WORLD2MOTION: TURNING VIDEO WORLD MODELS INTO 3D HUMAN MOTION GENERATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

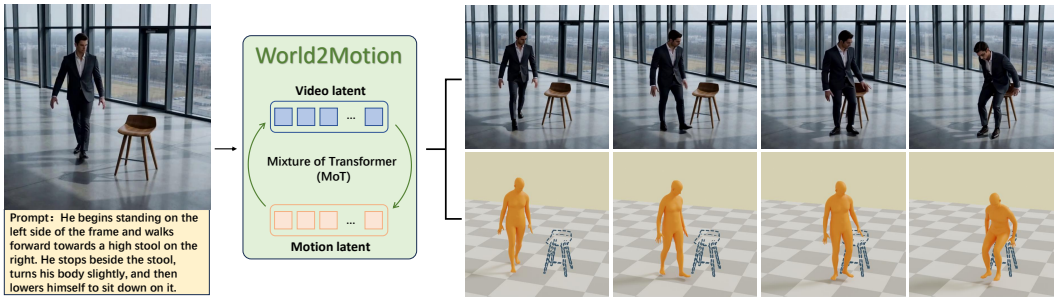

Figure 1: World2Motion adapts a video world model to jointly generate scene-aware 3D human motion and corresponding video from a single image and a text prompt. The example shows a person approaching a stool, turning, and sitting down. The top row shows generated video frames, while the bottom row shows the corresponding 3D motion rendered in a scene for visualization.

We present World2Motion, a framework that generates scene-aware 3D human motion and corresponding video from a single image and a text prompt. While existing 3D motion generators learn from motion datasets, their generalization is constrained by limited coverage of environments. In contrast, video world models such as Cosmos 3 offer broader environmental priors but are not designed for full-body motion generation; recovering motion from their generated videos requires costly two-stage inference. To address these, we turn Cosmos 3 into a single-stage 3D motion generator. This adaptation has two challenges: the scarcity of paired video–motion data and temporal instability in the generated motion. First, we construct a training dataset combining synthetic video–motion pairs with real videos paired with estimated 3D motion. Second, we propose a shift-decoupled noise schedule that assigns different noise levels to video and motion through shared denoising progress. This design accommodates the different denoising requirements of the two modalities, reducing motion jitter. Experiments on a multi-source interaction benchmark show that World2Motion has better motion–text alignment and scene interaction compared with the evaluated 3D motion generators. It also matches the interaction success rate of the two-stage baseline while achieving approximately 3.3× faster inference.

1 INTRODUCTION

Generating 3D human motion for interactive applications requires motions that follow text descriptions and remain consistent with the surrounding environment. Conventional text-to-motion (T2M) methods (Zhang et al., 2024; Tevet et al., 2023) synthesize human motion from textual descriptions, and recent approaches (Team, 2025; Rempe et al., 2026) have improved generation quality by scaling up motion datasets. However, the limited coverage of environments and interactions in available 3D motion datasets constrains their generalization to diverse scenes. Text alone also makes it difficult to specify detailed spatial relationships. For example, generating a person sitting on a chair requires the motion to match the chair's position and orientation. Human–object interaction (HOI)

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.

methods (Li et al., 2024; Jung et al., 2026) incorporate such spatial information through explicit 3D inputs, such as object meshes (Kulkarni et al., 2024) or object states (Li et al., 2024; Zhang et al., 2026). However, collecting the corresponding motion and scene data at scale remains difficult.

Video world models offer an alternative source of knowledge about human motion and environmental interactions (Wan et al., 2025; MiniMax, 2026). Trained on diverse videos, models such as Cosmos 3 (Agarwal et al., 2026) provide environmental priors that can complement the limited coverage of 3D motion datasets. However, they are not designed to directly produce the full-body 3D motion required by our task. A straightforward way to use these priors is to first generate a video and then recover motion using a human mesh recovery (HMR) model (Patel & Black, 2025; Zhao et al., 2026c). This two-stage pipeline enables scene-aware motion generation, but requires a separate motion-recovery pass over the generated video, increasing inference cost.

We introduce World2Motion, which adapts the existing Mixture-of-Transformers (MoT) backbone of Cosmos 3 (Agarwal et al., 2026) to jointly generate scene-aware 3D human motion and corresponding video from a single image and a text prompt. This single-stage framework avoids a separate motion-recovery stage on the generated sequence. The adaptation presents two challenges: the scarcity of paired video–motion training data and temporal instability in the generated motion. Table 1 summarizes the input requirements and generation pipelines of these approaches.

Table 1: Motion generation methods comparison.

Approach	Direct 3D motion output	No 3D object input	Scene conditioning
T2M	✓	✓	
I2V + HMR		✓	✓
HOI	✓		✓
World2Motion	✓	✓	✓

To address the scarcity of paired supervision, we construct a hybrid training dataset combining synthetic video–motion pairs with real videos paired with estimated 3D motion. For the synthetic component, we use motion sequences from existing interaction datasets (Li et al., 2023; Hassan et al., 2021a; Jiang et al., 2023b) to construct corresponding videos, providing paired supervision with known 3D motion. For the real component, we estimate 3D motion from real videos (Patel & Black, 2025) to expand the coverage of environments and appearances. These two sources combine motion supervision from existing 3D datasets with the visual diversity of real videos, supporting the adaptation of the video world model to joint video–motion generation.

Directly fine-tuning the backbone for joint video–motion generation produces noticeable motion jitter. To improve temporal stability, we employ root translations relative to the initial root position and 6D joint rotations (Zhao et al., 2026b). We propose a shift-decoupled noise schedule that assigns different noise levels to video and motion through shared denoising progress, accounting for their different denoising requirements. We maintain the same video–motion noise relationship during training and inference. Our ablation studies show that the shift-decoupled noise schedule reduces motion jitter.

