

FUSION: Full-Body Unified Motion Prior for Body and Hands via Diffusion

Supplementary Material

1. Architecture

Figure 1 shows the model architecture of our motion prior. It is a transformer-based diffusion model with 19.7M trainable parameters.

2. Optimization Hyperparameters

Tab. 1 lists the optimization hyperparameters used in evaluation and ablation studies. The chosen configuration balances task compliance (λ_{close}) with the plausibility of the resulting motion (λ_{foot} , λ_{ch} , λ_{lk}).

Stage	Epoch	λ_{lk}	λ_{foot}	λ_{ch}	λ_{close}	λ_{decorr}
1	800	1	0.1	0.5	1	2
2	800	0.1	0.0	0.1	1	2

Table 1. Optimization configuration. The same configuration applies across all ablation and evaluation settings.

3. Evaluation Metrics

Trajectory Error is the percentage of unsuccessful trajectories, defined as those with any keyframe location error exceeding a specified distance threshold. Location Error is the percentage of keyframe locations that exceed this threshold. Average Error measures the mean distance between the generated motion locations and the keyframe locations, measured at the keyframe motion steps.

4. Dataset Curation

Fig. 5 shows an example of random merging of the global body motion with local hand motion. Although it is rare, random merging of the local hand motion and the global body motion might introduce additional self-intersection. To address this issue, we perform a self-penetration test to ensure the obtained motions are physically possible. We compute face intersections between the hand mesh and the rest of the body mesh caused by copying the local hand motion. If there is an intersection exceeding a threshold, we discard this random merging in favor of another randomly sampled local hand motion.

Random merging inevitably incurs some information loss in hand-body coordination. This trade-off is intentional: it enables training on substantially more data while preserving meaningful cross-modal dependencies. To verify this, we conduct a Canonical Correlation Analysis (CCA) on local 6D joint rotations, comparing hand-body motions from Ground Truth (GT) sequences—i.e., data with inherent full-body annotations—against our curated dataset, which combines such sequences (by $\sim 85\%$, see

Fig. 3) and the motions that are obtained through augmenting with local hand motions. Fig. 2 reports the CCA and corresponding Mutual Information values for the first five canonical variables. The leading canonical component yields correlations of 0.97 and 0.94, respectively, indicating that despite some information loss, our curated data retains a strong cross-modal correlation.

5. Qualitative Results

For hand motion synthesis, we conducted a perceptual study comparing *FUSION* with both the GT and HMP [1] on the Keypoint Tracking task, which involves tracking all 10 fingertips on our test split. In total of 7 participants were asked to evaluate the realism of 120-frame test sequences involving the interaction of both hands with objects. Each participant evaluated 10 to 30 sequences, yielding 128 paired comparisons. Videos were presented side-by-side in random order as shown in Fig. 6. Objects were included only for visualization purposes, to help users better distinguish between the motions. The evaluation criteria included hand jitter, naturalness of articulation, and object penetration.

We then test the null hypothesis H_0 that two methods are perceived as equally realistic (*FUSION* vs. GT and *FUSION* vs. HMP [1]) using a two-sided exact sign test on ordinal preferences (ties ignored). Let n^+ and n^- denote preferences for *FUSION* and the baseline under H_0 , $n^+ \sim \text{Binomial}(n^+ + n^-, 0.5)$. We obtain $p = 2.26 \times 10^{-7}$ for *FUSION* vs. GT and $p = 7.11 \times 10^{-9}$ for *FUSION* vs. [1]. Subject preference ratios indicate that *FUSION* synthesizes more realistic hand motions than HMP.

6. Self-Interaction

We use our optimization framework with a prompted instance of GPT-o3 [2] to curate the self-interaction dataset. In total, there are 94 sequences with textual descriptions, contact conditions, and the corresponding human motions after optimization. This small dataset is shared in the codebase.

6.1. Sampling Body Vertices

We sample 236 vertices along with their semantic labels (e.g. *right_elbow_inside*, *left_biceps*) on the human body that may be important for self-interaction tasks. Figures 7 to 9 show the sampled vertices with a focus on the body, hands, and head, respectively.

