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

Anonymous CVPR submission

Paper ID ####

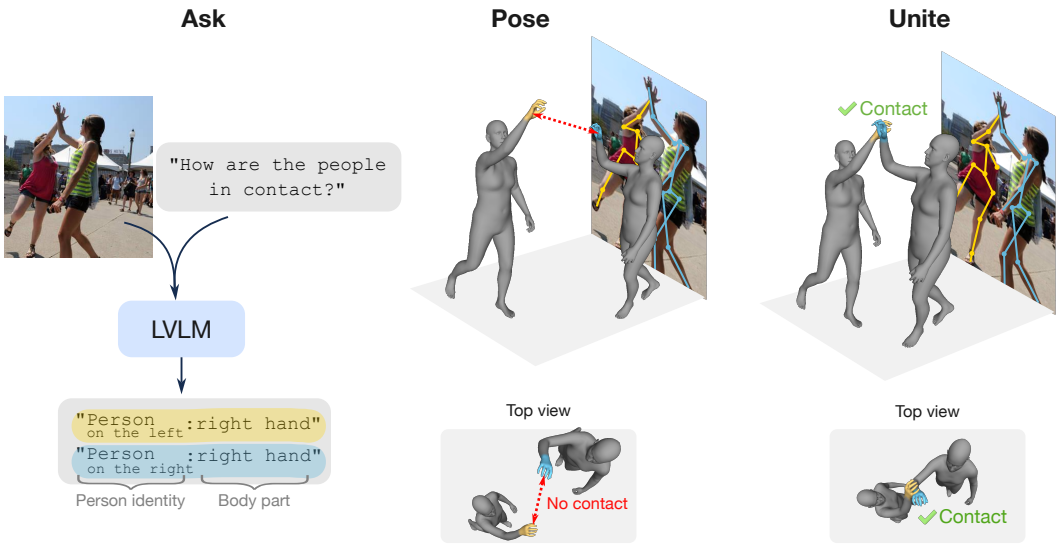


Figure 1. **Ask, Pose, Unite.** We scale data acquisition for close interactions by *Ask*ing a Large Vision Language Model (LVLM) to identify contact points between people via language descriptions of the body parts that are touching. We *Pose* 3D meshes in the scene with predicted 2D keypoints and *Unite* the meshes in 3D by constraining an optimization of the mesh parameters with the predicted contacts. Through our APU data generation method we curate a Human Mesh Estimation dataset for close interactions.

## Abstract

*Social dynamics in close human interactions pose significant challenges for Human Mesh Estimation (HME), particularly due to the complexity of physical contacts and the scarcity of training data. Addressing these challenges, we introduce a novel data generation method Ask, Pose, Unite (APU) which utilizes Large Vision Language Models (LVLMs) to annotate contact maps to guide test-time optimization. APU produces paired image and pseudo-ground truth meshes from monocular images. Our method not only alleviates the annotation burden but also enables the assembly of a comprehensive dataset specifically tailored for close interactions in HME. Our dataset, comprising over 6.2k human mesh pairs in contact covering diverse interaction types, is curated from images depicting naturalistic person-to-person scenes. We empirically show that using data from APU improves mesh estimation on unseen interactions when training a diffusion-based contact prior. Our work addresses longstanding challenges of data scarcity for close interactions in HME enhancing the field's capabilities of handling complex interaction scenarios. Our code, models and data will be made publicly available upon acceptance.*

## 1. Introduction

Understanding human behavior is fundamental for many fields, such as socially aware robotics, patient-caregiver interactions in healthcare, and parent-child interactions in psychology. Central to this pursuit is the study of close interactions between individuals, which are crucial for deciphering the complexities of human dynamics. The field of

Human Mesh Estimation (HME) has emerged as a promising approach to study these dynamics, leveraging parametric models to interpret intricate scenes involving multiple people. However, despite significant advancements in multi-person HME, current methods struggle when faced with close interactions. This is because accurately reasoning about contacts requires a deep understanding of their 3D nature—how people touch and interact in three-dimensional space. Capturing this type of training data is particularly challenging due to the scarcity of ground truth 3D meshes, which are essential for precise estimation and analysis. The lack of detailed 3D data leaves gaps in the ability to model the subtleties of human contact, resulting in less accurate and reliable HME outcomes. Addressing this data scarcity is crucial for advancing the field and improving the handling of complex interaction scenarios.

