SocialMirror: Reconstructing 3D Human Interaction Behaviors from Monocular Videos with Semantic and Geometric Guidance

Supplementary Material

We present additional implementation details, including model setup, dataset processing, diffusion process modification, and the two-branch network architecture, as well as VLM annotator details with prompting examples in Sec. 1. Additional experiments are provided in Sec. 2, including an ablation on the motion embedding layer in the Geometry Optimizer, performance breakdowns across Hi4D's action categories, cross-dataset results, and in-the-wild visualizations. Sec. 3 includes analyses of VLM limitations and failure cases, while also exploring the role of semantic information in limited-contact scenarios and outlining the framework's current limitations. The use of Large Language Models are declared in Sec.4

1. Additional Details

1.1. Implementation details

Our model was implemented using PyTorch and trained on an NVIDIA RTX 3090 GPU. The batch size was set to 32 for the Semantic-Guided Motion Infiller and 64 for the Geometry Optimizer. We employed the AdamW optimizer with CyclicLRWithRestarts, where the learning rate was initially set to 0.0001, with parameters restart_period=10, t_mult=2, and a "cosine" policy.

In the Motion Infiller and Motion Refiner, the dimension of human motion followed CloseInt [2] with D = 157. For the Geometry Optimizer, we utilized 24 SMPL joints to represent human motion, resulting in a human motion dimension of D' = 24 × 3. The text feature dimension F_{text}, encoded from CLIP [4], was 256.

For dataset implementation, original long motion sequences were divided into shorter clips with a length of L = 16 frames. Each clip was annotated with a corresponding text description using our LLM annotation module. For 3DPW, we established a new benchmark by selecting sequences involving two subjects: sequences captured in courtyard environments were used for training, and those captured in downtown settings were used for testing.

For multi-person scenes, we automatically detect and track individuals to obtain their bounding boxes and select the pair with the closest spatial proximity as the primary subjects. The original image is then cropped according to their bounding boxes, centering the region of interest to minimize background distractions and ensure the VLM focuses exclusively on the targets.

1.2. Diffusion with initial distributions

In prior approaches to diffusion-based pose estimation [1, 5], time-dependent Gaussian noise sampled from $\mathcal{N}(0, \mathrm{I})$ is incrementally injected into ground-truth motion sequences $\hat{x}_0$ through the forward process:

$$q(\mathbf{x}_t \mid \hat{\mathbf{x}}_0) = \sqrt{\hat{\alpha}_t}\hat{\mathbf{x}}_0 + \sqrt{1-\hat{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}(0,1) \quad (1)$$

where α_t denotes a constant hyper-parameter [3], and $\hat{\alpha}_t = \prod_{i=0}^{t} \alpha_i$. It was observed that x_t follows a standard Gaussian distribution, and the early iterative steps provide limited meaningful information for human motion. Additionally, the results should fully account for the initial prediction consistent with image characteristics.

To address these issues, we propose modifying the forward diffusion process to align with the initial distributions:

$$q(x_t|\hat{x}_0) = x + \sqrt{\hat{\alpha}_t}(\hat{x}_0 - x) + \sqrt{1-\hat{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma) \quad (2)$$

With this adjusted framework, a generative model is derived by reversing the diffusion process, starting from samples $x_t \sim \mathcal{N}(x, \sigma)$. The reverse process is defined as:

$$q(x_{t-1}|x_t, c) = \mathcal{N}\left(x_{t-1}; \mu_\alpha(x_t, c), \tilde{\beta}_t\sigma\right) \quad (3)$$

where $\mu_\alpha(x_t, c)$ represents the estimated mean from the diffusion model under condition c at timestep $t-1$, and $\tilde{\beta}_t$ denotes the variance calculated using the hyperparameters β_t, $\hat{\alpha}_t$, and $\hat{\alpha}_{t-1}$.