On a multi-source interaction benchmark, World2Motion improves motion–text alignment and scene interaction over the evaluated 3D motion generators. It also achieves comparable interaction success to the MiniMax-H3 + CameraHMR baseline with approximately 3.3× faster inference.

Our contributions are as follows:

- We adapt Cosmos 3 for single-stage joint generation of scene-aware 3D human motion and corresponding video from a single image and a text prompt, avoiding a separate motion-recovery stage on the generated video.
- We construct a hybrid training dataset combining synthetic video–motion pairs with real videos paired with estimated 3D motion.
- We propose a shift-decoupled noise schedule that uses modality-specific noise shifts with shared denoising progress and maintains the same video–motion noise relationship during training and inference, improving temporal stability.

2 Related Work

Text-conditioned human motion generation. Text-to-motion (T2M) methods generate 3D human motion from natural language. Diffusion-based methods such as MotionDiffuse (Zhang et al., 2024) and MDM (Tevet et al., 2023) denoise motion sequences, while MLD (Chen et al., 2023) performs diffusion in a learned motion latent space and ReMoDiffuse (Zhang et al., 2023b) incorporates retrieved motion examples. Autoregressive methods, including T2M-GPT (Zhang et al., 2023a) and MotionGPT (Jiang et al., 2023a), generate discrete motion tokens sequentially. Masked models such as MoMask (Guo et al., 2024), BAMM (Pinyoanuntapong et al., 2024), and MaskControl (Pinyoanuntapong et al., 2025) further explore token prediction for motion generation and control. MoMask uses residual vector quantization to represent motion with multiple token layers, while MaskControl incorporates spatial constraints. Recent work has also studied larger training datasets and more flexible motion constraints. HY-Motion (Team, 2025) scales DiT-based flow matching to a billion parameters, while Kimodo (Rempe et al., 2026) trains a motion diffusion model on large-scale optical motion capture data and supports pose and trajectory constraints. These methods improve motion generation and control, but do not directly use scene images to model spatial relationships between a person and the surrounding environment.

Scene-conditioned human motion generation. Scene-conditioned methods incorporate environmental information into human pose and motion generation. Methods such as POSA (Hassan et al., 2021b), SAMP (Hassan et al., 2021a), COINS (Zhao et al., 2022), and HUMANISE (Wang et al., 2022) use 3D scene information to model human placement, contact, or motion. Other approaches explore different forms of interaction conditions: OMOMO (Li et al., 2023) uses object trajectories, Move as You Say (Wang et al., 2024) uses scene affordances, and TRUMANS (Jiang et al., 2024) conditions motion generation on scene context and actions. SceneDiffuser (Huang et al., 2023) and InfBaGel (Zou et al., 2026) further study scene-conditioned generation and refinement for human interactions. These methods explore how scene information can guide motion beyond textual descriptions. Move-in-2D (Huang et al., 2025) generates motion from a background image and text, and uses the generated motion to guide subsequent video generation. Our work studies the adaptation of a pretrained video world model for image-conditioned joint video and 3D motion generation.

Human motion recovery from images and videos. Human motion recovery estimates 3D body pose and shape from observed images or videos, often using a parametric body model such as SMPL (Loper et al., 2015). HMR (Kanazawa et al., 2018), SPIN (Kolotouros et al., 2019), CLIFF (Li et al., 2022), and HybrIK (Li et al., 2021) improve recovery through learned pose priors, optimization during training, image-location cues, and kinematic constraints. More recent methods study transformer-based recovery, whole-body and multi-person reconstruction, and camera and temporal modeling, including Humans in 4D (Goel et al., 2023), SMPLer-X (Cai et al., 2023), Multi-HMR (Baradel et al., 2024), CameraHMR (Patel & Black, 2025), MetricHMSR (Song et al., 2026), and OnlineHMR (Zhao et al., 2026c). In our framework, CameraHMR provides motion estimates for real training videos and the initial pose from the input image. Motion recovery can also be applied to generated videos, forming a two-stage motion-generation pipeline, but the additional recovery stage increases inference cost. In contrast, World2Motion jointly generates motion and video, achieving approximately $3.3\times$ faster inference, better motion–text alignment, and lower motion jitter than the MiniMax-H3 (MiniMax, 2026) + CameraHMR (Patel & Black, 2025) baseline.

Video world models and joint video–motion generation. Large-scale video models, such as Wan (Wan et al., 2025) and MiniMax-H3 (MiniMax, 2026), learn visual priors about human motion and environmental interactions. Cosmos 3 (Agarwal et al., 2026) extends multimodal generation to video and action within a Mixture-of-Transformers architecture. Its action modeling includes egocentric head and hand motion, and its action training inherits the vision noise schedule. This provides a starting point for adapting video world models to full-body 3D motion generation. Related work has also explored joint video–motion generation. CoMoVi (Zhao et al., 2026a) uses separate diffusion branches for RGB video and 2D motion representations, together with 3D–2D cross-attention to predict 3D motion within the same denoising loop. World2Motion adapts the existing multimodal backbone of Cosmos 3 for joint video–motion generation. We combine synthetic and real video–motion training pairs and propose a shift-decoupled noise schedule that assigns modality-specific noise levels through shared denoising progress to improve temporal stability.

3 METHOD

Given a single image containing a person and a text prompt describing the intended action, World2Motion jointly generates a video sequence and corresponding 3D human motion. We estimate the initial human pose from the input image using CameraHMR (Patel & Black, 2025). To use the priors learned during video world model pretraining, we adapt the existing Cosmos 3 backbone for full-body motion generation. Figure 2 provides an overview of the framework.

