Supplementary Materials for Object-Aware 4D Human Motion Generation

Shurui Gui*[1,2†] Deep Patel*[1] Xiner Li[2] Martin Renqiang Min[1]

[1]NEC Laboratories America [2]Texas A&M University

{shurui.gui, lxe}@tamu.edu {dpatel, renqiang}@nec-labs.com

*Equal contribution.

1. Technical Appendices and Supplementary Material

1.1. Evaluation Metrics

Pose Plausibility. A lower KL divergence indicates that the pose is more similar to those seen during VPoser's training, and thus more plausible. The final Pose Plausibility score for a video is the average $\mathcal{L}_{\text{plaus},t}$ over all T frames:

$$M_{\text{Plausibility}} = \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_{\text{plaus},t}. \tag{1}$$

It is worth noting that pose plausibility utilizes a pretrained variational autoencoder, whose performance can be constrained by its original training data, potentially limiting generalization to out-of-distribution poses.

Pose Variation. we first compute the standard deviation σ_j for each of the K pose parameters across time:

$$\sigma_j = \text{std}(\{\phi_{1,j}, \phi_{2,j}, \ldots, \phi_{T,j}\}), \quad j = 1, \ldots, K. \tag{2}$$

A higher value indicates more significant changes in pose throughout the video, suggesting more dynamic motion. The Pose Variation metric is then the mean of these standard deviations:

$$M_{\text{Variation}} = \frac{1}{K}\sum_{j=1}^{K}\sigma_j. \tag{3}$$

Trajectory Length. To assess the extent of global character movement within the 3D space, we calculate the trajectory length of the root joint. For each frame t, HMR2.0 provides the 3D keypoint coordinates. We extract the root joint's 3D position $\mathbf{k}_t = (x_t, y_t, z_t)$. The total trajectory length is the sum of Euclidean distances between the root joint positions in consecutive frames:

$$M_{\text{Trajectory}} = \sum_{t=1}^{T-1}||\mathbf{k}_{t+1} - \mathbf{k}_t||_2. \tag{4}$$

A longer trajectory length suggests more substantial displacement of the character over time.

†Work done during internship at NEC Laboratories America.

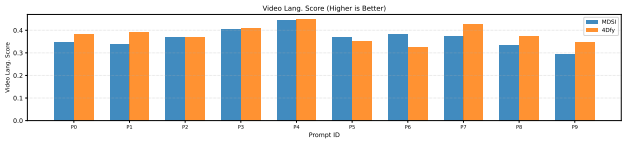

Figure 1. **Video Language Score** Comparison of MSDI and 4Dfy.

Video-Language Score. To measure the semantic alignment between the input text prompt and the generated video, we use InternVideo2 [4], a video-text foundation model. For each of the $N_{cv} = 4$ generated views, we compute the cosine similarity between the text embedding of the prompt and the video embedding. The final Video-Language Score is the average of these similarity scores across all views. A higher score indicates better prompt-video alignment.

4Dfy often achieves a higher **Video Language Score** (See Figure 1), this may stem from a bias in the metric towards static or motion-limited scenes. Consequently, the metric might prioritize overall scene-text alignment over nuanced motion quality, potentially favoring 4Dfy despite its weaker human motion dynamics. This observation is pertinent, as previous works have often relied on image-based metrics (e.g., CLIP [2] scores) for video-text alignment, which are arguably even less sensitive to temporal dynamics. Moreover, the human evaluation study shows that video generated by our method has high preference (75%) over 4Dfy for Text Alignment (TA). This shows that the video text alignment scores using video language models does not truly capture the human perception of motion quality.