1.3. Model details

We employ a two-branch network architecture to model human interactions, where each branch processes the actions of one individual and information sharing occurs between the branches. Specifically, x_a^t and x_b^t are first processed through a motion embedding layer and sequence position encoding to generate initial hidden states h_a^0 and h_b^0. These states are then fed into a two-branch transformer network with shared weights, composed of N transformer blocks. Within each block, self-attention (SA) and cross-attention (CA) mechanisms enable intra-agent and inter-agent information exchange, respectively. For the n-th transformer block in agent a's branch where $n \in [1, N]$:

The Self-Attention Block processes its own hidden state h_a^{n-1} to capture intra-agent dependencies. The query Q_a^{SA}, key K_a^{SA}, and value V_a^{SA} matrices are derived from h_a^{n-1} as:

$$Q_a^{\mathrm{SA}} = h_a^{n-1}W_Q^{\mathrm{SA}}, \quad K_a^{\mathrm{SA}} = h_a^{n-1}W_K^{\mathrm{SA}}, \quad V_a^{\mathrm{SA}} = h_a^{n-1}W_V^{\mathrm{SA}} \quad (4)$$

where $W_Q^{\text{SA}}, W_K^{\text{SA}}, W_V^{\text{SA}}$ are trainable weights. The self-attention output is calculated as:

$$\text{SA}(h_a^{n-1}) = \text{Softmax}\left(\frac{Q_a^{\text{SA}}(K_a^{\text{SA}})^T}{\sqrt{C}}\right)V_a^{\text{SA}} \quad (5)$$

where C is the number of channels in the attention layer. Then a Cross-Attention Block facilitates inter-agent information exchange. For agent a, the query matrix Q_a^{CA} is derived from h_a^{n-1}, while the key K_a^{CA} and value V_a^{CA} matrices come from h_b^{n-1}:

$$Q_a^{\text{CA}} = h_a^{n-1}W_Q^{\text{CA}}, \quad K_a^{\text{CA}} = h_b^{n-1}W_K^{\text{CA}}, \quad V_a^{\text{CA}} = h_b^{n-1}W_V^{\text{CA}} \quad (6)$$

The cross-attention output for agent a is:

$$\text{CA}(h_a^{n-1}, h_b^{n-1}) = \text{Softmax}\left(\frac{Q_a^{\text{CA}}(K_a^{\text{CA}})^T}{\sqrt{C}}\right)V_a^{\text{CA}} \quad (7)$$

A symmetric calculation for agent b, $\text{SA}(h_b^{n-1}), \text{CA}(h_b^{n-1}, h_a^{n-1})$, swaps the roles of h_a^{n-1} and h_b^{n-1}.The weight matrices $W_Q^{\text{SA}}, W_K^{\text{SA}}, W_V^{\text{SA}}$ and $W_Q^{\text{CA}}, W_K^{\text{CA}}, W_V^{\text{CA}}$ are shared across both branches. At the end of each block, the outputs of the SA and CA blocks are combined with residual connections and layer normalization, for agent a:

$$h_a^n = \text{LayerNorm}\left(h_a^{n-1} + \text{SA}(h_a^{n-1}) + \text{CA}(h_a^{n-1}, h_b^{n-1})\right) \quad (8)$$

This integrated hidden state h_a^n is then fed into subsequent transformer layers. The weight-sharing symmetry ensures balanced processing of inter-agent interactions, reducing model parameters while improving generalization capabilities.

ControlNet is a trainable copy of the N transformer blocks of the diffusion model, they share common inputs: h_a^0, h_b^0, t, and F_{img}. Additionally, it incorporates text features F_{text} encoded by CLIP. For each trained transformer block, the computation is defined as: $h_i = \mathcal{T}(h^{i-1}, F_{\text{img}}; \Theta)$, where Θ denotes the frozen parameters of the block.

The trainable copy of the model connects to the original model via zero linear layers. The output of the controlled diffusion network is therefore:

$$h_i^c = \mathcal{T}(h^{i-1}, F_{\text{img}}; \Theta) + \mathcal{Z}\left(\mathcal{T}\left(x + \mathcal{Z}(F_{\text{text}}; \Theta_{z1}), F_{\text{img}}; \Theta_c\right); \Theta_{z2}\right) \quad (9)$$

Here $\mathcal{T}$ represents the original model block and $\mathcal{Z}$ denotes the zero linear layers. At the start of training, the zero layers output zeros, so $h_i^c = \mathcal{T}(h_{i-1}^c; \Theta)$ matches the base model. As training proceeds, the zero layers gradually inject conditional signals.