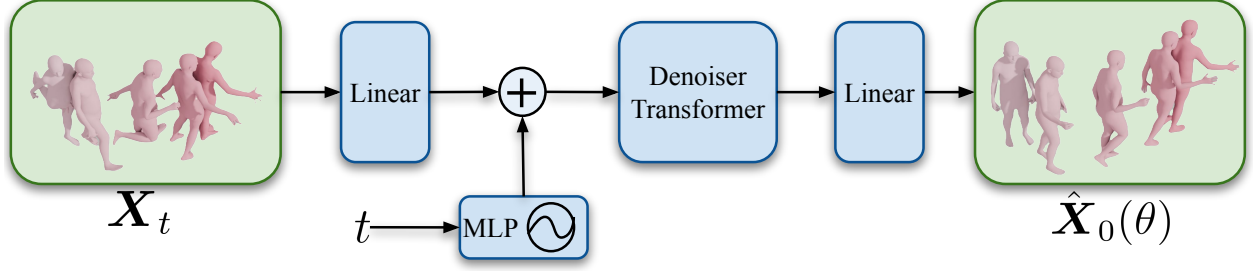


Figure 1. **FUSION Model Architecture:** Given a noisy motion X_t and the corresponding noising timestep t , *FUSION* first performs linear projection, and concatenates it with the time embeddings of the noising timestep. This concatenation forms the input to the Denoiser Transformer. Then it performs linear projection to the Denoiser Transformer output and obtains denoised motion $\hat{X}_0(\theta)$.

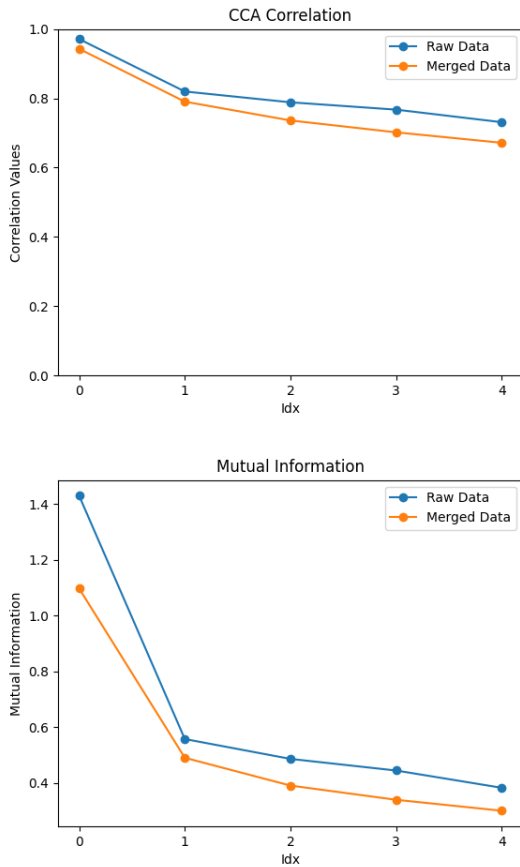


Figure 2. Canonical Correlation Analysis and Mutual Information between hand and body motions for the first five canonical variables, revealing the degree of linear and nonlinear statistical dependence across modalities.

6.2. Quantitative Evaluation

Fig. 10 shows the devised pipeline for the evaluation of LLM planning. We run this all 94 every input sentence in our self-interaction. Tab. 3 shows every input–predicted

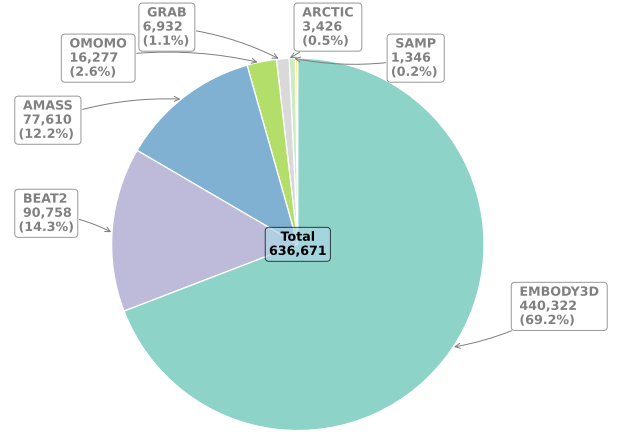


Figure 3. **Body Dataset Sequences:** Distribution of sequences across body motion datasets used for training. EMBODY3D, AMASS and BEAT2 constitute the majority of the training corpus, while smaller specialized datasets (OMOMO, GRAB, ARCTIC, and SAMP) provide additional diversity to the overall collection.

sentence pair and the corresponding BERTScore values [4]. The last two rows are for the pairs with the highest and the lowest F1 score.

The BERTScore analysis of the input–predicted sentence pairs shows generally high alignment between the intended actions and their corresponding descriptions, with most F1 scores clustering above 0.85, indicating strong semantic similarity. Actions involving direct tactile interactions, such as *rubbing the palms together* (F1 = 0.953), *rub your left calf* (F1 = 0.962), and *smell your armpit* (F1 = 0.977), achieve particularly high scores, reflecting clear and unambiguous mappings between input and evaluation sentences. Gestures that are slightly more abstract or involve sequential motion, such as *do a squat* (F1 = 0.932) or *crossing arms* (F1 = 0.873), show slightly lower but still reasonable alignment. Overall, the table reflects that the predictions are semantically faithful to the inputs, with minor variations in descriptive richness or specificity influencing BERTScore