1.2. User Study Methodology.

We followed human evaluation methodology established by 4D-fy [1] and MAV3D [3]. We collected responses from 11 human evaluators. For a diverse set of 10 text prompts, each evaluator was shown a pair of videos generated by MSDI and 4D-fy. Participants were asked to choose the superior video based on five criteria:

- **Appearance Quality (AQ):** The visual clarity and appeal

Prompt ID	Prompt Text
0	the human walks around the table in a circle and stops close to the start position
1	the human prepares to jump for 1 second then jumps over the fence
2	the human jumps from the stepstool onto the ground
3	the human walks on the clouds
4	the human walks towards the lamp
5	the human falls down from the stepstool
6	the human crawls under the table
7	the human prepares to jump for 1 second then jumps onto the table and stops on the surface of the table for 1 second
8	the human falls down on the ground
9	the human sits down on ground with legs cross

Table 1. List of text prompts used for evaluation.

of the generated human and object.

- **3D Structure Quality (SQ):** The realism and consistency of the 3D shapes across multiple viewpoints.
- **Motion Quality (MQ):** The naturalness, dynamism, and physical plausibility of the human's movements.
- **Text Alignment (TA):** How accurately the video's content reflects the input text prompt.
- **Overall Preference (OP):** The evaluator's subjective choice for the better video, considering all the above aspects.

1.3. Evaluation Prompts

Table 1 lists the text prompts used for the quantitative and qualitative evaluation.

1.4. Limitations

Despite its advancements, MSDI has several limitations offering avenues for future work.

First, the final output quality is tied to the pre-generated 3D assets and their initial placement and orientation. Suboptimal inputs or challenging initial setups (e.g., incorrect facing, distant objects) can hinder the generation of plausible interactions, as our framework doesn't currently optimize this initial scene layout.

Second, our reliance on LLMs for initial "coarse" trajectory generation can be a bottleneck. LLMs may produce suboptimal, physically impractical, or semantically incorrect paths for complex prompts or environments, providing a poor starting point for optimization.

Third, the framework struggles with fine-grained interactions, especially detailed hand and finger movements (e.g., realistically playing a drum, Figure 2). Current motion models and representations lack the specificity for such dexterous tasks, leading to generalized rather than precise contact.

Fourth, while our collision avoidance works for general movements, it may be less robust or efficient for highly complex object geometries or very intricate, close-quarters interactions.

Fifth, MSDI is currently designed for human interactions with static objects. Handling dynamic objects or multi-agent scenarios remains a future challenge.

Finally, the system's performance is dependent on the capabilities of the underlying pre-trained motion diffusion models, and the optimization process requires careful hyperparameter tuning to balance different objectives.

1.5. Compute Resources

All experiments were conducted on a system equipped with 1 NVIDIA A100 GPUs, 128 CPU cores, and 1TB of CPU memory. Generating a single 4D video clip with 4Dfy (all three of its stages) required approximately 24 hours. MSDI completed the generation of human and object artifacts followed by the optimization process in approximately 5 hours per prompt using the same computational resources.

1.6. Multi View Qualitative Results

Figures 3, 4, 5, 6, shows comparison of generated motion with 4Dfy and MSDI from different camera angles.

References

[1] Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7996–8006, 2024. 1

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[3] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. Text-to-4d dynamic scene generation. *arXiv preprint arXiv:2301.11280*, 2023. 1

[4] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024. 1

Figure 2. Generated motion for the prompt: *"the human is playing a drum"*. Top: 4Dfy. Bottom: MSDI