1.4. VLM annotator details

We further provide the details of VLM Annotation in Tab. 1. We also provide several generated textual descriptions and contact pairs in Fig.CameraReady 1; the text is well aligned with the images and supplies semantic guidance for human mesh reconstruction.

2. Additional Experiments

2.1. Ablation on Geometry Optimizer

Geometry Optimizer focuses on processing 3D joint positions to provide geometric guidance information. To validate the effectiveness of our encoding layer design for the auxiliary model, we conducted an ablation study by implementing the motion embedding layer with either STGCN or a Linear layer. The results are presented in Tab. 2.

The Geometry Optimizer that uses STGCN to encode 3D joint positions exhibits higher accuracy than the one using Linear. It successfully captures the 3D positional relationships of interacting humans and outperforms Motion Infiller in all metrics related solely to 3D joint positions. This indicates that it can effectively provide correct guidance information.

2.2. Additional experiments results on Hi4D

We further partition Hi4D into subsets by action label to assess performance across interaction categories. Tab. 3 presents our method's improvements over CloseInt across different subsets. Notably, our approach achieves the most significant gains on actions such as handshake, high-five, and kiss. In these actions, human behavioral patterns are relatively uniform, and occlusion levels are moderate. The model synthesizes plausible poses by integrating textual descriptions generated by VLM Annotator, while simultaneously mitigating mesh interpenetration issues and refining contact relationships. However, the method shows smaller gains on complex actions such as dancing and fighting. These activities involve intricate limb interactions and ambiguous joint-depth relationships, which can slightly undermine VLM annotation consistency and the precision of geometric guidance. Nonetheless, our method still outperforms the baseline.

2.3. Cross dataset evaluation

We also report both intra-domain and cross-domain results. SocialMirror outperforms prior methods in all settings. In our experiments, we observed that when not trained on the dataset, CloseInt may erroneously separate characters that should be in close contact. This results in the absence of even minor intended penetrations (e.g., slight mesh intersection between a palm and another person), leading to a relatively low penetration error—though this is not indicative of a good reconstruction outcome. After training on the

Table 1. Detailed prompting example for VLM Annotator.

Prompting Example
Given the image sequence of two human interaction, generate 0, 1 or more joint-joint contact pair(s) according to the following background information, rules, and examples. Joint-joint contact pair should exactly reflect the human interaction shown in the image sequence. [Start of background Information] Human has JOINTS: [‘pelvis’, ‘left_hip’, ‘right_hip’, ‘left_knee’, ‘right_knee’, ‘left_ankle’, ‘right_ankle’, ‘left_foot’, ‘right_foot’, ‘neck’, ‘left_collar’, ‘right_collar’, ‘head’, ‘left_shoulder’, ‘right_shoulder’, ‘left_elbow’, ‘right_elbow’, ‘left_wrist’, ‘right_wrist’]. [End of background Information] [Start of rules] 1.Each joint-joint pair should be formatted into {JOINT, JOINT, TIME-STEP, TIME-STEP}. JOINT should be replaced by JOINT in the background information. IMPORTANT: The first JOINT belongs to person 1, and the second JOINT belongs to person 2. Each joint-joint pair represents a contact of a joint of person 1 and a joint of person 2. The first TIME-STEP is the start frame number of contact, and the second TIME-STEP is the end frame number of contact. 2.Use one sentence to describe what action person 1 do and one sentence to describe what action person 2 do according to the image sequence. IMPORTANT: the sentence starts from ‘text 1:’ describing the action of person 1 from the perspective of person 1 and the sentence starts from ‘text 2:’ describing the action of person 2 from the perspective of person 2. Sentences should NOT contain words like ‘person 1’ or ‘person 2’, use ‘a person’ to refer to himself in the sentence and ‘others’ to refer to others. IMPORTANT: the sentence should be align with the joint-joint contact pair. IMPORTANT: the order of person 1 and person 2 should be the same in different joint-joint contact pair of the same image sequence. 3.IMPORTANT: Do NOT add explanations for the joint-joint contact pair. [End of rules] [Start of an example] [Start of sentences] Text 1: a person dance with others holding his left hand with the other’s right hand, putting his right hand on the other’s waist, and his shoulder being touched. Text 2: a person dance with other holding her right hand with the other’s left hand, with her waist being embraced, placing her left hand on the other’s shoulder. [End of sentences] [Start of joint-joint contact pair(s)] {left_wrist, right_wrist, 11, 15} {right_wrist, left_hip, 14, 15} {right_shoulder, left_wrist, 9, 15} [End of joint-joint contact pair(s)] [End of an example]

