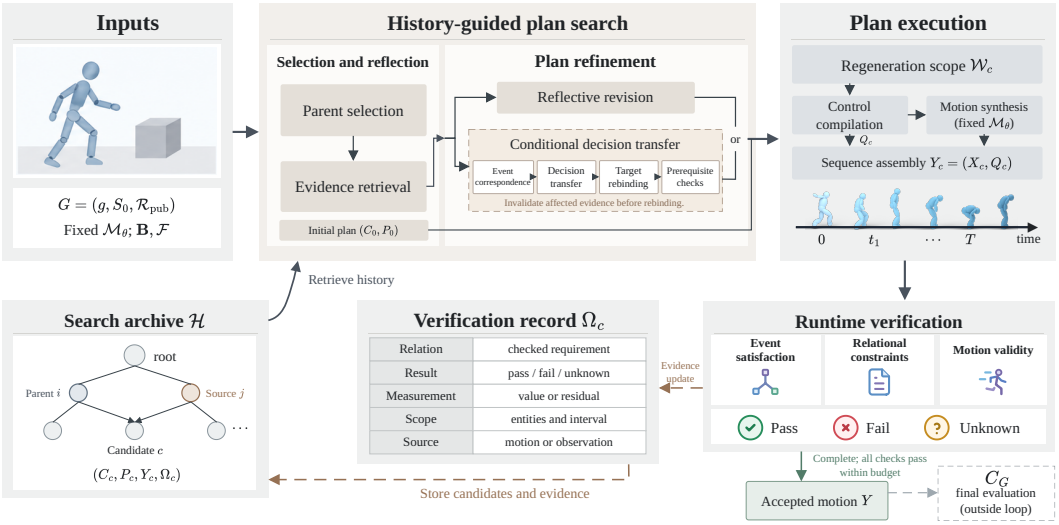

Figure 2: **MotionHarness framework.** Archived execution feedback guides plan revision and conditional decision transfer. Transfer invalidates affected evidence and rebinds targets before dependency-aware regeneration and reverification.

Throughout search, runtime feedback checks the candidate's event specification and public requirements; plan revisions cannot change those requirements. Final evaluation instead uses the fixed task specification C_G, independently of the agent's decomposition. Annotations and results reserved for final evaluation remain unavailable to search. Input variants and formal representations are detailed in Appendices B.1 and C.1.

3.2 Event specifications and execution evidence

The event specification defines *what must happen*; the joint plan describes *how to achieve it*. Each event identifies its participants, prerequisite conditions, conditions that must hold throughout its execution, and completion criteria. Dependencies link event outcomes to subsequent prerequisites; additional constraints express synchronization, overlap, and shared-resource limits without imposing a total order. For a surface-mediated handover, for example, release precedes acquisition, while the releasing actor remains clear of the pickup area during acquisition.

To realize these events, the plan assigns behaviors, spatial targets, event intervals, and a generation order. A reusable decision consists of one or more plan components, such as an action choice, an object-relative approach target, or an event interval. Each component retains explicit entity, coordinate-frame, and event-time references where applicable.

The resulting plan specifies intended outcomes, not observations. Execution evidence is recorded separately: a verification record associates each result with the measured relation, supporting motion, referenced entities, and time interval. The search archive retains these records with the corresponding plan and generated sequence. Revisions create new candidates rather than overwrite their parents, preserving the evidence associated with each attempt. Appendix C.1 defines the candidate representations shown in Figure 2.

3.3 Joint planning and motion synthesis

The initial planner proposes events and a joint plan from G and the available control interface; subsequent calls propose structured edits from a parent plan and execution feedback. In both cases, behavior, geometry, and timing are chosen jointly: B's waiting position must leave A's exit route clear, and B's approach must respect A's verified clearance. This coupling does not require simultaneous generation. Generation order is distinct from event time, so simultaneous actions can be synthesized sequentially. Only calls that consume measured outputs must wait for those observations; other targets remain provisional.