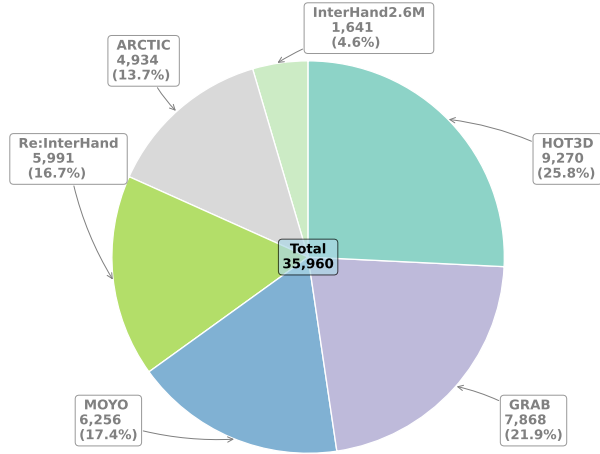


Figure 4. **Hand Dataset Sequences:** Distribution of sequences across hand motion datasets used for training. HOT3D constitutes the majority of the training corpus. GRAB, MOYO, Re:InterHand, ARCTIC, and InterHand2.6M datasets provide additional diversity to the overall collection.

Method \ Batch Size	1	2	4	8
TLControl [41]	91.0	154.0	247.8	565.8
<i>FUSION</i>	78.1	98.7	144.0	229.3

Table 2. Method runtime in seconds for different batch sizes.

variations.

6.3. Inference Runtime:

For a more comprehensive comparison with TLControl [3] + post-optimization, we report execution times on the Keypoint Tracking task on the HumanML3D dataset using an H100 GPU. *FUSION* performs DDIM sampling with only 5 denoising steps, making our iterative ODE solver lightweight. Our batched implementation scales sub-linearly with batch size (Tab. 2). Results show our method is 14.1%-59.5% faster than TLControl+post-optimization.

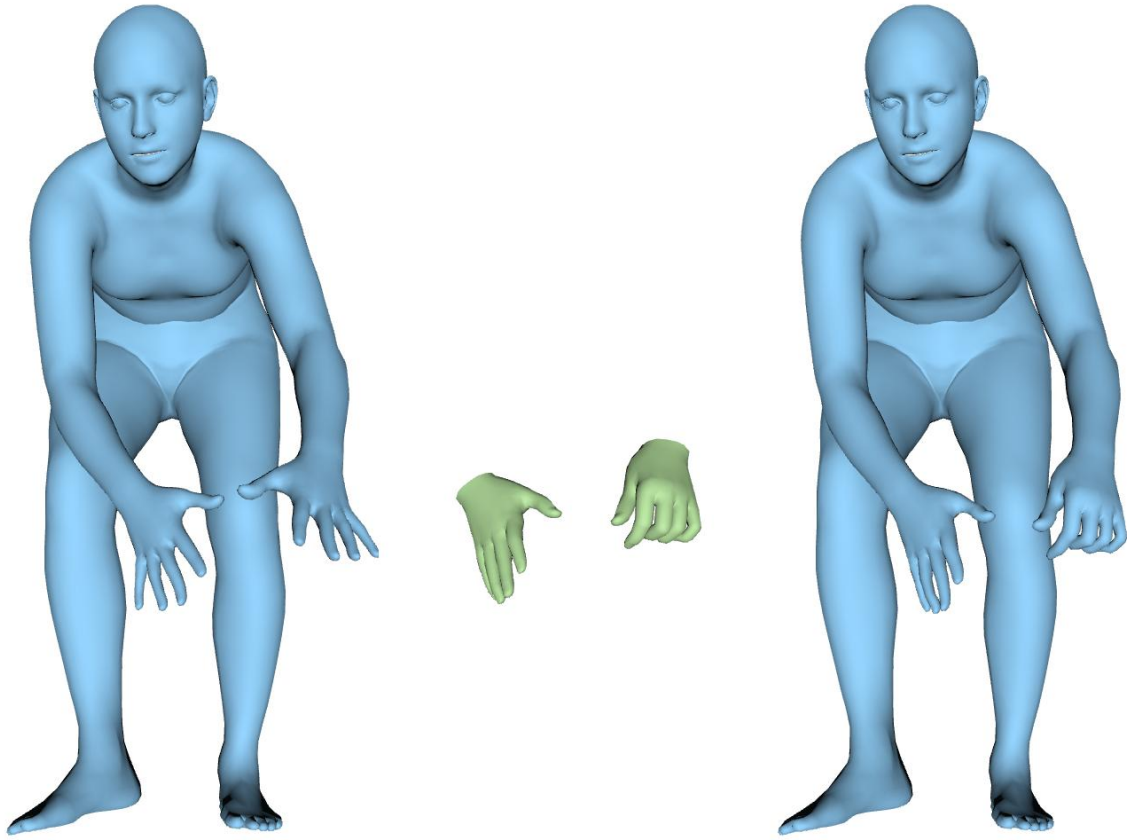


Figure 5. **FUSION Visualization of Random Body-Hand Pairing:** The mesh from the body dataset is shown on the left. The middle panel presents local hand meshes with randomly sampled poses, but with wrist orientations and locations taken from the GT data. On the right, we show the resulting full-body mesh obtained by combining the sampled hand poses with the corresponding body pose. Notably, self-penetration is effectively prevented, since the wrist orientations are directly derived from the body, ensuring consistent and plausible articulation.

