Ask, Pose, Unite: Scaling Data Acquisition for Close Interaction Meshes with Vision Language Models

Supplementary Material

In this supplementary material we provide details on our prompting strategy (Sec. A), additional qualitative examples (Sec. B), limitations and ethical concerns (Sec. C) and, information on the data sources we used to curate our dataset (Sec. D).

A. LVLM Prompts

Figure S7. In-context example from the TV Interactions dataset provided with the prompt.

To query the images in our dataset we use the prompt in Fig. S9. However, for the images that have interaction types we modify the prompt such that the instructions include the name of the interaction action (see Fig. S10).

For both cases we design the prompt such that the LVLM can make use of intermediate tasks to find the contacts. In particular, we: (1) provide an in-context example of two people hugging (Figure S7) for which we detail the expected output; (2) request a description of the type of interaction and people involved in it. Even though we crop out the image to the minimum bounding containing both people of interest, there are instances with other subjects present. (3) The orientation of the people w.r.t one another, which in our experiments improves the labeling of the body parts in terms of chirality and recall.

B. Additional qualitative results.

In section 4.2 we discussed how our contact prior successfully integrates the improved 3D meshes from our contact maps (Auto CM) (see Fig. 5) on close interactions from the NTU+RGBD 120 test set. In Fig. S8 we show additional qualitative examples. The contact maps as a stronger supervision of the interaction can enforce contact in cases that the prior does not (see the top row). However, mistakes in

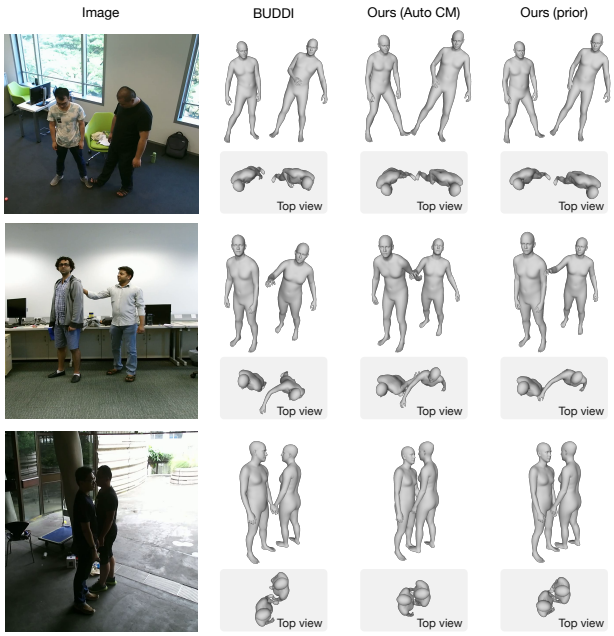

Figure S8. Examples of posed meshes on the close interactions NTU+RGBD 120 test set with BUDDI, Ours Auto CM, and Ours contact prior. *Top row* the contact maps (Auto CM) can enforce contacts when the prior-based methods do not (Ours prior and BUDDI). *Middle row* mistakes in the contact maps lead to incorrect reconstructions. *Bottom row* Occluded and dark scenes are challenging for all methods.

the contact maps can lead to incorrect reconstructions (see middle row). These errors can occur due to intrinsic biases from the LVLM, not specifically training the LVLM for the task, or challenging scenarios like occlusions and bad lighting (see bottom row). However, we note how in these cases a trained model like the contact prior can be robust to these mistakes.

C. Limitations & Ethical Concerns.

Close interactions in HME is an ongoing line of research. Our automatic data generation method filters out many close interaction images even if they appear suitable, yet if the 2D keypoints, initial mesh estimation, or automatic contact maps are not all accurate, the images can be excluded. As improvements in the models that produce each of these components are made, the diversity of interactions will increase.

To obtain the contacts for a pair of people with our

Table 4. Data sources for our APU dataset. Candidate pairs: possible people in contact from 2D keypoint distances. Final pairs: number of mesh pairs after automatic filtering.