Algorithm 1 History-guided motion coordination

Require: Task G, fixed model and controls, verifier, budget, history limit, and fixed prefix
Ensure: Acceptance status, returned motion Y, search archive, and costs

```
Initialize the archive, expansion queue, and cost record
while another proposal is affordable do
    if the expansion queue is empty then
        Propose an initial event specification and joint plan
    else
        Select a parent; retrieve relevant observations and feedback
        Apply a structured edit or conditional decision transfer (Section 3.5)
    end if
    if inadmissible or duplicate then record costs and diagnostics; continue
    Determine the affected generation calls; preserve the fixed prefix
    Compile controls, generate required motion, and assemble actor and object trajectories
    Verify task and motion requirements; record observations, results, and all costs
    Archive the candidate without modifying its parent; add it to the expansion queue
    if the sequence is complete, every runtime check passes, and the budget is respected then
        return accepted, motion, archive, and costs
    end if
end while
return exhausted, no accepted output, archive, and costs
```

Once targets are resolved, control compilation translates them into model-specific conditioning inputs without making scheduling decisions. Kimodo serves as the fixed pretrained motion generator $\mathcal{M}_\theta$, accepting text, root paths and headings, full-body keyframes, and end-effector targets (Rempe et al., 2026). Within the Kimodo adapter, inverse kinematics fits conditioning poses. Unsupported or inconsistent controls return to the planner; observed execution deviations likewise require plan revision before recompilation.

Actor segments are generated in overlapping windows using retained motion as context (Zhao et al., 2025), then assembled with the plan's object trajectories on a common timeline. Thus the motion model generates X, while object commands define Q; object-relative targets couple the two. Contact and object custody—which actor holds the object and how it is transferred—are checked in the assembled sequence rather than assumed to be satisfied by the commands alone. The indexed model interface, coordinate transforms, and boundary handling appear in Appendix C.2.

3.4 Execution feedback

From the assembled sequence, the verifier measures body occupancy, contact geometry, relative poses, and motion continuity. These measurements support event detection and a search for a consistent occurrence assignment satisfying all prerequisites and temporal constraints. Each check covers the required interval rather than the control activation mask: clearance must hold throughout acquisition, not only at its start.

Each check returns *pass*, *fail*, or *unknown*, together with measured residuals, entities, time intervals, and supporting observations. Insufficient evidence yields unknown rather than success. Local passes are retained even when the full candidate fails, allowing a successful release to inform later revisions of a conflicting approach. Separate kinematic validity checks assess penetration, floor contact, and boundary continuity. Appendix C.3 specifies occurrence matching and observation coverage.

3.5 History-guided branching and recombination

To use local execution evidence in later revisions, the search archive retains successful, failed, and partial attempts. Each expansion selects an archived candidate as the parent to be revised and retrieves relevant observations and feedback under a history limit. Reflection uses this history to propose changes to specified plan components, together with supporting evidence and expected effects. These edits may change behaviors, routes, timing, generation order, or sampling seeds. Retaining alternative parents allows search to explore multiple revisions without discarding useful earlier decisions. Selection and retrieval rules are given in Appendix C.5.

Table 1: **MotionCoordBench task families.** Each family contains ten tasks per difficulty level (30 total), split between offline and online generation.

ID	Task family	Coordination focus	Offline	Online
T01	Event ordering and timing	Event order, timing, and duration	30	0
T02	Shared-space and resource exclusion	Occupancy limits in shared regions	24	6
T03	Relative motion and synchronization	Sustained relative pose and phase	18	12
T04	Whole-body scene interaction	Posture, support, and body clearance	18	12
T05	Interpersonal contact and response	Contact timing and partner response	21	9
T06	Object manipulation and custody	Contact and custody during transport	30	0
T07	Handover and relay	Identity and release–acquire order	30	0
T08	Simultaneous co-manipulation	Concurrent contact on a shared object	30	0
T09	Articulated objects and tool use	Joint limits and tool-mediated state changes	30	0
T10	Object-state reconfiguration	Identity across support and binding changes	30	0
Total			**261**	**39**

Conditional decision transfer. Beyond reflective edits, search may reuse a decision from a *source candidate*: another archived solution to the same task with a locally satisfied requirement. The agent matches corresponding events and selects the behaviors, spatial targets, or event intervals to reuse. *Decision transfer* replaces these components in the parent plan with the source choices, retaining all unselected components. It transfers planning choices rather than the source's generated motion or verification results. Deterministic checks enforce participant roles, unique correspondence, and dependency compatibility. Applying the edit invalidates affected evidence before spatial targets are rebound to the revised plan. For an object-relative target $\bar{p}$, the world-space target in the new candidate c is

$$p_c = R_{c,o}(t_c)\bar{p} + q_{c,o}(t_c), \tag{1}$$

where $R_{c,o}(t_c)$ and $q_{c,o}(t_c)$ are object o's rotation and position at the corresponding event time t_c. World-frame targets and unedited event times remain unchanged. Target rebinding alone does not establish event prerequisites: these are checked using only unaffected observations, and unresolved conditions require new evidence. Thus, B's approach can follow a newly placed object, but B must still wait until the revised pickup area is clear.