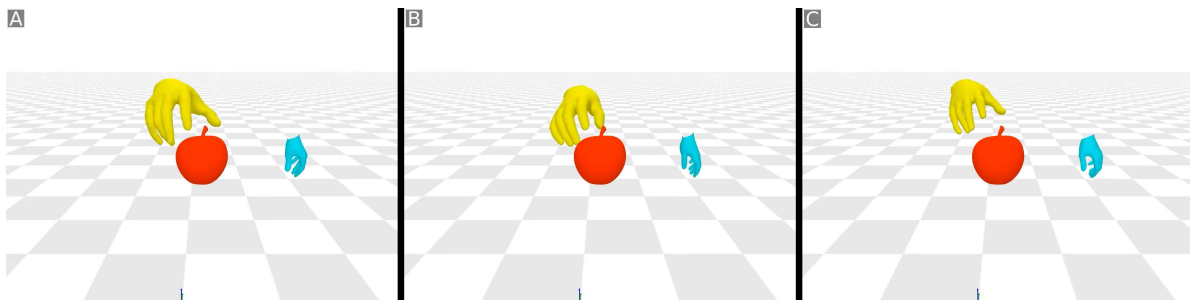


Figure 6. **User Study Graphical Interface:** Subjects were tasked with ranking the realism of the three options A, B, and C. Ordering is random for each interaction sample.

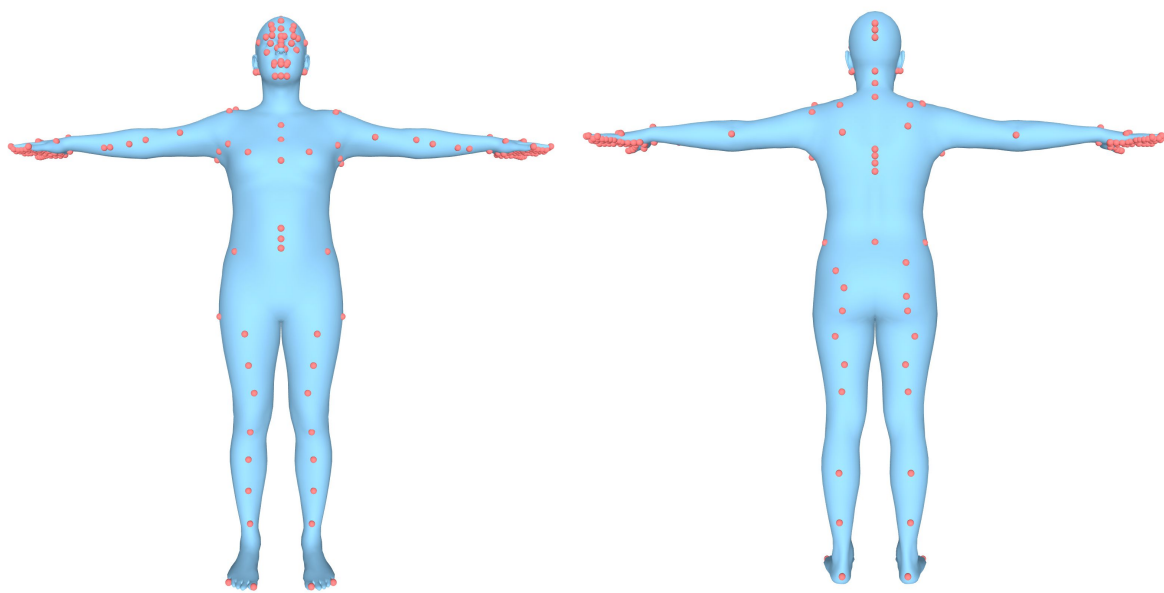


Figure 7. **Sampled vertices for the Self-Interaction Task:** We manually select 236 body vertices together with their semantic labels (*e.g.* *right_fingertip*, *left_biceps*, ...). The selected vertices are distributed across major body parts but are denser in the hands, as fine-grained hand articulations are more likely to come into contact during interaction. This manual selection ensures that the set covers contact-relevant regions, while keeping the number of vertices computationally manageable.