Data source	Images	Subjects	Actions	# Actions	# Subjects	Candidate pairs	Final pairs
TV Interactions [40]	8445	adults	✓	4	≥ 2	5970	1679
Human Interaction Images [53]	1177	all ages	✓	7	≥ 2	954	282
Relative Human [51]	8740	all ages	×	in-the-wild	≥ 2	10577	2725
NTU RGB+D 120 train [32]	3000	adults	✓	11	2	2347	1523

method we query once per image using a single LVLM, GPT-4V [1]. This approach may replicate the existing biases in the LVLM. Producing multiple outputs per image with a higher temperature and/or querying multiple LVLMs could provide a measure of uncertainty to the predicted contacts, which could be easily integrated into our soft contact maps to improve robustness in the predictions. We do not foresee significant risks of security threats or human rights violations in our work. However, the advancements in close interactions HME could be misused for creating misleading visual content, leading to potential harm or deception.

D. Data source details

In this section we provide more information on the dataset generated with our method. We use an abridged version of the dataset datasheet format (some questions have been removed for conciseness and to preserve anonymity).

D.1. Motivation

For what purpose was the dataset created? We created the dataset from our data generation method to diversify the paired image and mesh available for closely interacting humans.

D.2. Composition

What do the instances that comprise the dataset represent? The basic data element is an image of a pair of people. This image can be complete or a portion of a larger image. For each pair of people we provide their posed meshes in SMPL-XA format, their keypoints and bounding boxes predicted by VitPose and Openpose, and the LVLM's output which includes the interaction type, description of the people and list of body parts in contact.

How many instances are there in total? 6209 instances of pairs of people interacting sourced from in-the-wild and laboratory images.

Does the dataset contain all possible instances? We source the images for the dataset from 4 existing datasets (Tab. 4): TV interactions, Human interaction images, Relative human, and close interaction classes from NTU RGB+D 120. Each image may contain from 0 to multiple pairs of people interacting, we provide the instances with reconstructions that have a keypoint reprojection error less than 20.0.

For the NTU RGB+D 120 train set we randomly selected 3000 initial from a complete set of 3 frames per sequence that could contain people in contact. For the test set where we included a keyframe from each sequence where the subjects were in contact. We manually inspected all images from the final test set only.

What data does each instance consist of? The raw data are the images from each data source.

Is there a label or target associated with each instance? For each pair of people we provide the reconstructed meshes from our data generation method.

Are there recommended data splits (e.g., training, development/validation, testing)? We use all pairs for training except those from the NTU RGB+D test set.

Are there any errors, sources of noise, or redundancies in the dataset? As we generate pseudo-ground truth meshes these and the contacts and the keypoints may not correspond exactly to what is depicted in the image. We noticed that some images from Relative Human are duplicated under different names.

Is the dataset self-contained, or does it link to or otherwise rely on external resources? We provide the links to download the original images and annotations from the source datasets. All other products are self-contained.

Does the dataset contain data that might be considered confidential? The images depict people and their faces which makes the data identifiable but all are within the public domain.

Does the dataset identify any subpopulations? No. We only specify that the data contains all ages instead of only adults.