Dependency-aware regeneration. Both reflective edits and decision transfer invalidate dependent evidence and may affect additional motion windows. For a plan edit Δ, let $\mathcal{W}_\Delta$ contain ungenerated calls and calls whose controls, seeds, boundary motion, or observed inputs have changed. An execution-dependency graph $\mathcal{G}_{\text{exec}}$ links calls that consume one another's outputs; these edges describe computational dependencies rather than event prerequisites. Regeneration therefore covers the initially affected calls and their descendants:

$$\mathcal{W}_c = \mathcal{W}_\Delta \cup \mathrm{Desc}_{\mathcal{G}_{\text{exec}}}(\mathcal{W}_\Delta), \tag{2}$$

where Desc includes all downstream calls that depend on the changed outputs. Regenerated motion replaces affected intervals while committed prefixes remain fixed. Reverification also covers retained boundaries and changed interactions, including conflicts not represented by existing dependency edges. Reusing a decision therefore does not require reusing its original motion.

Algorithm 1 integrates these operations into the search loop. All attempts, including rejected proposals, count toward the budget. Search returns the first complete candidate that passes every runtime check within budget; if none is found, it reports exhaustion with no accepted output and retains the archive. This runtime acceptance remains distinct from final success under C_G. Detailed transfer and stopping rules appear in Appendix C.5.

4 MotionCoordBench

MotionCoordBench evaluates spatial, temporal, and actor–object coordination through relation-constrained tasks (Chang et al., 2025). The benchmark comprises 300 tasks, with ten instances for each combination of ten task families and three difficulty levels. Each task pairs a public goal and a 3D scene specification with a fixed evaluation specification C_G.

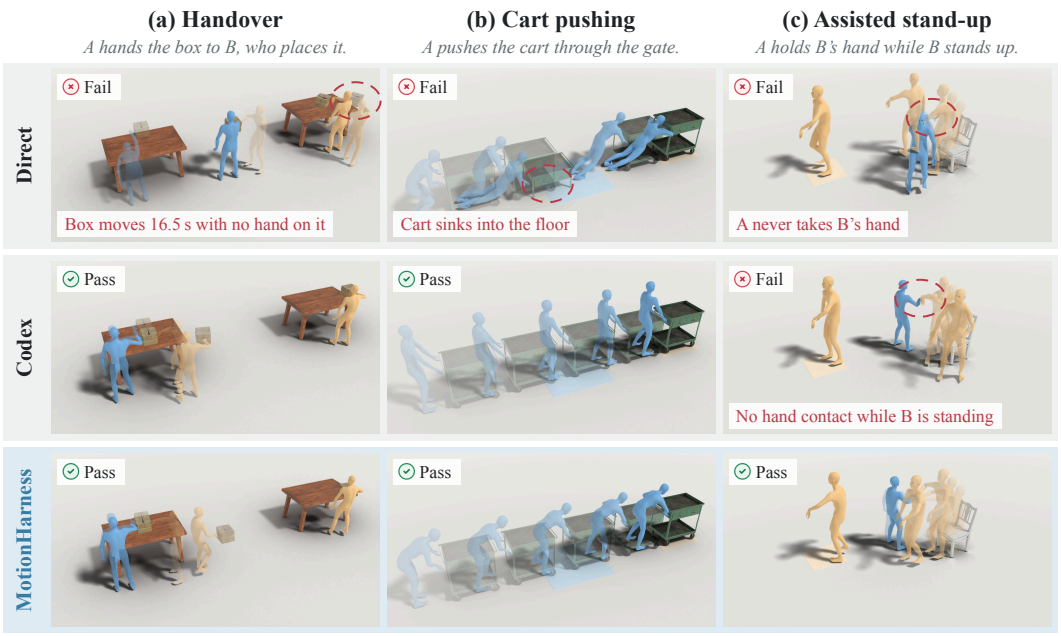

Figure 3: **Qualitative comparison with GPT-6 Sol.** Rows compare Direct, Codex, and MotionHarness; columns show three task types. A pass satisfies both task-completion and kinematic-validity checks.

4.1 Task families and difficulty

Table 1 summarizes tasks spanning temporal organization, full-body interaction, and persistent multi-entity state. These families combine overlapping relation types rather than partitioning them. Within each family, the three levels (L1–L3) vary temporal tolerances, spatial clearance, contact precision, and dependency complexity. Difficulty is thus specified by task parameters rather than model performance. The three versions of each source scenario remain grouped, yielding 100 groups.