Table 2. Ablation studies on the impact of motion embedding layer.

Embedding Layer	↓R-MPJPE	↓G-MPJPE	↓Int	↓MPJPE	↓PA-MPJPE
Linear	102.3	110.0	84.9	81.9	66.8
STGCN	**81.7**	**93.2**	**62.5**	**60.8**	**47.8**

dataset, CloseInt's errors in character placement are reduced, but it correspondingly exhibits more interpenetration, which explains why the penetration loss increases post-training. Our method, in both scenarios, produces more accurate relative positions of characters (as reflected in RE and Int) while ensuring less interpenetration, demonstrating the positive effect of the proposed method in reducing interpenetration.

2.4. Results on Harmony4D

For completeness, we also conducted training experiments on the Harmony4D dataset, which further confirms the effectiveness of our approach. Specifically, our method achieves significant improvements in interaction-related metrics: it yields decreases of 8.2%, 3.5%, and 3.2% in RE, GE, and Int, respectively. Meanwhile, it maintains nearly unchanged performance on single-person reconstruction metrics (i.e., MPJPE, PA, and VPE). This result demonstrates the robust capability of our method in capturing human interaction relationships.

2.5. Additional visualization results

We present additional in-the-wild reconstructions in Fig. 1; the supplementary video includes further comparisons and demonstrations.

3. Discussions

3.1. Reconstruction under VLM Limitations

Based on our user study, the text descriptions generated by the VLM are, on average, superior to those produced by human annotators. As shown in Fig.1, VLM annotations can capture not only macroscopic actions but also fine-grained contact relationships between specific joints (e.g., “A person leads the dance, extending his left arm to hold the other's

Table 3. Comparison of CloseInt and our method, CloseInt/Ours (Improvement), across different actions on Hi4D.

Action	handshake	high-five	kiss	dance	fight
↓R-MPJPE	78.0/65.8 (20.0)	60.5/53.5 (19.2)	81.7/67.7 (20.9)	96.4/87.6 (9.4)	110.3/100.3 (8.2)
↓G-MPJPE	93.2/72.5 (23.8)	84.9/70.9 (19.7)	98.9/79.7 (19.6)	109.2/97.9 (9.4)	131.4/120.1 (6.5)
↓Int	36.9/31.1 (15.6)	26.7/25.1 (5.9)	33.3/23.0 (63.9)	39.9/32.4 (18.8)	46.7/41.1 (11.9)
↓Pen	194.8/71.9 (63.1)	107.4/50.1 (53.3)	15409.3/5570.1 (63.9)	5477.1/2455.0 (55.2)	636.6/226.2 (64.5)

Table 4. Cross Dataset Evaluation on Hi4D and 3DPW.