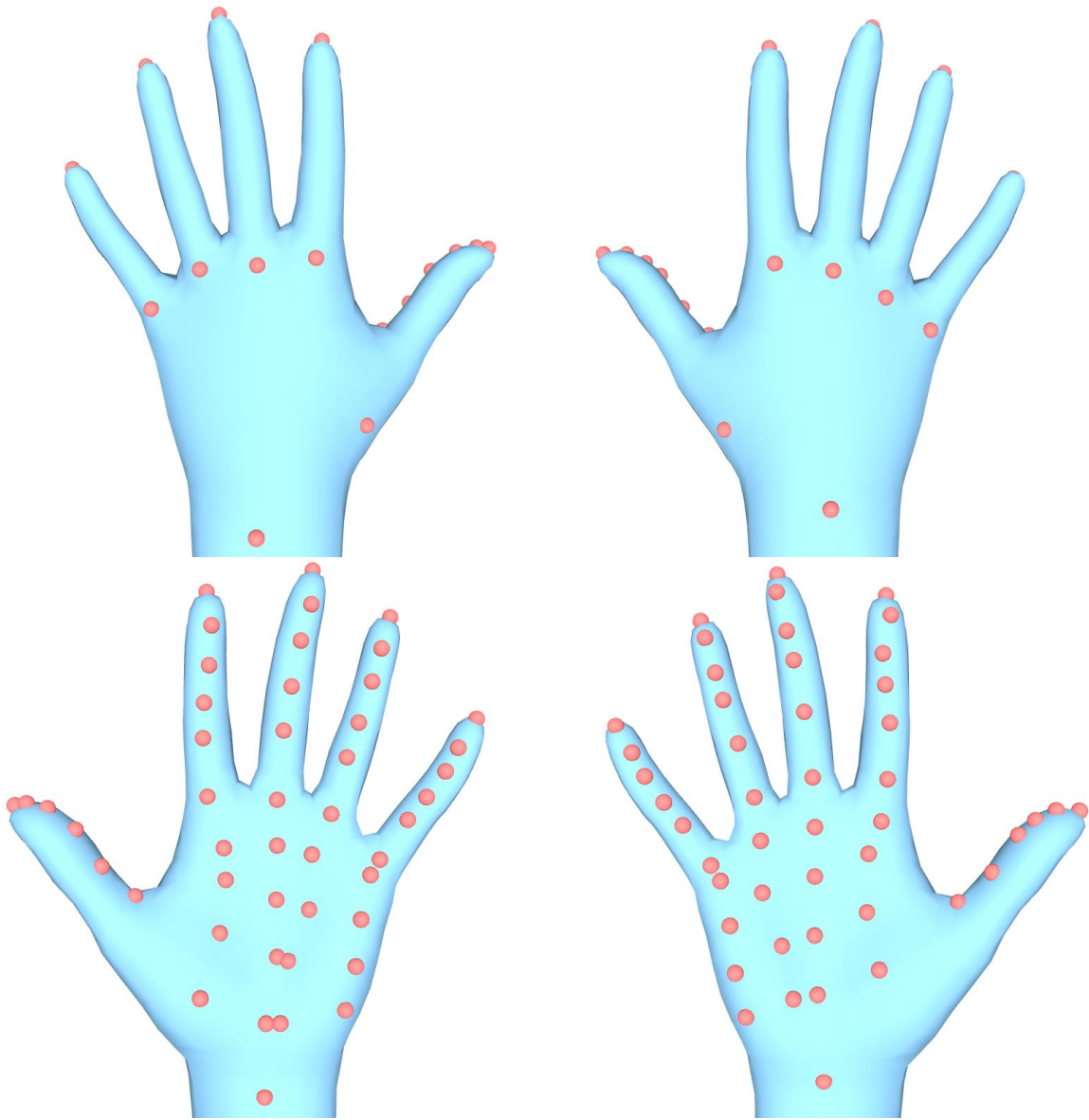


Figure 8. **Sampled hand vertices for the Self-Interaction Task:** The palms contain more sampled points than the back of hands to better capture contact in relevant tasks.