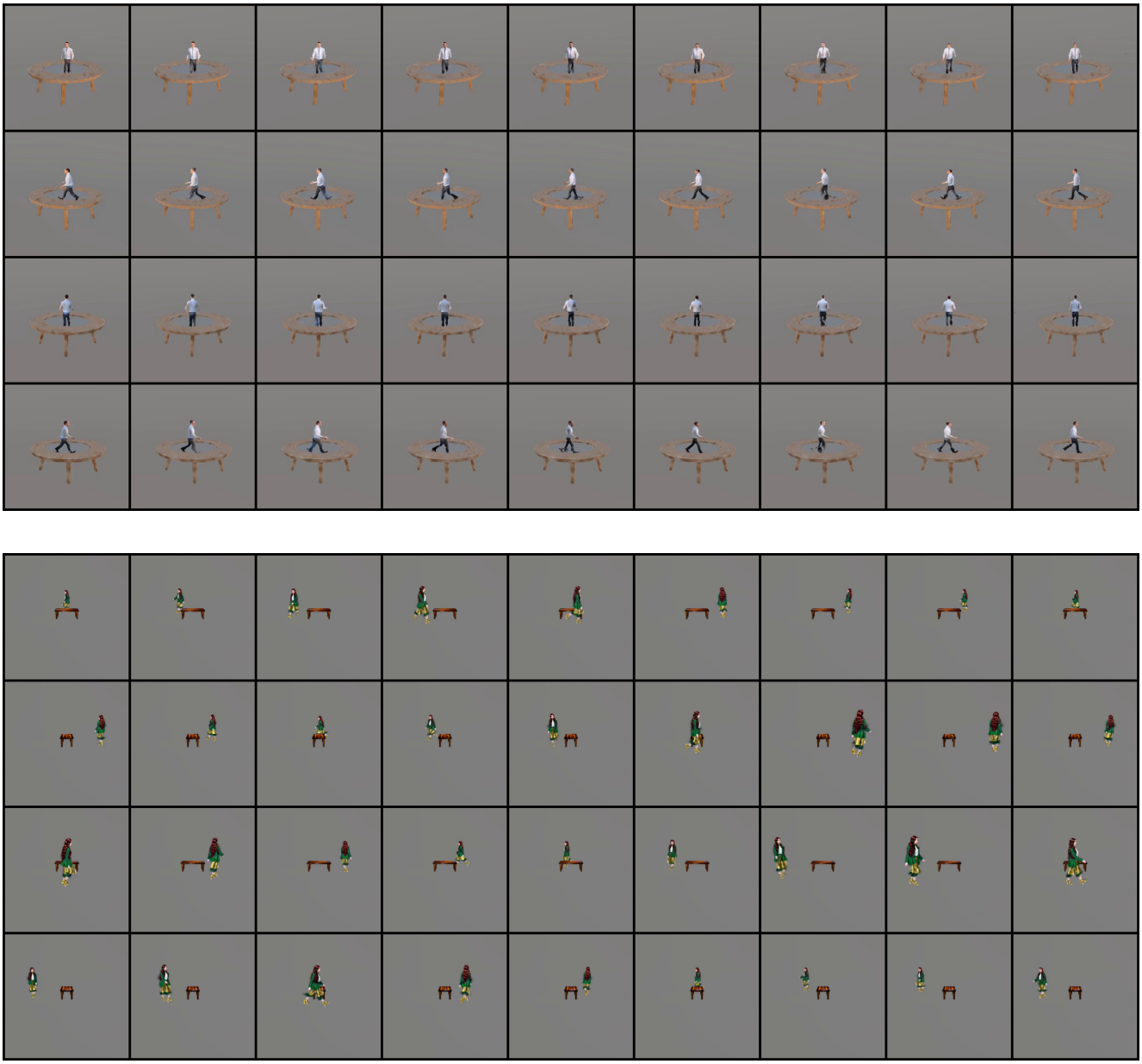

Figure 3. Generated motion for the prompt: *"a human walks around a table in a circle and stops close to the start position"*. Top: 4Dfy. Bottom: MSDI. The four rows illustrate the motion from different camera viewpoints