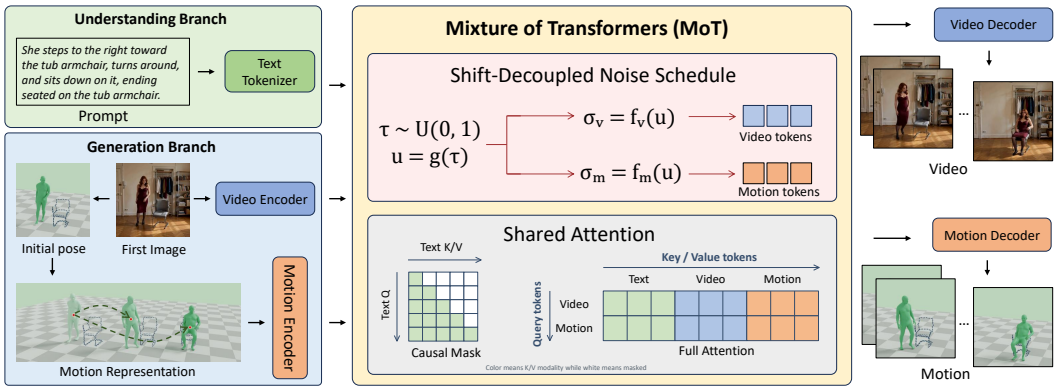

Figure 2: Overview of World2Motion. From an image and a text prompt, the adapted Cosmos 3 backbone jointly generates 3D human motion and video. CameraHMR estimates the initial pose from the image. The shift-decoupled schedule uses modality-specific noise shifts with shared denoising progress.

3.1 REVIEW OF COSMOS 3

Cosmos 3 (Agarwal et al., 2026) uses a Mixture-of-Transformers (MoT) backbone with an understanding branch and a generation branch. We retain its text and video encoding components, shared attention mechanism, positional encoding, and train both generated modalities with rectified-flow objectives.

3.2 HYBRID VIDEO–MOTION TRAINING DATA

To address the scarcity of paired supervision, we construct HOI-mix, a hybrid dataset combining synthetic video–motion pairs with real videos paired with estimated 3D motion. For the synthetic component, we render motion sequences from existing interaction datasets, including OMOMO (Li et al., 2023), CHAIRs (Jiang et al., 2023b), and SAMP (Hassan et al., 2021a), and use the rendered animations to guide MiniMax-H3 (MiniMax, 2026) in generating corresponding RGB videos. The original motion sequences provide known 3D supervision. For the real component, we estimate 3D motion from curated Bilibili videos using CameraHMR. We use Gemini 2.5 (Comanici et al., 2025) to generate text descriptions for both components. These sources combine motion supervision from existing 3D datasets with the diversity of environments and appearances in real videos.

3.3 FULL-BODY MOTION ADAPTATION

We represent each SMPL (Loper et al., 2015) motion frame using root translation and 6D joint rotations, including the root orientation. This representation allows generation to be conditioned on an initial pose without requiring velocity estimates. Following ARDY (Zhao et al., 2026b), we subtract the initial root position from the root translation at each frame. We convert the axis-angle joint rotations to 6D rotations, normalize the combined translation and rotation features, and project them into the Transformer hidden dimension using a motion encoder. A learnable modality embedding distinguishes motion tokens from video tokens. Motion time coordinates are scaled to match the video VAE's fourfold temporal compression, aligning the two modalities in time.

We train the motion encoder and decoder from scratch and initialize the remaining model weights from Cosmos 3. The understanding branch remains frozen during training. Our ablation study compares relative and absolute root translations and shows lower jerk and foot sliding with the relative representation (Table 3).

3.4 Shift-Decoupled Noise Schedule

To improve temporal stability in joint video–motion generation, we propose a shift-decoupled noise schedule. It uses separate noise shifts for video and motion while retaining shared denoising progress $u \in [0, 1]$. At each training step, we sample one value of u and compute the video and motion noise levels as

$$\sigma_v(u) = \frac{s_v u}{1 + (s_v - 1)u}, \qquad \sigma_m(u) = \frac{s_m u}{1 + (s_m - 1)u}, \tag{1}$$

where s_v and s_m are positive shifts for video and motion, respectively. The shifts control how shared progress maps to each modality's noise level. Setting $s_v > s_m$ keeps video at a higher noise level than motion during intermediate denoising steps, as illustrated in Figure 3(a). Each modality receives a timestep embedding corresponding to its own noise level. The shared progress constrains the training noise pairs to a single curve in the video–motion noise space, as illustrated in Figure 3(b).

During inference, both modalities follow a common progress grid from $u = 1$ to $u = 0$, using the same shifts and mappings as in training. Both noise schedules reach zero at the final step. Among the tested configurations, $s_v = 10$ and $s_m = 1$ achieve the lowest jerk and foot sliding rate. Matching the inference shifts to the training shifts also improves both metrics compared with mismatched settings (Table 4).

After joint generation, we inverse-normalize the motion output and recover global root translation by adding the initial root position to the predicted relative translation. We then apply a Savitzky–Golay filter (Savitzky & Golay, 1964) as post-processing. We disable this filter in all ablation experiments to evaluate the effects of the motion representation and noise schedule directly.