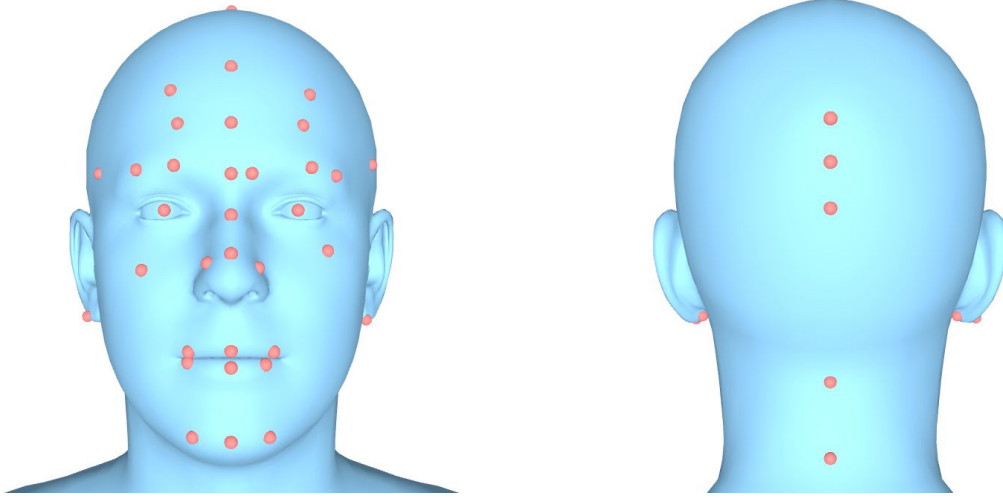


Figure 9. **Sampled head vertices for the Self-Interaction Task:** The face contains more sampled points than the back of the head to better capture contact in tasks such as touching the face.

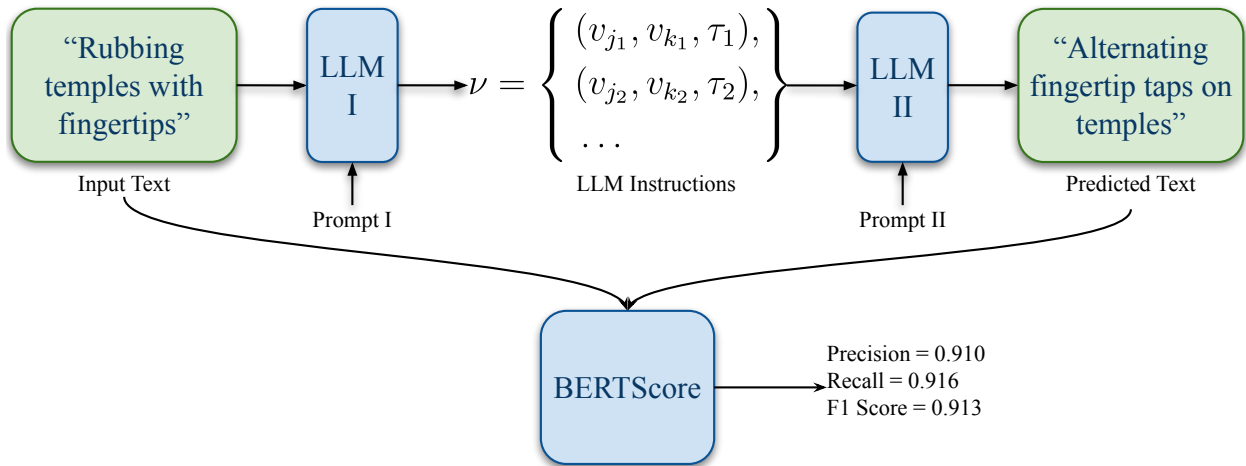


Figure 10. **Quantitative Analysis of Large Language Models (LLM) Outputs:** We employ a pipeline to verify the quality of LLM-generated plans. Specifically, we prompt a second instance of GPT-o3 with only four samples from the interaction dataset to generate a sentence describing the instruction. The second instance is then used in tandem with the first one to perform reverse reasoning. Then for each resulting input–prediction sentence pair, BERT is used to compute precision, recall, and F1 score based on semantic similarity [4].