4.2 Evaluation settings and inputs

We distinguish two evaluation settings. In *offline generation* (261 tasks), the method returns a complete motion sequence and may revise it before submission. In *online generation* (39 tasks), the method receives observations and commits the next motion segment every 0.1 s of simulation time; committed motion cannot be revised, and future external signals and actor motion are not provided. Reported experiments use offline generation.

Within these settings, the goal-conditioned input includes instructions, requirement text, scene geometry, public tolerances, and control access. The event-graph-conditioned diagnostic additionally provides the task event graph.

4.3 Metrics

TaskSR requires all mandatory events, temporal dependencies, interval relations, and final-state conditions to pass under a consistent event assignment. ValidTaskSR additionally requires all kinematic validity checks to pass, covering bone lengths, speed, penetration, floor contact, and foot sliding. Missing or unresolved mandatory evidence does not count as success. The main model comparison and ablation study use all 261 offline-generation tasks. Table S2 further reports valid completions by task family and difficulty for the model comparison. To account for the resources used to obtain these outcomes, costs include planning, failed proposals, and every execution. Appendix B.3 specifies scoring, and Appendix B.4 defines budget accounting.

Table 2: **Model comparison on 261 offline-generation tasks.** Methods share an executor. Costs include all attempts and checker calls; runtime is end-to-end. Means use all tasks, except Codex token means, which require a completed agent turn.

Method	Base LLM	TaskSR (%) ↑	ValidTaskSR (%) ↑	Tokens (k) ↓	Generated motion (s) ↓	Runtime (s) ↓
Direct	GPT-6 Astra	39.8	29.9	**7.6**	146	**89**
Direct	GPT-6 Sol	32.6	20.3	9.2	146	121
Direct	GPT-5.6 Sol	33.7	21.5	9.1	146	140
Direct	Claude Opus 5	40.2	21.1	34.9	146	436
Direct	Gemini 3.1 Pro	37.9	24.9	29.0	146	384
Direct	Qwen3-Max	18.0	7.7	23.0	146	423
Codex	GPT-6 Sol	38.3	32.6	681	763	506
Pi	Qwen3-Max	14.2	11.5	372	745	211
MotionHarness	GPT-6 Astra	**51.3**	**42.1**	90.8	858	612
MotionHarness	GPT-6 Sol	42.1	30.7	113	812	815
MotionHarness	Qwen3-Max	25.7	21.1	234	954	3083

4.4 Scope and validation

The benchmark has no held-out evaluation split; related scenarios remain grouped. Structural validation checks references, body geometry, prerequisites, and timing; synthetic tests assess verifier behavior. Benchmark construction and evaluation details are provided in Appendix B.

5 Experiments

5.1 Evaluation setup

For quantitative evaluation, candidate plans are executed and verified using a 60 Hz kinematic executor. Kimodo synthesizes motion conditioned on selected rollout keypoints; the reported metrics and costs concern executor-based plan evaluation.

Table 2 is uncapped; Table 3 uses difficulty-dependent budgets. All methods receive public task inputs. Model settings, feedback access, and cost accounting appear in Appendix B.4; metrics follow Section 4.3.

5.2 Overall coordination performance

MotionHarness improves both TaskSR and ValidTaskSR over direct generation across all three evaluated backbones (Table 2). With GPT-6 Astra, it achieves the highest reported TaskSR (51.3%) and ValidTaskSR (42.1%), compared with 39.8% and 29.9% for direct generation.

With GPT-6 Sol, MotionHarness achieves higher TaskSR than Codex (42.1% versus 38.3%) but lower ValidTaskSR (30.7% versus 32.6%), illustrating that task satisfaction and motion validity can diverge. With Qwen3-Max, MotionHarness reaches 21.1% ValidTaskSR, versus 7.7% for direct generation and 11.5% for Pi; only Pi disables thinking (Appendix B.4). Figure 3 illustrates spatial coordination and event timing across three task types for the compared methods.

5.3 Ablation study

At low reasoning effort, MotionHarness reaches 49.0% TaskSR and 38.3% ValidTaskSR (Table 3). Removing joint planning or iterative refinement reduces ValidTaskSR by 14.9 and 7.3 percentage points. Single-lineage refinement and branching without recombination reach 35.2% and 36.4%; full regeneration yields 48.3% TaskSR and 34.5% ValidTaskSR (Appendix B.5).