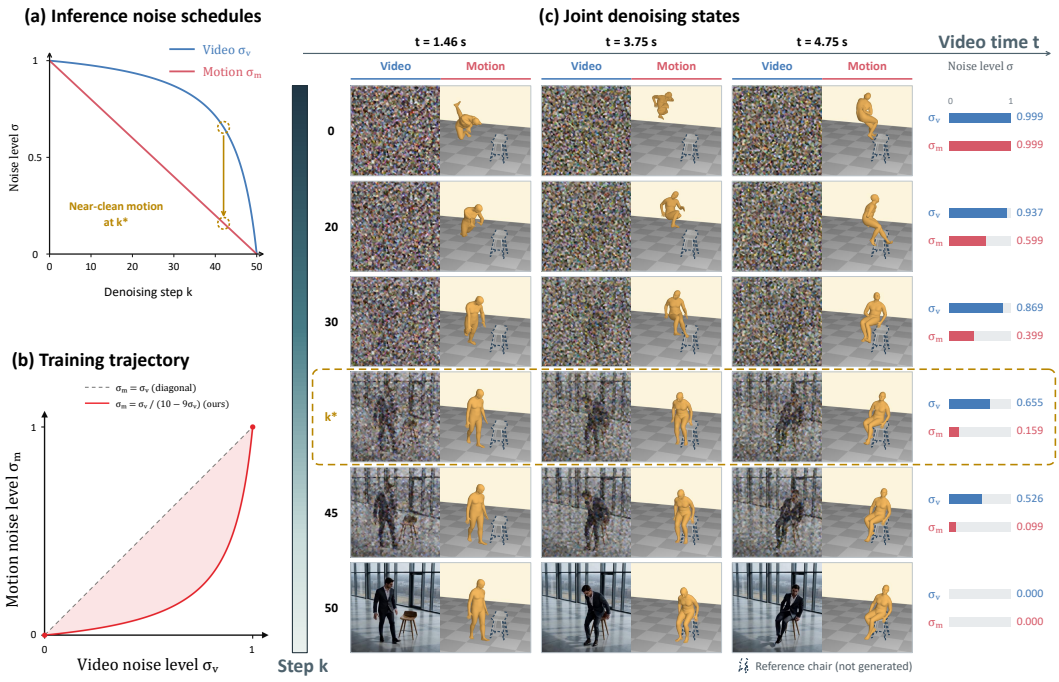

Figure 3: Shift-decoupled video–motion denoising. (a) Video and motion noise levels during inference with $s_v = 10$ and $s_m = 1$. (b) The corresponding training noise pairs follow a shared curve, with lower motion noise than video noise at intermediate steps. (c) Intermediate video frames and motion states. Rows show denoising steps, and column groups show video times. The chair mesh is reference geometry for visualization.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Dataset Construction and Splits. Following Section 3.2, HOI-mix combines SAMP (Hassan et al., 2021a), CHAIRs (Jiang et al., 2023b), and OMOMO (Li et al., 2023) motions paired with synthesized RGB videos and curated Bilibili videos with CameraHMR (Patel & Black, 2025) motion estimates. The training set contains 20,812 video–motion–text triplets totaling 29.15 hours, each with 121 frames at 24 fps. These triplets represent 10,847 distinct motion clips, with two appearance variants per interaction motion and one video per Bilibili clip. Motions share a common camera coordinate system; training uses sample-level shuffling across the merged dataset without source-specific reweighting. The legacy validation split contains 30 interaction and 20 non-interaction clips. Interaction data are held out by source sequence, and Bilibili data by clip. The final interaction benchmark contains 100 cases: 12 from CHAIRs, 12 from OMOMO, 6 from SAMP, and 70 additional cases from COUCH (Zhang et al., 2022). All benchmark source sequences are disjoint from the current training manifest. However, the benchmark retains the 30 interaction cases from the legacy validation split and is therefore not independent of that split. The separate 35-case ablation set is described in Section 4.5.

Implementation Details. We initialize World2Motion from Cosmos 3 NANO (16B) (Agarwal et al., 2026), which includes a Qwen3-VL (8B) text backbone (Bai et al., 2025), 36 Mixture-of-Transformers layers, and a Wan2.2 causal VAE (Wan et al., 2025). The motion encoder and decoder are trained from scratch, and the understanding branch remains frozen. We train for 20,000 iterations with AdamW (Loshchilov & Hutter, 2019), a learning rate of 5×10^{-5}, a cosine decay schedule, and 200 warmup steps. Text conditioning is dropped with probability 0.1 for classifier-free guidance (Ho & Salimans, 2022). Training uses bfloat16 precision, a batch size of 16, and 16 NVIDIA H100 GPUs, and takes approximately 44 hours.

Baselines. We compare with text-to-motion methods HY-Motion (Team, 2025), Kimodo (Rempe et al., 2026), and Go-to-Zero (Fan et al., 2025); interaction methods LIGHT (Wang et al., 2026) and HOIFHLI (Wu et al., 2025); MiniMax-H3 (MiniMax, 2026) + CameraHMR; and CoMoVi (Zhao et al., 2026a), fine-tuned on our data. We initialize Kimodo, HOIFHLI, CoMoVi, and World2Motion with poses estimated by CameraHMR (Patel & Black, 2025). LIGHT and HOIFHLI additionally require 3D object information, while World2Motion uses the scene image.

4.2 EVALUATION

Motion Quality and Text Alignment. Following MoMask (Guo et al., 2024), we report Fréchet Inception Distance (FID) for motion distributions, and R-Precision at rank 1 (R@1) and multimodal distance (MM-Dist) for motion–text alignment. Jerk and foot sliding rate (FSR) follow WHAM (Shin et al., 2024) and OmniControl (Xie et al., 2024). In Table 2, jerk uses native frame rates and units of $10\,\mathrm{m/s^3}$; FSR measures the percentage of adjacent frame pairs with sliding on either contacting foot after resampling to 20 fps. Ablation protocols differ (Section 4.5); table arrows indicate the preferred direction.

Scene Interaction. Following OMOMO (Li et al., 2023), we report Contact-F1. We render generated motions in the corresponding 3D scenes and use Qwen3-VL (Bai et al., 2025) to assess whether the requested interaction is completed. Interaction Success Rate (ISR) is the fraction of test cases judged successful. These measures complement motion quality: a smooth motion can match the action description while failing to reach the intended object.

Video–Motion Consistency. For joint generation methods, we recover motion from the generated video using CameraHMR (Patel & Black, 2025) and compare it with the directly generated motion. PoseError is the pelvis-relative mean Euclidean distance over 22 corresponding body joints, in millimeters. We exclude invalid CameraHMR frames and average within each sequence, then across sequences. This measures cross-modal pose agreement, rather than ground-truth 3D accuracy or agreement of the root trajectories.