Index	Input Sentence	Evaluation Sentence	Precision	Recall	F1
1	Interlacing fingers.	Press both palms and matching fingertips together in a prayer gesture.	0.862	0.852	0.857
2	Cracking knuckles.	Slide your left thumb across the right knuckles, then press palms together.	0.842	0.859	0.85
3	Self fist bumping.	Tap the knuckles of your hands together.	0.868	0.852	0.86
4	Rubbing the palms together.	Rub your palms together.	0.954	0.952	0.953
5	Rubbing fingertips together.	Alternately tap your left thumb with your index and middle fingertips.	0.854	0.885	0.869
6	Pressing thumb against palm.	Slide your right thumb across your palm.	0.906	0.908	0.907
7	Rolling a ring around a finger.	Rub your right ring finger with your thumb and middle fingertip.	0.862	0.9	0.881
8	Flicking one finger with another.	Right-hand index taps between thumb and middle finger.	0.876	0.866	0.871
9	Massaging palm with opposite hand.	Rub the left palm with the right thumb and index finger.	0.892	0.902	0.897
10	Tapping fingers against each other.	Bring all corresponding fingertips of both hands together in a prayer-like touch.	0.86	0.875	0.867
11	Snap your fingers to make some noise.	Snap your fingers with your right hand.	0.937	0.916	0.926
12	Pulling fingers one by one with the opposite hand.	Right hand pinches each left fingertip in sequence.	0.884	0.891	0.887
13	Drumming left fingers on the back of the right hand.	Tap your left fingertips on the knuckles of your right hand.	0.919	0.908	0.914
14	Tracing a circle or shape on the palm with a finger.	Run your right index finger across your left palm.	0.889	0.881	0.885
15	Scratching forehead.	Slide your left index finger across your forehead.	0.853	0.851	0.852
16	Rub your back of head.	Scratch the back of your head with the left hand.	0.91	0.922	0.916
17	Tapping cheek with fingers.	Touch your right cheek with your right fingertips.	0.896	0.911	0.903
18	Scratching chin or jawline.	Thoughtful chin stroking.	0.88	0.86	0.87
19	Wiping sweat from forehead.	Wipe your forehead with your right middle knuckle.	0.88	0.922	0.901
20	Resting fingertips against forehead.	Touch your forehead with right fingertips.	0.918	0.921	0.92
21	Gently pinching and releasing fingers together.	Do the OK gesture with the right hand.	0.883	0.854	0.868
22	Fixing hair.	Slide right fingertips from temple over head to neck.	0.845	0.873	0.859
23	Touch you ear.	Gently pinch your left earlobe.	0.881	0.907	0.894
24	Grab your ear.	Gently pinch your right earlobe with your right hand.	0.867	0.924	0.894
25	Cover both ears.	Massage your earlobes with both palms.	0.881	0.912	0.896
26	Pinching own skin.	Pinch your left forearm with your right thumb and index finger.	0.82	0.856	0.837
27	Adjusting glasses.	Right hand pinches nose then earlobe.	0.859	0.862	0.861
28	Rubbing your knee.	Stroke your left knee with your left hand.	0.871	0.904	0.887
29	Rub your left calf.	Rub your left calf with your left hand.	0.945	0.98	0.962
30	Clapping hands twice.	Rub your hands together.	0.921	0.893	0.907
31	Rubbing hands together.	Rubbing palms together.	0.978	0.972	0.975
32	Grab your opposite ear.	Pinch your right earlobe with the left hand.	0.891	0.921	0.905

Index	Input Sentence	Evaluation Sentence	Precision	Recall	F1
33	Put your hands on your hips.	Place your hands on your hips.	0.981	0.977	0.979
34	Both hands grabbing earlobes.	Gently pinch both earlobes with thumb and index fingers.	0.882	0.912	0.897
35	Place your hands on your waists.	Stand with hands on hips.	0.925	0.886	0.905
36	Fixing sleeves or rolling them up.	Slide your right fingertips up your left arm.	0.861	0.86	0.86
37	Hands on knees to rest after running.	Rest both palms on your knees.	0.898	0.882	0.89
38	Clean your armpit with your left hand.	Scratch your left armpit with your left hand.	0.958	0.965	0.961
39	Scretch yourself by touching your toes.	Slide your hands down your legs to touch your toes.	0.898	0.876	0.887
40	Touch your left heel with the left hand.	Rub your left heel with your left hand.	0.971	0.972	0.971
41	Crossing arms and tapping arm with fingers.	Lightly tap both biceps with opposite fingertips.	0.872	0.866	0.869
42	Placing both hands in front of the abdomen.	Rest both hands on your belly.	0.935	0.919	0.927
43	Wrapping fingers around the opposite wrist.	Hold your right wrist with your left hand.	0.887	0.902	0.894
44	Touching the back of the neck with one hand.	Gently massage your neck with the left hand.	0.927	0.919	0.923
45	Secure your head in the moment of emergency.	Scalp massage.	0.873	0.871	0.872
46	Pressing fingers against biceps or shoulders.	Gently scratching left upper arm.	0.888	0.866	0.877
47	Scratching forearm or wrist with opposite fingers.	Brush the left forearm with right fingertips.	0.914	0.891	0.902
48	One heel to the other knee.	Rub your right shin with your left heel.	0.858	0.875	0.866
49	Running fingers across lips.	Tap your lips with right index and middle fingers.	0.867	0.906	0.886
50	Fist bump your hands.	Press your palms together in a prayer gesture.	0.884	0.869	0.877
51	Tracing a circle or shape on the palm with your thumb.	Rub your left palm with your thumb.	0.935	0.9	0.917
52	Smell your armpit.	Smell your right armpit.	0.97	0.984	0.977
53	Scratch your eyes with thumbs.	Gently rub your eyes with both thumbs.	0.914	0.93	0.922
54	Gently pinching bridge of nose.	Pinch your nose with the right hand.	0.917	0.89	0.903
55	Rubbing temples with fingertips.	Alternating fingertip taps on temples.	0.91	0.916	0.913
56	Running ring and middle fingertips across lips.	Slide right middle and ring fingertips across your lips.	0.919	0.942	0.93
57	Cover your mouth with your left hand over surprising news.	Left hand gently touches chin and lips.	0.882	0.868	0.875
58	Kiss your hand palm.	Wipe your lips with the right palm.	0.903	0.909	0.906
59	Send a kiss to your valentine.	Right-hand smoking gesture.	0.882	0.857	0.87
60	Crossing the arms over the chest.	Cross arms over chest in a self-hug.	0.906	0.939	0.922
61	Resting hand on chin.	Hold your chin with the right hand thoughtfully.	0.905	0.914	0.909
62	Stay in attention position.	Stand with both palms on your back thighs, heels pressed together.	0.804	0.83	0.817
63	Clean your right armpit with your left hand.	Scratch your right armpit with your left hand.	0.968	0.973	0.97
64	Tracing a shape on own forearm with the right index finger.	Scratch your left forearm with your right index finger.	0.932	0.911	0.921