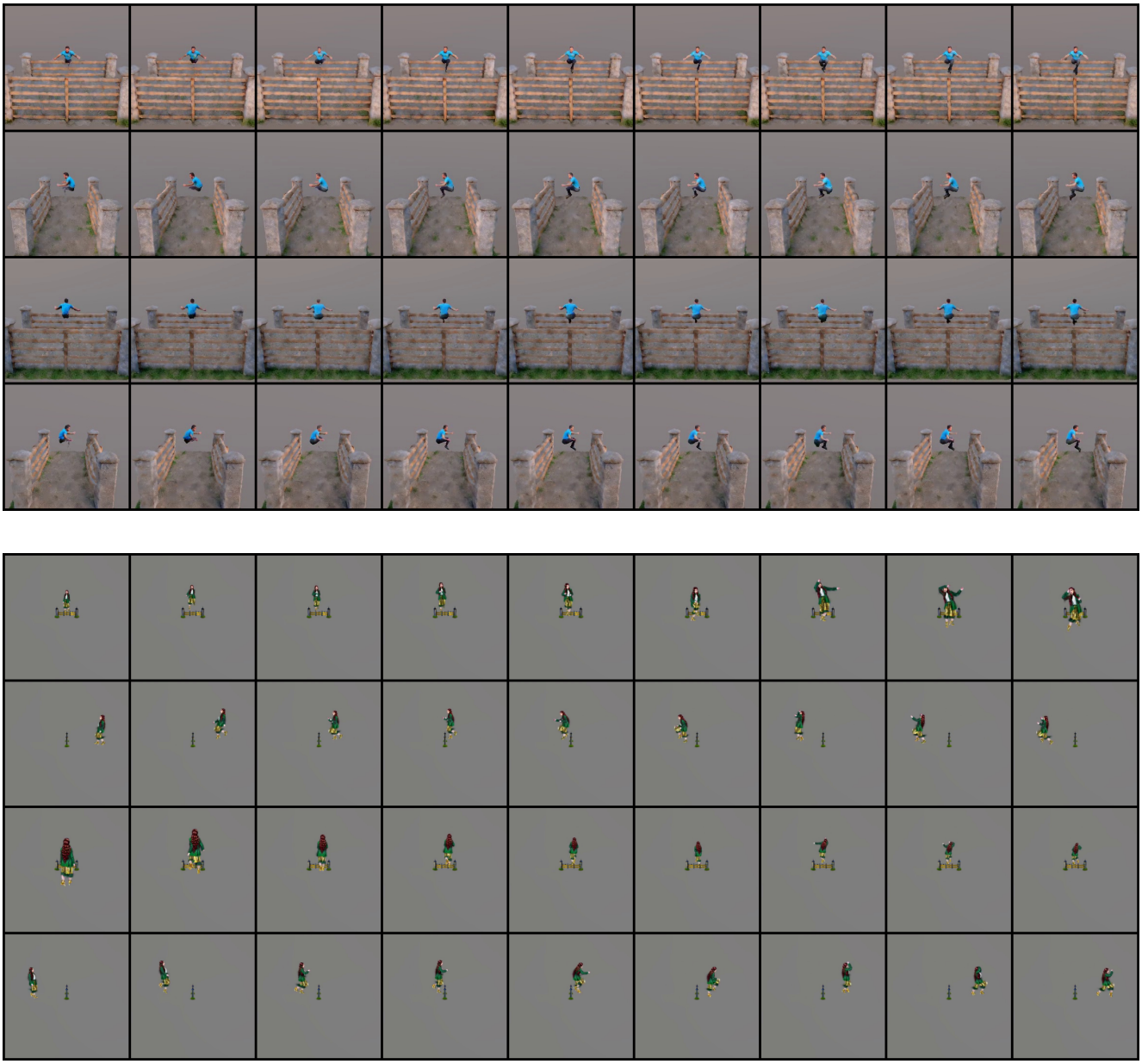

Figure 4. Generated motion for the prompt: *"the human prepares to jump for 1 second then jumps over the fence"*. Top: 4Dfy. Bottom: MSDI. The four rows illustrate the motion from different camera viewpoints