Inference Efficiency. We report mean inference time in seconds. Main-comparison World2Motion outputs use Savitzky–Golay (Savitzky & Golay, 1964) post-processing (Section 3); all ablations disable this filter.

4.3 Quantitative Comparisons

Table 2: Quantitative comparison on 100 interaction cases. FID measures motion distribution quality; R@1 and MM-Dist measure text alignment; Contact-F1 and ISR evaluate scene interaction. Jerk and FSR measure motion stability, and PoseError measures video–motion agreement. Time is mean inference time in seconds. Bold denotes the best reported value per column; dashes indicate unreported scores.

Method	FID↓	R@1↑	MM-Dist↓	Contact-F1↑	ISR↑	Jerk↓ ($10\,\mathrm{m/s^3}$)	FSR↓ (%)	PoseError↓ (mm)	Time (s)↓
Text-to-motion									
HY-Motion (Team, 2025)	0.46	0.10	1.21	0.09	0.51	3.43	**1.25**	–	2.83
Kimodo (Rempe et al., 2026)	0.66	0.05	1.30	0.03	0.23	35.14	6.07	–	**1.93**
Go-to-Zero (Fan et al., 2025)	0.51	0.04	1.35	0.07	0.36	17.50	3.42	–	2.57
Human–object interaction									
HOIFHLI (Wu et al., 2025)	0.82	0.02	1.37	0.09	0.12	10.62	6.28	–	13.94
LIGHT (Wang et al., 2026)	0.62	0.03	1.35	0.17	0.46	**1.37**	11.60	–	25.88
Two-stage video generation and motion recovery									
MiniMax-H3 (MiniMax, 2026) + CameraHMR (Patel & Black, 2025)	0.69	0.10	1.32	**0.24**	**0.91**	37.22	6.41	–	523.67
Joint video–motion generation									
CoMoVi (Zhao et al., 2026a)	0.71	0.12	1.25	0.09	0.69	16.10	24.45	81.28	270.45
World2Motion	**0.27**	**0.24**	**1.13**	0.22	**0.91**	8.35	2.80	**67.50**	156.59

Motion Quality and Text Alignment. In Table 2, World2Motion achieves the lowest FID and MM-Dist and the highest R@1 among the compared methods. Relative to HY-Motion, FID decreases from 0.46 to 0.27 and R@1 increases from 0.10 to 0.24. It also improves over CoMoVi, whose FID and R@1 are 0.71 and 0.12, respectively.

Scene Interaction. World2Motion reaches an ISR of 0.91, compared with 0.23–0.51 for the evaluated text-to-motion methods and 0.12–0.46 for the human–object interaction methods. Its Contact-F1 of 0.22 also exceeds those of these baselines. The two-stage pipeline achieves the same ISR and a higher Contact-F1 of 0.24. Successful completion and contact accuracy are therefore distinct: matching the overall interaction success of the two-stage method does not imply identical contact quality. The remaining contact gap motivates more precise supervision for human–object interactions.

Motion Stability. World2Motion has lower jerk and FSR than Kimodo, Go-to-Zero, HOIFHLI, and the two video-based baselines. HY-Motion achieves lower values on both stability metrics, while LIGHT achieves the lowest jerk but has a higher FSR of 11.60. Thus, the best result depends on which aspect of stability is measured. World2Motion improves stability over the compared video-based pipelines, but it does not surpass every motion-only method. We retain both jerk and FSR to distinguish temporal irregularity from sliding artifacts.

Video–Motion Consistency. Compared with CoMoVi, World2Motion reduces PoseError from 81.28 to 67.50, a 17.0% reduction. This indicates closer agreement between its generated video and directly predicted 3D motion under the motion-recovery evaluation. It complements the improvements in text alignment and scene interaction, which assess the generated motion against the prompt and environment rather than against the other output modality.

Inference Efficiency. World2Motion takes 156.59 seconds on average, compared with 523.67 seconds for MiniMax-H3 + CameraHMR and 270.45 seconds for CoMoVi: speedups of approximately $3.34\times$ and $1.73\times$, respectively. The text-to-motion and human–object interaction methods remain faster. Our efficiency gain concerns video-based motion generation. World2Motion provides both a video and an explicit 3D motion sequence, whereas the faster motion-only baselines do not produce the corresponding video. The results support reducing the cost of obtaining these paired outputs while preserving competitive interaction performance.