Index	Input Sentence	Evaluation Sentence	Precision	Recall	F1
65	Making a 'spider walk' motion with index and middle fingertips on the opposite forearm.	Scratch your left forearm with the right index and middle fingertips.	0.93	0.894	0.911
66	Do a squat.	Do a deep squat, sitting on your heels.	0.899	0.969	0.932
67	Holding one wrist with the opposite hand.	Right hand grasps left wrist.	0.915	0.91	0.912
68	Scratch your eyes.	Rub your right eye with your right fingertips.	0.893	0.924	0.909
69	Blow nose with your right.	Pinch your nose with your right hand.	0.935	0.931	0.933
70	Make delicious gesture.	Blow a kiss with your right hand.	0.859	0.85	0.855
71	Crossing arms.	Cross arms, fingertips resting on opposite biceps.	0.843	0.906	0.873
72	Scratching head.	Gently scratch your head with right fingertips.	0.869	0.863	0.866
73	Holding stomach.	Rest both hands on your belly.	0.871	0.885	0.878
74	Adjusting a tie.	Touch chest with right fingertips.	0.875	0.871	0.873
75	Hugging yourself.	Scratching your shoulder blades with both hands.	0.856	0.904	0.879
76	Gripping shoulder.	Scratch left shoulder with right hand.	0.87	0.874	0.872
77	Tighten your belt.	Gently touch belly and waist with both hands.	0.865	0.866	0.865
78	Salute a commander.	Rub your right forehead with your right index and middle fingers.	0.811	0.849	0.83
79	Button your trousers.	Pinch your lower abdomen using right thumb and index finger.	0.852	0.902	0.876
80	Adjusting a necklace.	Straightening shirt collar with alternating hands.	0.87	0.909	0.889
81	Rub your outer elbow.	Scratch left forearm.	0.904	0.88	0.892
82	Place hand over heart.	Place right hand on chest.	0.942	0.936	0.939
83	Gently touch your back.	Scratch your mid-back with your right hand.	0.91	0.916	0.913
84	Gently scratch your shin.	Gently scratching right shin.	0.952	0.946	0.949
85	Adjusting a tie or necklace.	Lightly tap up and down your sternum with right fingertips.	0.843	0.875	0.859
86	Running fingers along the arm.	Run right fingertips down left arm.	0.909	0.939	0.924
87	Zippering or buttoning a jacket.	Pinched right hand slides upward from belly to chest.	0.853	0.851	0.852
88	Brushing off dust from clothes.	Slide your right fingertips across the chest.	0.874	0.85	0.862
89	Roll up both sleeves one by one.	Gently stroke your opposite arms.	0.893	0.884	0.889
90	Trace a circle around your belly.	Circle your navel with right index and middle fingertips.	0.857	0.917	0.886
91	Touching chin with the left hand.	Rub your chin with left fingertips.	0.94	0.939	0.939
92	Rubbing belly with your left palm.	Left-hand circular belly rub.	0.919	0.906	0.913
93	Clap your belly with your left hand.	Left palm resting on stomach.	0.912	0.879	0.895
94	Resting one hand on the opposite shoulder.	Left hand on right shoulder.	0.938	0.916	0.927
62	Stay in attention position.	Stand with both palms on your back thighs, heels pressed together.	0.804	0.83	0.817
33	Put your hands on your hips.	Place your hands on your hips.	0.981	0.977	0.979

Table 3. **BERTScores for input–predicted sentence pairs:** Pairs with lowest and the highest scores shown in the last two rows. As seen in the figure the semantic similarity between input and predicted text pairs are high, indicating the effectiveness of the high-level LLM planning.

References

- [1] Enes Duran, Muhammed Kocabas, Vasileios Choutas, Zicong Fan, and Michael J. Black. Hmp: Hand motion priors for pose and shape estimation from video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6353–6363, 2024. 1
- [2] OpenAI. Introducing o3 and o4-mini, 2024. Accessed: 2025-05-16. 1
- [3] Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. Tlcontrol: Trajectory and & language control for human motion synthesis. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXXVII*, page 37–54, 2024. 3
- [4] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *ICLR*, 2020. 2, 7