D.3. Collection Process

How was the data associated with each instance acquired? The source images were directly observable and the products of our dataset were derived from our data generation method. We validated the quality of the posed meshes with a threshold on the 2D keypoint reprojection error. For the NTU RGB+D 120 test set we manually verified the quality of the 3D ground truth joints from the original annotations.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual

```
Overview:
- You are participating in an image annotation project.
Your task is to annotate images where two people are interacting, specifically identifying
where their bodies touch.
Example:
- The first image is an example.
{"interaction": "hugging",
 "people": {"person_left": "woman wearing red hugging a child",
            "person_right": "child in a plaid shirt hugging a woman"},
 "orientation": "front to front",
 "contacts": [
    {"body_part_person_left":
        {"part_name: "upper arm", "body_side": "right"},
    "body_part_person_right":
        {"part_name: "forearm", "body_side": "left"},
    "confidence": 0.8},
    {"body_part_person_left":
        {"part_name: "hand", "body_side": "left"},
    "body_part_person_right":
        {"part_name: "back", "body_side": "right"},
    "confidence": 0.7}]
}
Instructions:
1. Examine the second image carefully.
2. Annotate each point where body parts from the two individuals make contact.
3. For each annotation, clearly specify:
    - Indicate which person (e.g., person on the left, person on the
    right) the body part belongs to.
    - The body part involved for each person and body side (either right
    or left or both)
    - The confidence level of that the contact is happening (0.0 - 1.0)

Output Requirements:
- Provide annotations in the following format:
{"interaction": "type of interaction",
 "people": {"person_left": "description of the person on the left",
            "person_right": "description of the person on the right"},
 "orientation": "orientation of the people (e.g., front to front, back to front, back to
 back, side to side)",
 "contacts": [
    {"body_part_person_left":
            {"part_name: "...", "body_side": "..."},
        "body_part_person_right":
            {"part_name: "...", "body_side": "..."},
        "confidence": 0.0 - 1.0},
    // More annotations here ]
}
- Use only this list of body part name: {body_parts}
Note:
- Aim for comprehensive coverage of all contact points, even those that might appear
minimal.
```

Figure S9. Example of the complete LVLM prompt.

human curation, software program, software API)? The source images were collected from existing image datasets and the products of our dataset were derived from our data generation method.

If the dataset is a sample from a larger set, what was the sampling strategy? The pairs of interacting people are a subsample of all existing pairs in the images. Our strategy was to use our data generation method to select the pairs in contact with valid posed meshes. For more details see Tab. 4.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data

```
Instructions:
1. Examine the second image of two people performing the action {action}
carefully.
2. Annotate each point where body parts from the two individuals
make contact.
3. For each annotation, clearly specify:
  - Indicate which person (e.g., person on the left, person on the right) the body part belongs to.
  - The body part involved for each person and body side (either right or left or both)
  - The confidence level of that the contact is happening (0.0 - 1.0)
```

Figure S10. Example of the modifications to the LVLM prompt when there is an annotation of the action depicted in the image.

associated with the instances (e.g., recent crawl of old news articles)? This dataset was collected in 2024, the original images correspond to publications from 2012 (TV Interactions), 2016 (Human Interaction Images), 2016/2019 (NTU RGB+D 120), and 2019/2022 (Relative Human).

D.4. Uses

Has the dataset been used for any tasks already? In the paper we show how the data can be used to train a contact prior for Human Mesh Estimation.

What (other) tasks could the dataset be used for? We used the data from a 3D application, but it can also be used for 2D tasks like image generation and general 2D understanding of person-to-person interactions.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? The dataset was not compiled to have an equal ethnic or demographic distribution, as such, downstream tasks should be aware of the possible sampling biases in the data.

Are there tasks for which the dataset should not be used? The dataset focuses on broadly on human interactions. It should not be used to generate any explicit or harmful content from the subjects in the images or any other subjects.

D.5. Distribution

How will the dataset will be distributed Through the project website. In the code we will detail the process for accessing the data, including a form where users agree to the license and terms of use. Users must apply separately for access to the NTU RGB+D 120 subset of the dataset through that dataset's webpage: https://rose1.ntu.edu.sg/dataset/actionRecognition.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? We will distribute the data with a CC BY-NC 4.0 license after filling a form where they agree to the license and terms of use. Users must apply separately for access to the NTU RGB+D 120 subset of the dataset.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? The data from the NTU RGB+D 120 dataset have their own restrictions including redistribution, derivation or generation of a new dataset without permission and commercial usage.