4.4 QUALITATIVE COMPARISONS

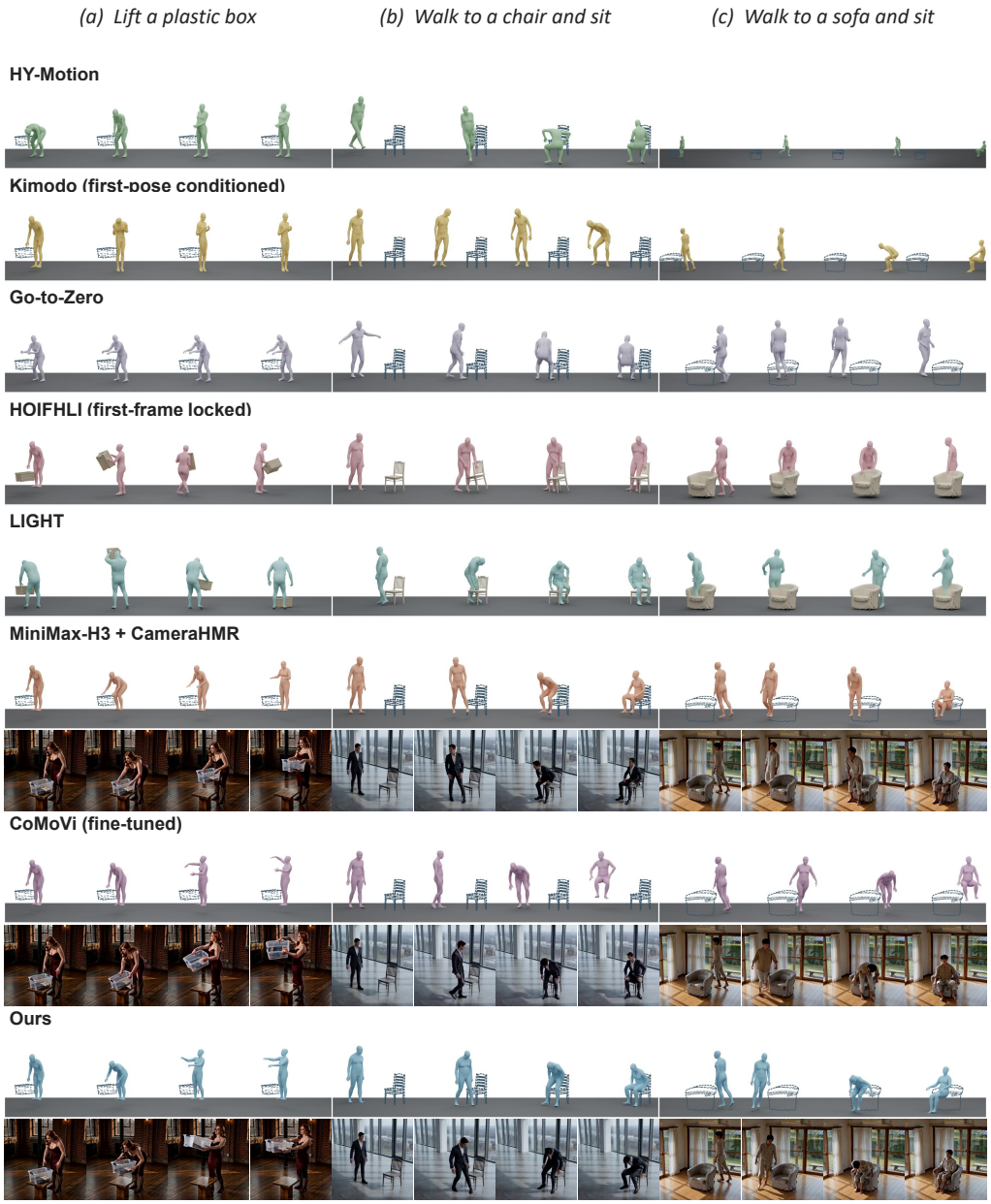

Figure 4: Qualitative comparison of scene interactions: (a) lifting a plastic box, (b) walking to a chair and sitting, and (c) walking to a sofa and sitting. Each method shows successive motion states; video-generating methods also show corresponding video frames. Human–object interaction baselines include their predicted object trajectories.

Figure 4 compares actions that depend on both body motion and placement relative to an object. In the sitting examples, producing a plausible sitting pose is insufficient: the character must first approach the intended seat and then align the body with it. The box example also involves coordinating the body with object motion. Viewing the approach and final interaction together helps distinguish action matching from scene-consistent execution. For video-based methods, the paired rows additionally allow the reader to inspect whether the generated video and recovered or directly generated 3D motion describe the same interaction.

4.5 Ablation Studies

All ablations use 10,000-iteration checkpoints and 35 clips (15 demonstration clips and 20 internet-video clips), without post-processing; Table 2 uses the smoothed 20,000-iteration model on 100 interaction cases. Ablation jerk uses 25 fps. FSR pools sliding observations over contacting foot-joint/frame pairs from four foot joints on a normalized skeleton. Unlike Table 2, its denominator excludes non-contact observations, so the scores are not directly comparable across protocols. The relative-translation row in Table 3 and the matched $s_m = 1$ entry in Table 4 share the same outputs.

Root translation. With other settings fixed, relative root translation reduces jerk from 4.355 to 1.251 (71.3%) and FSR from 0.766 to 0.457 (40.3%) compared with absolute translation (Table 3). Both configurations use the same 6D joint-rotation representation. These results support the relative coordinate choice for temporal stability in the tested setting, without root velocity features or post-processing.

Table 3: Root translation ablation on 35 cases without post-processing. Raw Jerk uses 25 fps ($10^3\,\mathrm{m/s^3}$). Contact FSR is the sliding fraction among contacting foot-joint/frame pairs on a normalized skeleton (0–1); its definition differs from Table 2.

Translation	Raw Jerk ↓ ($10^3\,\mathrm{m/s^3}$)	Contact FSR ↓ (0–1)
Absolute	4.355	0.766
Relative	**1.251**	**0.457**

Shift-decoupled noise schedule. In Table 4, matched training and inference shifts of $s_m = 1$ reduce jerk and FSR by 23.5% and 28.0% relative to matched $s_m = 10$. Matching shifts gives the lowest values within every row. After training with $s_m = 1$, changing only the inference shift to 10 increases jerk from 1.251 to 8.003. These results support modality-specific shifts and training–inference consistency for the tested stability metrics.

Table 4: Noise-schedule ablation on the same 35 cases, using Raw Jerk and contact-conditioned FSR as defined in Table 3. These unfiltered ablation scores are not directly comparable to Table 2. Columns give the inference motion shift s_m; $s_v = 10$ is fixed. Bold marks the lowest value per metric within each row.