Method	Hi4D							3DPW						
	↓RE	↓GE	↓Int	↓Pen	↓MPJPE	↓PA.	↓VPE	↓RE	↓GE	↓Int	↓Pen	↓MPJPE	↓PA.	↓VPE
CloseInt	99.0	114.9	81.4	3947.6	63.1	**47.5**	**76.4**	135.7	**159.1**	95.5	342.7	79.9	52.9	95.1
Ours	**83.6**	**95.2**	**68.5**	**2380.5**	**62.2**	**47.5**	79.3	**104.8**	162.7	**89.9**	**109.7**	**65.1**	**49.0**	**79.7**
CloseInt (Eval. Only)	181.1	232.1	182.7	**1973.8**	109.1	**62.5**	132.0	194.4	340.2	128.4	**101.6**	88.6	63.6	110.7
Ours (Eval. Only)	**165.2**	**184.1**	**153.0**	2380.3	**105.2**	63.6	**129.4**	**174.7**	**307.4**	**125.4**	109.7	**87.5**	**63.3**	**109.8**

Table 5. Comparisons on Harmony4D.

Method	↓RE	↓GE	↓Int	↓Pen	↓MPJPE	↓PA.	↓VPE
CloseInt	134.8	297.5	182.5	482.6	70.2	38.6	82.6
Ours	**123.8**	**287.2**	**176.8**	**480.3**	**69.8**	**39.7**	**80.8**

right hand and guiding her movements with his right hand on her back"), whereas a human annotator might simply describe it as "two people dancing ballroom dance. While VLM Annotator demonstrates satisfactory performance in describing human interaction under most circumstances, its accuracy tends to decline when confronted with complex limb interactions, affecting the precision of both textual descriptions and contact pair annotations. By prioritizing visual feature extraction over textual inputs, our proposed method maintains reconstruction fidelity even when text-image alignment is compromised. As illustrated in Fig.2, despite VLM Annotator's failure to correctly identify the human action, our approach successfully reconstructs accurate motion patterns by leveraging visual information.

3.2. Failure cases

Our approach remains limited under prolonged, severe occlusions. Fig. 3 shows a case where both visual and semantic cues are unreliable. Inaccurate text and contact predictions from the VLM annotator then propagate misleading guidance and large reconstruction errors. This observation underscores the necessity of complementary mechanisms to handle extreme occlusion scenarios in future work.

3.3. The effect of semantic information on limited contact scenarios

Even when contact is absent, the VLM can still produce high-level scene descriptions (e.g., two people stand and face each other), which are encoded as semantic features. These provide contextual cues about interaction and spatial layout beyond direct contact information. In addition, our approach does not rely solely on contact labels. The semantic features guide the Motion Infiller to infer plausible poses for ambiguous regions, and the Temporal Motion Refiner and geometric constraints based on 3D joint prediction from the Auxiliary Module ensure motion smoothness and spatial plausibility. Table 3 further shows gains in interaction metrics even for actions with mild occlusion and limited contact.

3.4. Limitations and future works

Our current pipeline targets two-person interaction. For reconstructing interactions involving more participants, further improvements to the network architecture and annotation protocols are required.

Improving the reliability of semantic guidance is another important direction for future work. Promising steps include estimating confidence from the VLM annotator, adaptively reweighting text conditioning when captions are uncertain, and explicitly checking semantic-visual agreement before feeding language into reconstruction.

4. The Use of Large Language Models (LLMs)

We declare that vision-language models (VLMs) in this paper are used primarily as a VLM Annotator to produce textual descriptions of interactions in image sequences and spatio-temporal joint contact pairs. LLMs are used only for light text polishing and grammar fixes. The research approach, core ideas, reasoning, and conclusions remain the authors' own work. All VLM/LLM-assisted content generation is documented together with how and where it was applied.

Text 1: A person leads the dance, extending his left arm to hold the other's right hand.
Text 2: A person follows the lead, extending her right arm to hold the other's left hand.

{left_wrist, right_wrist, 0, 5}

Text 1: A person extends his right hand to hold the other's left hand and moves his left arm upwards, guiding others in a dance.
Text 2: A person holds her left hand with the other's right hand and raises her right arm, following the lead in the dance.