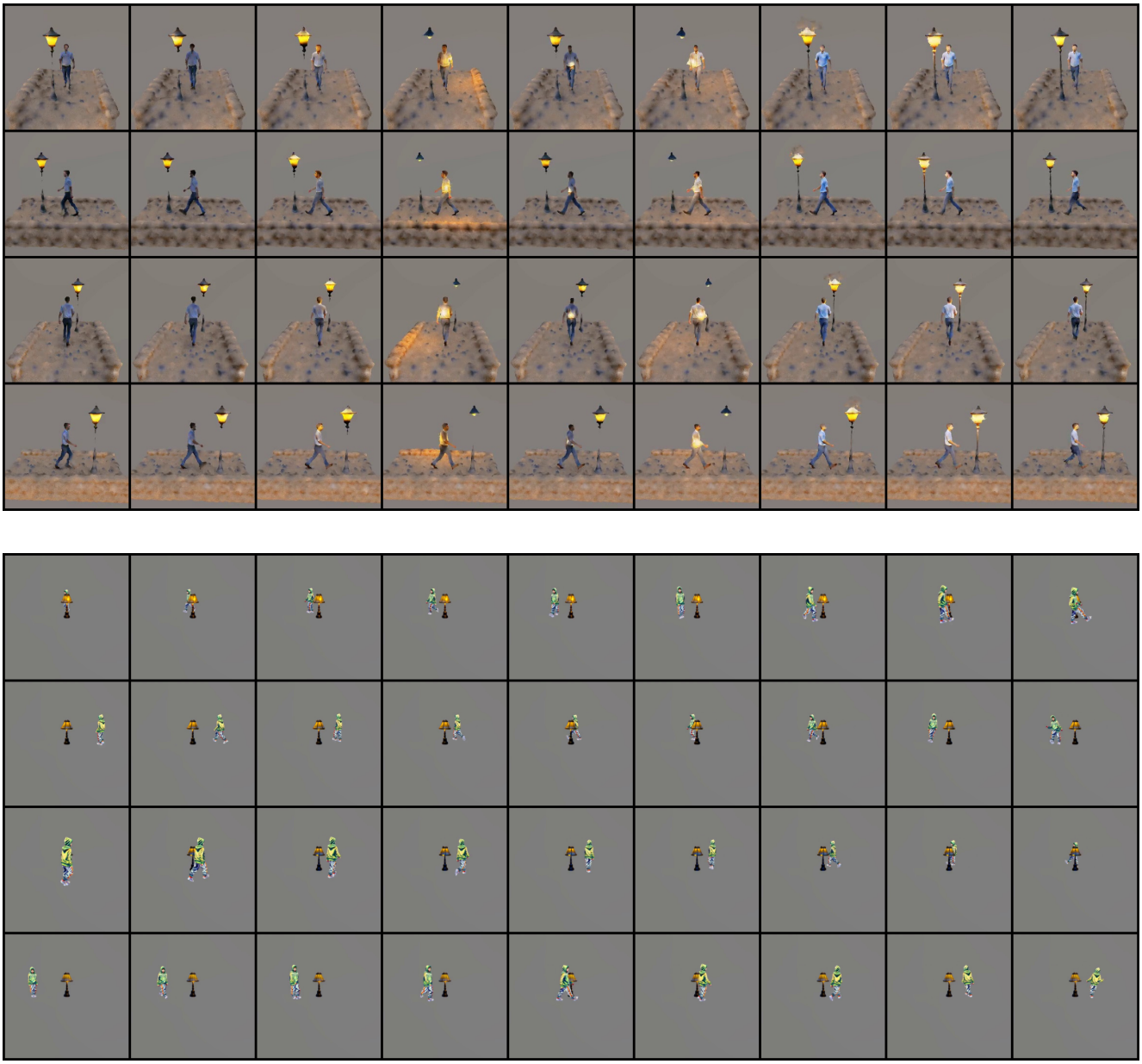

Figure 5. Generated motion for the prompt: *"the human walks towards the lamp"*. Top: 4Dfy. Bottom: MSDI. The four rows illustrate the motion from different camera viewpoints