	Raw Jerk ($10^3\,\mathrm{m/s^3}$) ↓			Contact FSR (0–1) ↓		
Training s_m	1	10	15	1	10	15
1	**1.251**	8.003	9.703	**0.457**	0.986	1.000
10	7.021	**1.635**	2.557	0.985	**0.635**	0.830
15	8.413	2.127	**1.809**	0.981	0.700	**0.649**

5 Limitations

World2Motion provides a baseline for adapting pretrained video world models to 3D human motion generation, but several limitations remain. First, our training data mainly cover single-person motion and simple interactions, leaving more complex scenarios largely unexplored. Second, despite the improvement in inference speed, World2Motion is still too slow for real-time use. Third, accurate root placement and contact remain challenging. Our evaluation also has limits: a high ISR or low PoseError does not imply precise contact. Future work could expand interaction coverage, incorporate scene geometry and object states, and reduce inference time. For now, generated motions can serve as starting points for further editing. The supplementary video shows examples of agent-assisted smoothing and contact refinement.

6 Conclusion

We present World2Motion, which adapts Cosmos 3 to jointly generate scene-aware 3D human motion and video from a single image and a text prompt. We train the model on synthetic video–motion pairs and real videos with estimated motion, and use a shift-decoupled noise schedule to reduce motion jitter. On our interaction benchmark, World2Motion improves motion–text alignment over the compared methods and matches the interaction success rate of the two-stage baseline.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript editing and organization. Uses of generative models in training-data construction and evaluation are described in Sections 3.2 and 4.1.

REFERENCES

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Fabien Baradel, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas. Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In *European Conference on Computer Vision*, pp. 202–218. Springer, 2024.

Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36:11454–11468, 2023.

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.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Ke Fan, Shunlin Lu, Minyue Dai, Runyi Yu, Lixing Xiao, Zhiyang Dou, Junting Dong, Lizhuang Ma, and Jingbo Wang. Go to zero: Towards zero-shot motion generation with million-scale data. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13336–13348. IEEE, 2025.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14737–14748. IEEE, 2023.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1900–1910. IEEE, 2024.

Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11374–11384, 2021a.

Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14708–14718, 2021b.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Hsin-Ping Huang, Yang Zhou, Jui-Hsien Wang, Difan Liu, Feng Liu, Ming-Hsuan Yang, and Zhan Xu. Move-in-2d: 2d-conditioned human motion generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22766–22775. IEEE, 2025.

Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16750–16761, June 2023.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023a.

Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9331–9342. IEEE, 2023b.

Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1737–1747. IEEE, 2024.

Hwanhee Jung, Seunggwan Lee, Jeongyoon Yoon, SeungHyeon Kim, Giljoo Nam, Qixing Huang, and Sangpil Kim. Decoupled generative modeling for human-object interaction synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2253–2263, 2026.

Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7122–7131. IEEE, 2018.

Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2252–2261, 2019.

Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas. Nifty: Neural object interaction fields for guided human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 947–957, 2024.

Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023.

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*, pp. 54–72. Springer, 2024.

Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3382–3392. IEEE, 2021.

Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, pp. 590–606. Springer, 2022.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

MiniMax. Minimax h3: An open model breaking the boundaries between tasks and modalities. https://www.minimax.io/blog/minimax-h3, July 2026.

Priyanka Patel and Michael J Black. Camerahmr: Aligning people with perspective. In *2025 International Conference on 3D Vision (3DV)*, pp. 1562–1571. IEEE, 2025.

Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Pu Wang, Minwoo Lee, Srijan Das, and Chen Chen. Bamm: Bidirectional autoregressive motion model. In *European Conference on Computer Vision*, pp. 172–190. Springer, 2024.

Ekkasit Pinyoanuntapong, Muhammad Saleem, Korrawe Karunratanakul, Pu Wang, Hongfei Xue, Chen Chen, Chuan Guo, Junli Cao, Jian Ren, and Sergey Tulyakov. Maskcontrol: Spatio-temporal control for masked motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9955–9965, 2025.

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:2603.15546*, 2026.

Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2070–2080. IEEE, 2024.

Chentao Song, He Zhang, Haolei Yuan, Haozhe Lin, Jianhua Tao, Hongwen Zhang, and Tao Yu. Metrichmsr: Metric human mesh and scene recovery from monocular images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21132–21142, 2026.

Tencent Hunyuan 3D Digital Human Team. Hy-motion 1.0: Scaling flow matching models for text-to-motion generation. *arXiv preprint arXiv:2512.23464*, 2025.

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Mdm: Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. *Advances in Neural Information Processing Systems*, 35:14959–14971, 2022.

Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say, interact as you can: Language-guided human motion generation with scene affordance. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 433–444. IEEE, 2024.

Ziyin Wang, Sirui Xu, Bing Zhou, Jiangshan Gong, Jian Wang, Yu-Xiong Wang, Liang-Yan Gui, et al. Unleashing guidance without classifiers for human-object interaction animation. In *International Conference on Learning Representations*, volume 2026, pp. 46121–46141, 2026.

Zhen Wu, Jiaman Li, Pei Xu, and C Karen Liu. Human-object interaction from human-level instructions. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11176–11186. IEEE, 2025.

Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. In *International Conference on Learning Representations*, volume 2024, pp. 28176–28194, 2024.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023a.

Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 364–373. IEEE, 2023b.

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, 2024.

Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. In *European Conference on Computer Vision*, pp. 518–535. Springer, 2022.

Xiaohan Zhang, Sebastian Starke, Alexander Winkler, Federica Bogo, Samir Aroudj, and Yuting Ye. Giraf: Towards generalizable human interactions with articulated objects. *arXiv preprint arXiv:2607.07880*, 2026.

Chengfeng Zhao, Jiazhi Shu, Yubo Zhao, Tianyu Huang, Jiahao Lu, Zekai Gu, Chengwei Ren, Zhiyang Dou, Qing Shuai, and Yuan Liu. Comovi: Co-generation of 3d human motions and realistic videos. *arXiv preprint arXiv:2601.10632*, 2026a.

Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. In *European Conference on Computer Vision*, pp. 311–327. Springer, 2022.