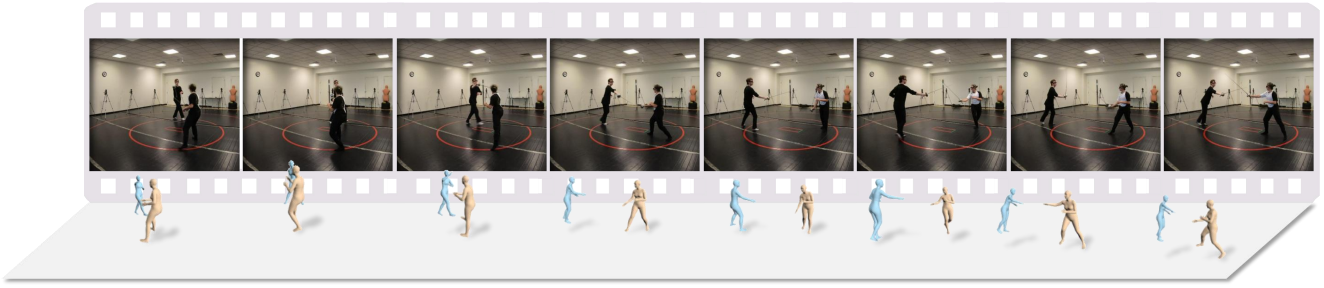
{right_wrist, left_wrist, 0, 5}
{left_elbow, right_shoulder, 3, 5}

Text 1: A person leads the dance, extending his left arm to hold the other's right hand and guiding her movements with his right hand on her back.
Text 2: A person follows the lead, holding the other's left hand with her right hand and feeling his guidance on her back as she moves gracefully.

{left_wrist, right_wrist, 0, 4}

Text 1: A person extends their right hand to hold the other's left hand while dancing.
Text 2: A person extends their left hand to hold the other's right hand while dancing.

{right_wrist, left_wrist, 0, 16}



Text 1: A person walks towards others, passing by them without any physical contact.
Text 2: A person stands still as others walk past without touching.

Text 1: A person moves towards others, extending his right hand to touch the other's left shoulder.
Text 2: A person receives a touch on her left shoulder from the other's right hand while moving away.

{right_wrist, left_shoulder, 5, 8}

Text 1: A person walks towards others, extending their right hand forward.
Text 2: A person stands still as others approach and extend their left hand towards A person's right hand.

Text 1: A person moves his right hand towards others.
Text 2: A person extends her left hand towards the other's right hand.

Text 1: A person strikes with his right hand, making contact with the other's left hand.
Text 2: A person blocks the strike with his left hand, keeping his right hand ready for defense.

{right_wrist, left_wrist, 1, 4}

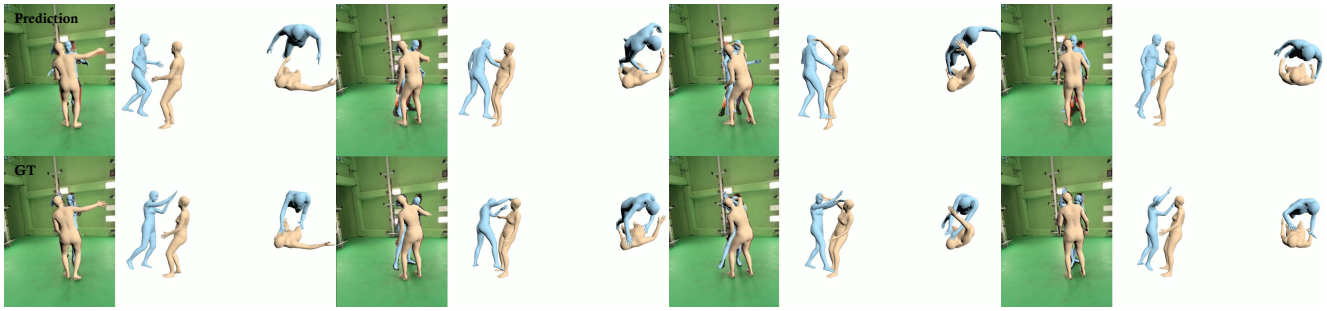
Text 1: A person practices martial arts with others, keeping his hands ready for defense.
Text 2: A person practices martial arts with others, leaning slightly forward.