Table 3: **Ablations with GPT-6 Astra (low reasoning effort).** All variants use 261 offline-generation tasks, a shared executor, and common budget limits. Costs are per-task means over all attempts, including initial planning.

Variant	TaskSR (%) ↑	ValidTaskSR (%) ↑	Generated motion (s) ↓	Runtime (s) ↓
MotionHarness	**49.0**	**38.3**	284	245
w/o joint planning	34.9	23.4	141	**115**
w/o verification-guided refinement	46.7	31.0	**139**	119
w/o plan recombination	46.7	36.4	276	224
Single-lineage refinement	46.4	35.2	281	231
Full regeneration	48.3	34.5	222	198

5.4 Performance across task families

Across 261 offline-generation tasks (Table S2), direct baselines achieve motion-valid completions only in T01–T03 and T05. MotionHarness with GPT-6 Astra also records valid completions in T04, T06, T07, T09, and T10, totaling 110. No evaluated configuration achieves a motion-valid completion in co-manipulation (T08).

5.5 Computational cost

Under bounded budgets, MotionHarness uses 284 s of generated motion and 245 s of runtime, costing 8 s and 21 s more than branching without recombination for a 1.9-point ValidTaskSR gain. Full regeneration is cheaper but yields lower ValidTaskSR. Uncapped Astra-based MotionHarness uses 90.8k tokens and 612 s, versus 7.6k and 89 s for direct generation. These comparisons show a success–cost tradeoff; uncapped methods are not resource-matched, and aggregate costs do not isolate per-edit regeneration efficiency.

6 Limitations

MotionHarness inherits the control limitations of its motion backbone, particularly for sustained contact. Although execution feedback guides refinement, kinematic checks alone cannot establish physical stability or perceived motion naturalness. Coupled motion dependencies can also require extensive regeneration after local edits, restricting both the number of candidate plans explored and the depth of refinement within a fixed computational budget.

7 Conclusion

We presented MotionHarness, an agentic framework for spatiotemporal coordination without task-specific backbone retraining. Event specifications link joint behavior, trajectory, and timing decisions to execution feedback, enabling history-guided search and conditional reuse of locally successful decisions. Target rebinding, evidence invalidation, and dependency-aware regeneration adapt these decisions to revised contexts, separating their reuse from that of the original trajectories and verification outcomes. MotionCoordBench complements the framework with 300 tasks across ten families, supporting joint assessment of task completion, motion validity, and computational cost. Evaluation shows higher task success and motion-valid completion than direct generation across three language-model backbones. These findings highlight the value of planning and execution feedback for multi-character and human–object coordination.

AI USE STATEMENT

Generative AI tools assisted with the literature review, experimental design, synthetic benchmark construction, implementation and debugging, interpretation and presentation of results, and refinement of the manuscript language. The authors reviewed the AI-assisted work and take responsibility for the final manuscript, scientific claims, citations, code, benchmark, and reported results.

ETHICS STATEMENT

This work studies synthetic character motion in virtual scenes and involves no new human-participant studies or collection of personal data. Third-party motion data, models, and scene assets are subject to their respective licenses and attribution requirements. Kinematic evaluation does not establish the safety of physical deployment.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix C describe the method and algorithms. Appendix B details benchmark construction, evaluation criteria, experimental controls, execution backends, system configurations, and budget accounting. Success rates include all allocated tasks, and computational costs include unsuccessful attempts. Code accompanies the supplementary material. Our code and benchmark will be publicly released upon acceptance.

REFERENCES

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. GEPA: Reflective prompt evolution can outperform reinforcement learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0e9e708b6f48e14fd0ac29e167413f76-Abstract-Conference.html.

Bones Studio. BONES-SEED: Skeletal everyday embodiment dataset. Hugging Face dataset, 2026. URL https://huggingface.co/datasets/bones-studio/seed.

Samarth Brahmbhatt, Cusuh Ham, Charles C. Kemp, and James Hays. ContactDB: Analyzing and predicting grasp contact via thermal imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8709–8719, 2019. doi: 10.1109/CVPR.2019.00891.

Xingyuan Bu, Chengru Song, Hao Zhou, Tao Zhou, Dong Li, Wei Li, Shilong Li, Hao Shi, Yongxin Guo, Donghao Zhou, Qiangpeng Yang, and Shilei Wen. AgenticGen: Reward-guided agentic video generation for advertising. *arXiv preprint arXiv:2609.09187*, 2026. URL https://arxiv.org/abs/2609.09187.