Kaifeng Zhao, Mathis Petrovich, Haotian Zhang, Tingwu Wang, Siyu Tang, and Davis Rempe. Autoregressive diffusion with hybrid representation for interactive human motion generation. *ACM Transactions on Graphics (TOG)*, 45(4):1–14, 2026b.

Yiwen Zhao, Ce Zheng, Yufu Wang, Hsueh-Han Daniel Yang, Liting Wen, and Laszlo A Jeni. Onlinehmr: Video-based online world-grounded human mesh recovery. *arXiv preprint arXiv:2603.17355*, 2026c.

Yude Zou, Junji Gong, Xing Gao, Zixuan Li, Tianxing Chen, and Guanjie Zheng. Infbagel: Human-object-scene interaction generation with dynamic perception and iterative refinement. *arXiv preprint arXiv:2604.04843*, 2026.

A Additional Qualitative Results

Figure 5 presents box lifting and carrying, approaching and sitting on a stool or armchair, and two yoga movements. The sitting examples show the approach to a seat as well as the transition from standing to sitting, allowing the reader to inspect body placement throughout the sequence. The yoga examples show full-body pose changes without object manipulation.

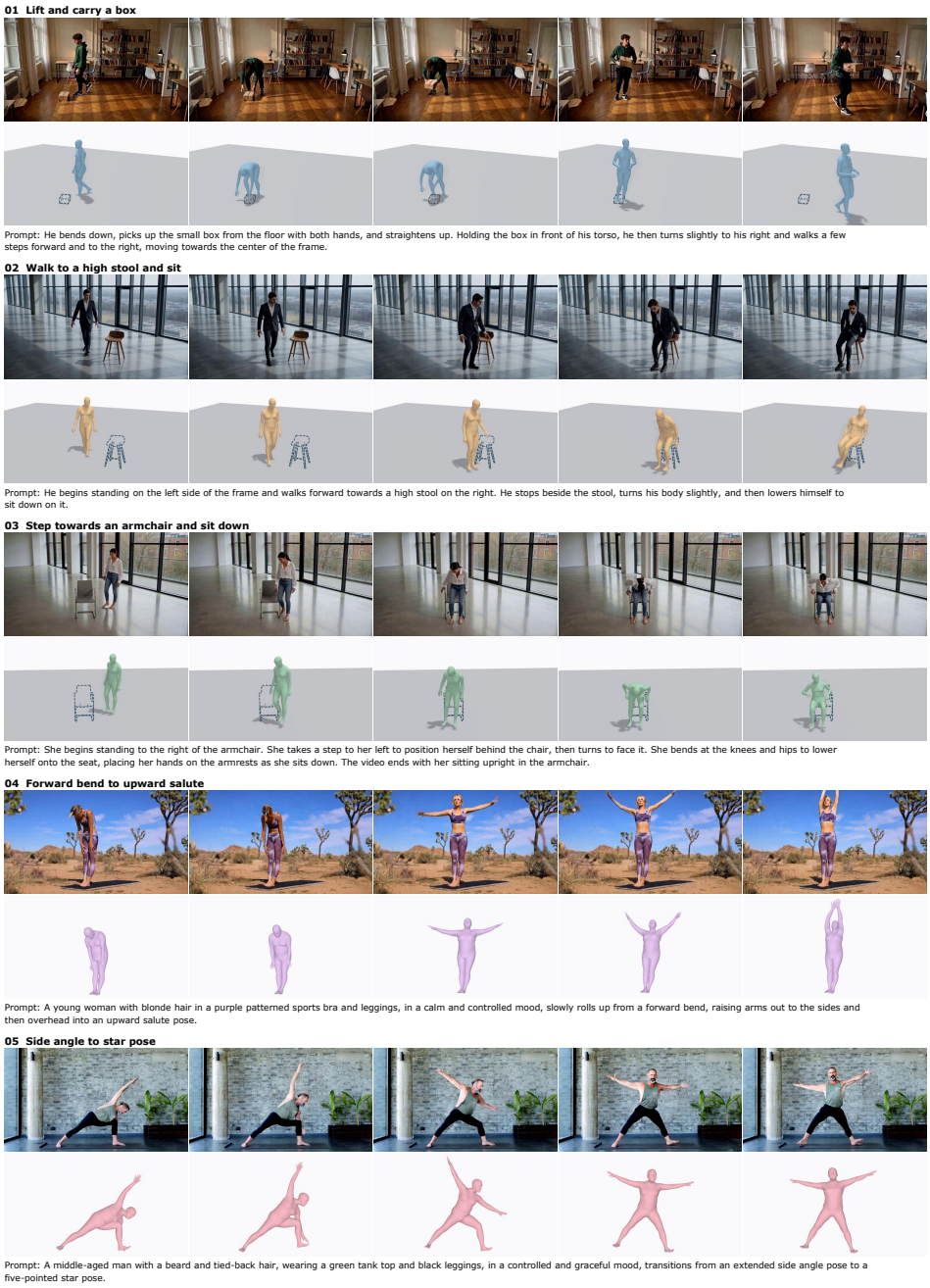

Figure 5: Qualitative results of World2Motion for box lifting and carrying, sitting on a stool or armchair, and two yoga movements. Each example shows five successive video frames (top) and corresponding 3D motion renderings (bottom), ordered from left to right. The input text prompt appears below each example.

Figure 6 adds sofa sitting, picking up and carrying a vacuum cleaner, standing up from a stool, and two dance sequences. These examples include both changes in body position relative to furniture and coordinated limb movements. In both figures, the paired video and motion rows show five corresponding time points, making it possible to compare how the requested action progresses in the two outputs. The text prompt is provided below each example.



Figure 6: Additional qualitative results of World2Motion for sofa sitting, picking up and carrying a vacuum cleaner, standing up from a stool, and two dance sequences. Each example shows five successive video frames (top) and corresponding 3D motion renderings (bottom), ordered from left to right. The input text prompt appears below each example.