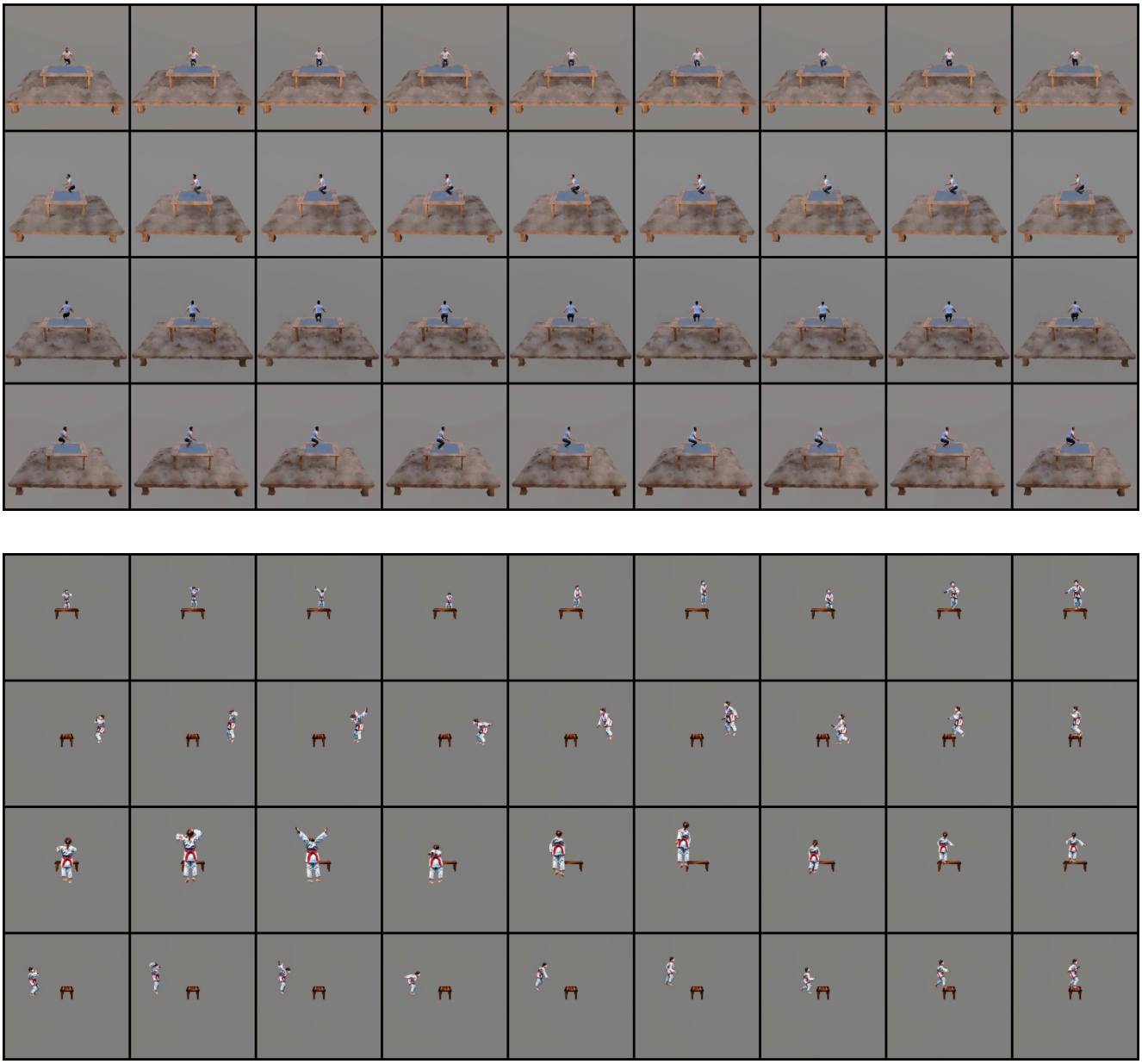

Figure 6. Generated motion for the prompt: *"the human prepares to jump for 1 second then jumps onto the table and stops on the surface of the table for 1 second"*. Top: 4Dfy. Bottom: MSDI. The four rows illustrate the motion from different camera viewpoints