

SimMotionEdit: Text-Based Human Motion Editing with Motion Similarity Prediction

Supplementary Material

A. Perceptual Study

In addition to our quantitative and qualitative comparisons, we have also conducted a perceptual study with human participants to understand how human observers view the quality of motions edited by SimMotionEdit, particularly in comparison to the best available baseline TMED [1] and the corresponding ground truth edited motions.

A.1. Setup

Each participant in our study observed 10 sample sets corresponding to 10 randomly selected pairings of source motions and textual edit instructions. Within each sample set, a participant observed the source motion, the text instructions, and three versions of the edited motion — one each for the ground truth, SimMotionEdit, and TMED. We put these three edited motions in a random order within each sample set to avoid any positional biases in the study. We asked the participants to respond on two metrics:

- *Alignment of edited motions with text: How well does the edited motion follow the edit instructions, irrespective of the motion quality?* On this metric, the participants responded in a 3-point Likert Scale with the following descriptions: “Edited motion is completely different from the edit instructions” (1), “Edited motion only follows some parts of the edit instructions” (2), and “Edited motion fully follows the edit instructions” (3),
- *Plausibility of edited motions: How is the quality of the edited motion, irrespective of whether it follows the edit instructions?* On this metric, the participants responded in a 3-point Likert Scale with the following descriptions: “Edited motion has severe issues, e.g., self-intersections, unrealistic poses or movements, etc.” (1), “Edited motion has minor issues, e.g., foot sliding, plausible but non-humanlike poses or movements, etc.” (2), and “Edited motion has no issues” (3).

We show an example layout with one source motion, one edit instruction, and one edited motion in Fig. A.1a and the scoring instructions and scoring area for the participants in Fig. A.1b.

A.2. Results

A total of 15 participants, consisting of students and staff from a University campus, responded to our perceptual study, leading to a total of 150 responses across the 10 sample sets. We show the distribution of participant scores aggregated over all the responses in Fig. A.2, and report the mean statistics in Tab. A.1. We observe that SimMotionEdit

Table A.1. **Perceptual Evaluation Mean Statistics.** We report the mean scores achieved by all three candidates in the perceptual study, averaging the aggregated responses across all the participants and sample sets. SimMotionEdit achieves scores that are 0.3 to 0.5 points higher than TMED on a 3-point Likert Scale.

| Metric | Candidate | Mean Score \uparrow |
|--------------|----------------------|-----------------------------------|
| Alignment | Ground Truth | 2.53 ± 0.65 |
| | SimMotionEdit (ours) | 2.17 ± 0.78 |
| | TMED [1] | 1.70 ± 0.81 |
| Plausibility | Ground Truth | 2.79 ± 0.51 |
| | SimMotionEdit (ours) | 2.37 ± 0.70 |
| | TMED [1] | 1.99 ± 0.81 |

Table B.1. **Additional Performance and Efficiency Evaluations.** Compared to TMED, we report comparable generated-to-source accuracy measures and lower L@ distance and FID measures.

| Method | Generated-to-Source (Batch) | | | L2 Dist. | FID |
|--------|-----------------------------|----------------|----------------|------------------|--------------|
| | R@1 \uparrow | R@2 \uparrow | R@3 \uparrow | (m) \downarrow | \downarrow |
| GT | 74.01 | 84.52 | 89.91 | — | — |
| TMED | 71.77 | 84.07 | 89.52 | 0.278 | 0.167 |
| Ours | 72.71 | 83.54 | 87.50 | 0.253 | 0.110 |

is consistently scored higher than TMED [1], outperforming it by 40% to 50% on score distributions and by an absolute 0.3 to 0.5 points on the mean score on a 3-point Likert Scale across the two metrics.

B. Additional Quantitative Evaluations

For the preservation of the source motion, we follow TMED and use generated-to-source retrieval to measure motion preservation. Tab. B.1 shows the comparable performance of our method to TMED. For other fidelity-related metrics, since TMED does not provide the implementation of FID and L2 distance, we implement our own for a fair comparison. We see that our method outperforms TMED.


C. Additional Qualitative Results

As shown in Fig. C.1, our method can edit dance and crawling motions. It also successfully follows complex edit instructions.

Set 1/10


Please look at the source motion (in blue) and the following edit instruction, and evaluate the qualities of the three different versions of the edited motion (in orange) below.

Source Motion:



Edit instructions:
take shorter steps in the same direction

Edited Motion 1/3:



(a) Upper Part

Scoring Instructions

Alignment: How well does the edited motion follow the edit instructions, irrespective of the motion quality?

1: Edited motion is completely different from the edit instructions
2: Edited motion only follows some parts of the edit instructions
3: Edited motion fully follows the edit instructions

Plausibility: How is the quality of the edited motion, irrespective of whether it follows the edit instructions?

1: Edited motion has severe issues, e.g., self-intersections, unrealistic poses or movements, etc.
2: Edited motion has minor issues, e.g., foot sliding, plausible but non-humanlike poses or movements, etc.
3: Edited motion has no issues

| | 1 | 2 | 3 |
|--------------|-----------------------|-----------------------|-----------------------|
| Alignment | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| Plausibility | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

(b) Lower Part

Figure A.1. **Perceptual Study Layout.** (*Upper Part*) We show an example of our study layout with one source motion, one edit instruction, and one edited motion. (*Lower Part*) We show the scoring instructions and scoring area for all the samples in the perceptual study.