Yukang Cao, Xinying Guo, Mingyuan Zhang, Haozhe Xie, Chenyang Gu, and Ziwei Liu. CrowdMoGen: Zero-shot text-driven collective motion generation. *arXiv preprint arXiv:2407.06188*, 2024. URL https://arxiv.org/abs/2407.06188.

Matthew Chang, Gunjan Chhablani, Alexander Clegg, Mikael Dallaire Cote, Ruta Desai, Michal Hlavac, Vladimir Karashchuk, Jacob Krantz, Roozbeh Mottaghi, Priyam Parashar, Siddharth Patki, Ishita Prasad, Xavier Puig, Akshara Rai, Ram Ramrakhya, Daniel Tran, Joanne Truong, John M. Turner, Eric Undersander, and Tsung-Yen Yang. PARTNR: A benchmark for planning and reasoning in embodied multi-agent tasks. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a3cf318fbeec1126da21e9185ae9908c-Abstract-Conference.html.

Jianqi Chen, Panwen Hu, Xiaojun Chang, Zhenwei Shi, Michael Kampffmeyer, and Xiaodan Liang. Sitcom-crafter: A plot-driven human motion generation system in 3D scenes. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/de3d2bb604cfc43c81edd2a31b257f03-Abstract-Conference.html.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18000–18010, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Chen_Executing_Your_Commands_via_Motion_Diffusion_in_Latent_Space_CVPR_2023_paper.html.

Christian Diller and Angela Dai. CG-HOI: Contact-guided 3D Human-Object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19888–19901, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Diller_CG-HOI_Contact-Guided_3D_Human-Object_Interaction_Generation_CVPR_2024_paper.html.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3D human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5152–5161, 2022a. URL https://openaccess.thecvf.com/content/CVPR2022/html/Guo_Generating_Diverse_and_Natural_3D_Human_Motions_From_Text_CVPR_2022_paper.html.

Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. TM2T: Stochastic and tokenized modeling for the reciprocal generation of 3D human motions and texts. In *European Conference on Computer Vision*, pp. 580–597, 2022b. doi: 10.1007/978-3-031-19833-5_34.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. MoMask: Generative masked modeling of 3D human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1900–1910, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Guo_MoMask_Generative_Masked_Modeling_of_3D_Human_Motions_CVPR_2024_paper.html.

Ziyan Guo, Zeyu Hu, De Wen Soh, and Na Zhao. MotionLab: Unified human motion generation and editing via the motion-condition-motion paradigm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13869–13879, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Guo_MotionLab_Unified_Human_Motion_Generation_and_Editing_via_the_Motion-Condition-Motion_ICCV_2025_paper.html.

Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, Dmitry Kalashnikov, Sergey Levine, Yao Lu, Carolina Parada, Kanishka Rao, Pierre Sermanet, Alexander T. Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Mengyuan Yan, Noah Brown, Michael Ahn, Omar Cortes, Nicolas Sievers, Clayton Tan, Sichun Xu, Diego Reyes, Jarek Rettinghouse, Jornell Quiambao, Peter Pastor, Linda Luu, Kuang-Huei Lee, Yuheng Kuang, Sally Jesmonth, Nikhil J. Joshi, Kyle Jeffrey, Rosario Jauregui Ruano, Jasmine Hsu, Keerthana Gopalakrishnan, Byron David, Andy Zeng, and Chuyuan Kelly Fu. Do as I can, not as I say: Grounding language in robotic affordances. In *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 287–318. PMLR, 2023. URL https://proceedings.mlr.press/v205/ichter23a.html.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. MotionGPT: Human motion as a foreign language. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/3fbf0c1ea0716c03dea93bb6be78dd6f-Abstract.html.

Biao Jiang, Xin Chen, Chi Zhang, Fukun Yin, Zhuoyuan Li, Gang Yu, and Jiayuan Fan. MotionChain: Conversational motion controllers via multimodal prompts. In *European Conference on Computer Vision*, pp. 54–74, 2024. doi: 10.1007/978-3-031-73347-5_4.

Chuqiao Li, Xianghui Xie, Yong Cao, Andreas Geiger, and Gerard Pons-Moll. FrankenMotion: Part-level human motion generation and composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a. URL https://arxiv.org/abs/2601.10909v1.

Jiaman Li, Jiajun Wu, and C. Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics*, 2023. doi: 10.1145/3618333.

Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C. Karen Liu. Controllable human-object interaction synthesis. In *European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/5720_ECCV_2024_paper.php.

Siyao Li, Jiawei Gu, Shuai Liu, Kairui Hu, Zekun Li, Linjie Li, Chengcheng Tang, Po-Chen Wu, Ivan Shugurov, Lingni Ma, Michael Zollhoefer, Sizhe An, Abhay Mittal, Amy Zhao, Ranjay Krishna, Manling Li, Ziwei Liu, and Chuan Guo. HumanCLAW: Can vision-language models act through a body? *arXiv preprint arXiv:2607.27180*, 2026b. URL https://arxiv.org/abs/2607.27180.

Yuan-Ming Li, Qize Yang, Nan Lei, Shenghao Fu, Ling-An Zeng, Jian-Fang Hu, Xihan Wei, and Wei-Shi Zheng. IRG-MotionLLM: Interleaving motion generation, assessment and refinement for text-to-motion generation. In *European Conference on Computer Vision*, 2026c. URL https://arxiv.org/abs/2512.10730.

Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. InterGen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, 132:3463–3483, 2024. doi: 10.1007/s11263-024-02042-6.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *IEEE International Conference on Robotics and Automation*, pp. 9493–9500, 2023. doi: 10.1109/ICRA48891.2023.10160591.

Hanchao Liu, Xiaohang Zhan, Shaoli Huang, Tai-Jiang Mu, and Ying Shan. Programmable motion generation for open-set motion control tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1399–1408, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_Programmable_Motion_Generation_for_Open-Set_Motion_Control_Tasks_CVPR_2024_paper.html.

Jiaju Ma and Maneesh Agrawala. MoVer: Motion verification for motion graphics animations. *ACM Transactions on Graphics*, 2025. doi: 10.1145/3731209.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Léopold Maillard, Francis Engelmann, Tom Durand, Boxiao Pan, Yang You, Leonidas Guibas, and Maks Ovsjanikov. SceneTeract: Probing and improving agent-aware activity reasoning in 3D indoor scenes. *arXiv preprint arXiv:2603.29798v2*, 2026. URL https://arxiv.org/abs/2603.29798v2.

Aymen Mir, Xavier Puig, Angjoo Kanazawa, and Gerard Pons-Moll. Generating continual human motion in diverse 3D scenes. In *International Conference on 3D Vision*, pp. 903–913, 2024. doi: 10.1109/3DV62453.2024.00061.

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. DeepMimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions on Graphics*, 37(4):143:1–143:14, 2018. doi: 10.1145/3197517.3201311.

Mathis Petrovich, Or Litany, Umar Iqbal, Michael J. Black, Gül Varol, Xue Bin Peng, and Davis Rempe. Multi-track timeline control for text-driven 3D human motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1911–1921, 2024. URL https://openaccess.thecvf.com/content/CVPR2024W/HuMoGen/html/Petrovich_Multi-Track_Timeline_Control_for_Text-Driven_3D_Human_Motion_Generation_CVPRW_2024_paper.html.

Zhongfei Qing, Zhongang Cai, Zhitao Yang, and Lei Yang. Story-to-motion: Synthesizing infinite and controllable character animation from long text. In *SIGGRAPH Asia 2023 Technical Communications*, 2023. doi: 10.1145/3610543.3626176.

Davis Rempe, Mathis Petrovich, Ye Yuan, Haotian Zhang, Xue Bin Peng, Yifeng Jiang, Tingwu Wang, Umar Iqbal, David Minor, Michael de Ruyter, Jiefeng Li, Chen Tessler, Edy Lim, Eugene Jeong, Sam Wu, Ehsan Hassani, Michael Huang, Jin-Bey Yu, Chaeyeon Chung, Lina Song, Olivier Dionne, Jan Kautz, Simon Yuen, and Sanja Fidler. Kimodo: Scaling controllable human motion generation. *arXiv preprint arXiv:2603.15546*, 2026. URL https://arxiv.org/abs/2603.15546.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision*, pp. 581–600, 2020. doi: 10.1007/978-3-030-58548-8_34.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=SJ1kSyO2jwu.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024a. URL https://openreview.net/forum?id=ehfRiF0R3a.