Text 1: A person swings his left arm and makes contact with the other's right shoulder while keeping his right hand near his own waist.
Text 2: A person stands facing others, receiving a hit on her right shoulder from the other's left arm while maintaining her stance.

{left_elbow, right_shoulder, 5, 8}

Text 1: A person practices martial arts with others, keeping his left hand raised.
Text 2: A person practices martial arts with others, keeping his right arm up for defense.

Figure 1. Visualization results on in-the-wild data.



Text 1: A person holds the other's left hand with their right hand while moving their right arm outwards.
Text 2: A person holds the other's right hand with their left hand while moving their left arm outwards.

Text 1: A person holds the other's left hand with their right hand.
Text 2: A person holds the other's right hand with their left hand.

{RElbow, LWrist, 7, 11}

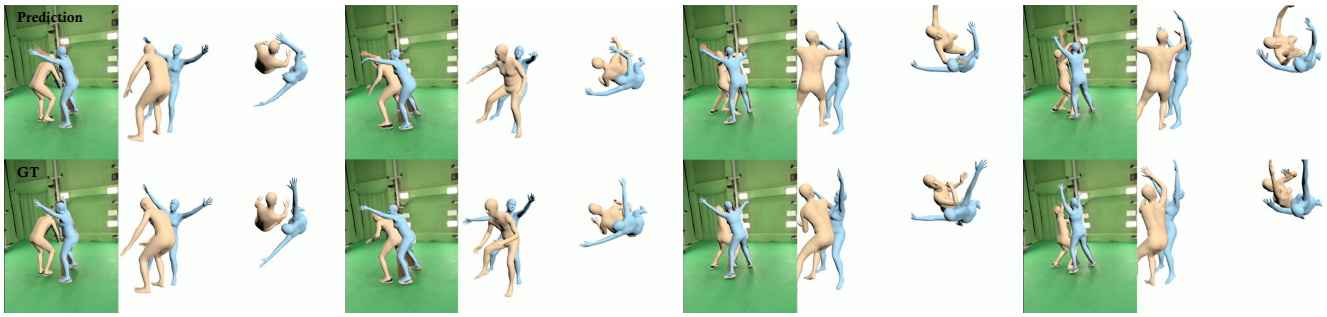
Text 1: A person holds the other's left wrist with their right hand.
Text 2: A person holds the other's left wrist with their right hand.

{RShoulder, LWrist, 5, 5}

Text 1: A person is being lifted off the ground by another person.
Text 2: A person is lifting another person off the ground.

{RSmallToe, LAnkle, 1, 2}
{RSmallToe, LHeel, 1, 3}

Figure 2. VLM Annotator failed to describe human interaction.



Text 1: A person extends their arms outwards while bending slightly forward.
Text 2: A person raises their arms above their head while leaning back slightly.

Text 1: A person is leaning forward with their arm extended.
Text 2: A person is leaning backward with their arm extended.

Text 1: A person leans forward with their arms extended, appearing to push against another person.
Text 2: A person leans backward, supporting themselves with one arm while the other arm is extended outward.

Text 1: A person is dancing with another person, extending their arm forward.
Text 2: A person is dancing with another person, leaning back slightly.

Figure 3. Challenging case with prolonged, severe occlusions.

References

[1] Runyang Feng, Yixing Gao, Tze Ho Elden Tse, Xueqing Ma, and Hyung Jin Chang. Diffpose: Spatiotemporal diffusion model for video-based human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14861–14872, 2023. 1

[2] Buzhen Huang, Chen Li, Chongyang Xu, Liang Pan, Yangang Wang, and Gim Hee Lee. Closely interactive human reconstruction with proxemics and physics-guided adaption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1011–1021, 2024. 1

[3] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 1

[4] 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

[5] Cédric Rommel, Eduardo Valle, Mickaël Chen, Souhaiel Khalfaoui, Renaud Marlet, Matthieu Cord, and Patrick Pérez. Diffhpe: Robust, coherent 3d human pose lifting with diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3220–3229, 2023. 1