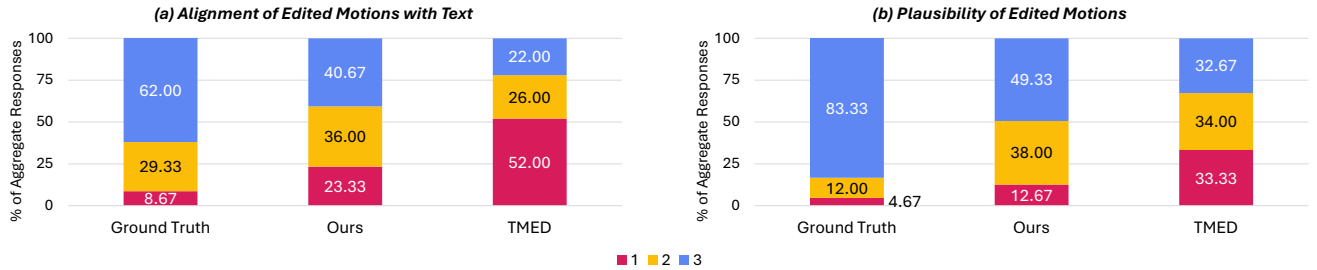


Figure A.2. **Perceptual Evaluation Score Distributions.** We show the distributions of aggregate responses from participants on the three versions of edited motions — Ground Truth, SimMotionEdit, and TMED [1] — on the two metrics of *Alignment* and *Plausibility*. We observe that participants have marked 3 for SimMotionEdit about 40% to 50% more times than TMED across the two metrics.

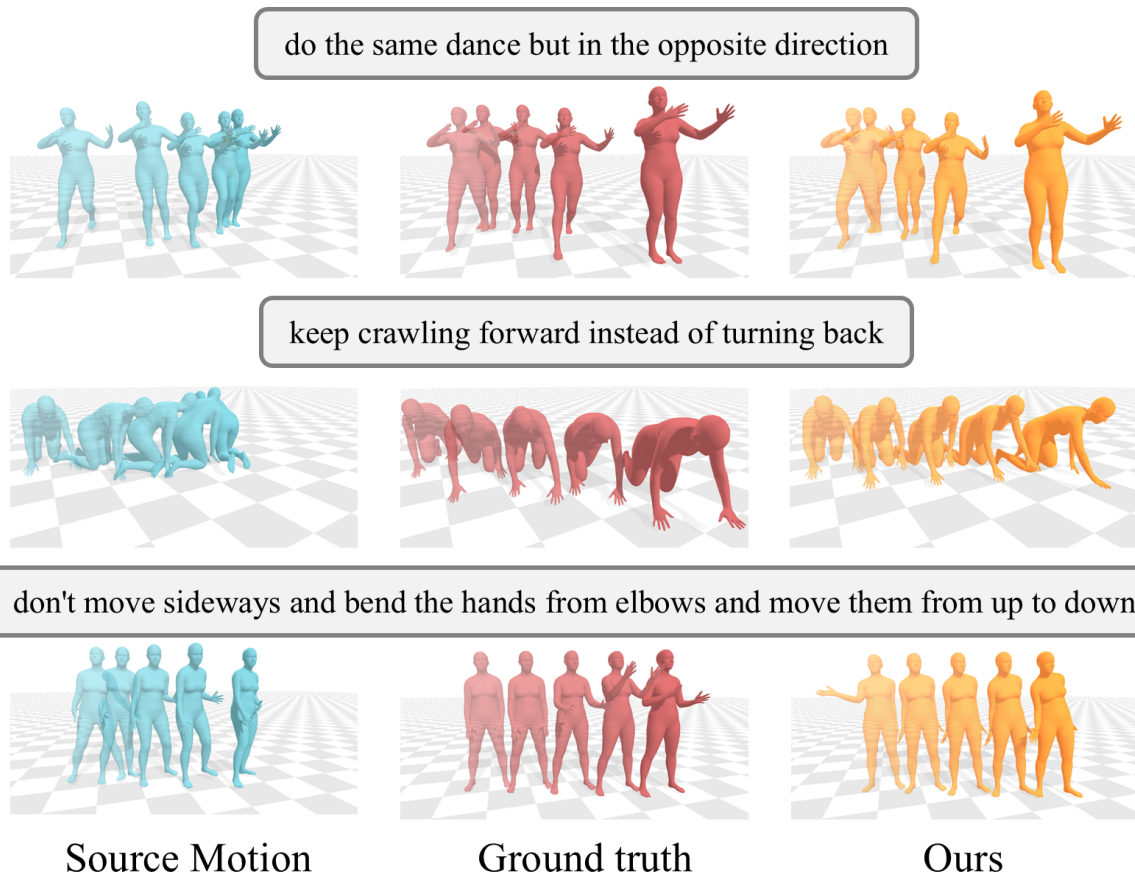


Figure C.1. **More Qualitative Results.**

References

- [1] Nikos Athanasiou, Alpár Ceske, Markos Diomataris, Michael J. Black, and Gül Varol. MotionFix: Text-driven 3d human motion editing. In *ACM SIGGRAPH Asia*, 2024. 1, 2