Tingwu Wang, Olivier Dionne, Michael De Ruyter, David Minor, Davis Rempe, Kaifeng Zhao, Mathis Petrovich, Ye Yuan, Chenran Li, Zhengyi Luo, Brian Robison, Xavier Blackwell, Bernardo Antoniazzi, Xue Bin Peng, Yuke Zhu, and Simon Yuen. Motionbricks: Scalable real-time motions with modular latent generative model and smart primitives. *ACM Transactions on Graphics*, 45(4):85:1–85:22, 2026. doi: 10.1145/3811334.

Zhenzhi Wang, Jingbo Wang, Yixuan Li, Dahua Lin, and Bo Dai. InterControl: Zero-shot human interaction generation by controlling every joint. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/be41269a9fe258f1ecba663b0b402322-Abstract-Conference.html.

Zhen Wu, Jiaman Li, Pei Xu, and C. Karen Liu. Human-object interaction from Human-Level instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11176–11186, 2025. doi: 10.1109/ICCV51701.2025.01040.

Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. InterDiff: Generating 3D Human-Object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14928–14940, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Xu_InterDiff_Generating_3D_Human-Object_Interactions_with_Physics-Informed_Diffusion_ICCV_2023_paper.html.

Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. InterMimic: Towards universal whole-body control for physics-based Human-Object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12266–12277, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Xu_InterMimic_Towards_Universal_Whole-Body_Control_for_Physics-Based_Human-Object_Interactions_CVPR_2025_paper.html.

Wenning Xu, Shiyu Fan, Paul Henderson, and Edmond S. L. Ho. Multi-person interaction generation from two-person motion priors. In *ACM SIGGRAPH 2025 Conference Papers*, 2025b. doi: 10.1145/3721238.3730688.

Zhenran Xu, Longyue Wang, Jifang Wang, Zhouyi Li, Senbao Shi, Xue Yang, Yiyu Wang, Baotian Hu, Jun Yu, and Min Zhang. FilmAgent: A multi-agent framework for end-to-end film automation in virtual 3D spaces. *arXiv preprint arXiv:2501.12909*, 2025c. URL https://arxiv.org/abs/2501.12909.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 11809–11822, 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=WE_vluYUL-X.

Heng Yu, Juze Zhang, Changan Chen, Tiange Xiang, Yusu Fang, Juan Carlos Niebles, and Ehsan Adeli. SocialGen: Modeling multi-human social interaction with language models. In *International Conference on 3D Vision*, pp. 1844–1860, 2026. doi: 10.1109/3DV69130.2026.00175.

Yanjie Ze, João Pedro Araújo, Jiajun Wu, and C. Karen Liu. GMR: General motion retargeting, 2025. URL https://github.com/YanjieZe/GMR. GitHub repository.

Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingya Wang, Lan Xu, and Jingyi Yu. HOI-M3: Capture multiple humans and objects interaction within contextual environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 516–526, 2024a. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zhang_HOI-M3_Capture_Multiple_Humans_and_Objects_Interaction_within_Contextual_Environment_CVPR_2024_paper.html.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. MotionDiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4115–4128, 2024b. doi: 10.1109/TPAMI.2024.3355414.

Zeyi Zhang, Tenglong Ao, Yuyao Zhang, Qingzhe Gao, Chuan Lin, Baoquan Chen, and Libin Liu. Semantic Gesticulator: Semantics-aware co-speech gesture synthesis. *ACM Transactions on Graphics*, 43(4):1–17, 2024c. doi: 10.1145/3658134.

Zeyi Zhang, Yanju Zhou, Heyuan Yao, Tenglong Ao, Xiaohang Zhan, and Libin Liu. Social Agent: Mastering dyadic nonverbal behavior generation via conversational LLM agents. In *SIGGRAPH Asia 2025 Conference Papers*, 2025a. doi: 10.1145/3757377.3763879.

Zeyi Zhang, Zixi Kang, Ruijie Zhao, Yusen Feng, Biao Jiang, Hanyu Ji, and Libin Liu. ProAct: Harnessing streaming motion generation and agentic reasoning for real-time embodied social interaction. *ACM Transactions on Graphics*, 2026. doi: 10.1145/3842576.

Zhiyuan Zhang, Dongdong Chen, and Jing Liao. I2V3D: Controllable image-to-video generation with 3D guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13360–13371, 2025b. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_I2V3D_Controllable_Image-to-video_Generation_with_3D_Guidance_ICCV_2025_paper.html.

Kaifeng Zhao, Gen Li, and Siyu Tang. DartControl: A diffusion-based autoregressive motion model for real-time text-driven motion control. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/3b057de5a2e38bd8fa10201866c20dbf-Abstract-Conference.html.