Recent efforts [14, 17, 67] have successfully acquired ground truth data for closely interacting scenes using motion capture (mocap) systems. Although effective, these systems are costly and limit the dataset's scope. Typically, these datasets feature only two subjects at a time, are confined to indoor lab environments, and cover a restricted set of predefined actions. Other approaches [14, 38] have proposed using weak supervision to avoid the need for mocap by formulating contact as a matching problem between surface body regions of the SMPL model [33] in the form of binary contact matrices. While promising, this approach requires manual annotation of each contact region, limiting the scalability and integration into existing HME pipelines.

More recently, Müller et al. [38] successfully generated pseudo-ground truth meshes from manually annotated image-contact matrix pairs from the FlickrCI3D dataset [14], which they used to train a diffusion-based contact prior for HME. The key insight for this approach was that the additional data enabled the contact prior to learn more meaningful contacts, significantly enriching the training set. However, despite these advancements, the contact prior still faces challenges with complex or out-of-distribution interaction scenarios. This problem is compounded by the high cost of sourcing relevant images and manually annotating people interacting with their contact matrices. This highlights the ongoing need for solutions that can reflect the wide range of interactions found in in-the-wild scenes. Further progress in HME will be facilitated by datasets that accurately mirror natural interactions, capturing the full spectrum of human dynamics in diverse and unstructured environments.

In this work, we introduce a novel data generation method and dataset to increase the diversity of posed meshes interacting closely. We develop an innovative approach *Ask, Pose, Unite* (APU) to automatically create paired pseudo-ground truth meshes for scenes with closely interacting individuals (see Figure 1). First, we *Ask* a Large Vision Language Model (LVLMs) to produce contact maps from an image, then we use an optimization based on 2D keypoint reprojection to *Pose* the 3D meshes. Finally, we *Unite* the meshes by enforcing contact between them using the predicted contacts.

Table 1. Multi-person Human Mesh Estimation datasets. ✕: absent, ✓present, ✓̸: has some examples. Size: images or unique sequences. CI: close interactions.

| Dataset | Source | Size | Subjects | Subjects | Actions | CI | Contact anns. |
|---|---|---|---|---|---|---|---|
| AGORA [39] | synth. | 18k | all ages | 5 - 15 | - | ✕ | ✕ |
| BEDLAM [4] | synth. | 380k | adults | 1 - 10 | - | ✕ | ✕ |
| 3DPW [57] | wild | 60 | adults | $\geq 2$ | - | ✓̸ | ✕ |
| MuPoTS-3D [36] | lab & wild | 20 | adults | 3 | - | ✓̸ | ✕ |
| MultiHuman [70] | lab | 150 | adults | 1-3 | - | ✓̸ | ✕ |
| ExPI[17] | lab | 60k | adults | 2 | 16 | ✓ | ✕ |
| Harmony4D[26] | controlled | 21 | adults | 24 | 6 | ✓ | ✓ |
| FlickrCI3D [14] | wild | 10k | all ages | $> 100$ | - | ✓ | ✓ |
| CHI3D [14] | lab | 631 | adults | 2 | 8 | ✓ | ✓ |
| Hi4D [67] | lab | 100 | adults | 2 | 22 | ✓ | ✓ |
| **APU (Ours)** | lab & wild | 6209 | all ages | $> 2$ | $> 100$ | ✓ | ✓ |

Our APU method enables the generation of detailed 3D representations of interactions without the need for costly and labor-intensive motion capture systems. To support this, we curate a dataset, which features a large diversity of natural interaction types. We collect these types from in-the-wild images depicting people involved in close contact, capturing the complexity and variability of real-world human dynamics. By including a wide range of interaction scenarios, our method and dataset provide a robust foundation for training Human Mesh Estimation (HME) models for interacting humans. We validate the effectiveness of our data generation method by improving a contact prior for HME on new interactions through generating in-domain pseudo data. We also quantify the quality of our predictions on the established lab dataset Hi4D. This validation shows that our approach not only enhances the quality of HME models but also ensures their adaptability to diverse interaction scenarios.

Our contributions can be summarized as follows: **(1)** We propose APU a novel data generation method for close interactions that leverages noisy automatic annotations to scale data acquisition, producing pseudo-ground truth meshes from in-the-wild images. **(2)** We curate a dataset of paired images and pseudo-ground truth meshes featuring a diverse array of close interaction types and subjects. **(3)** We demonstrate that the data generated with APU significantly enriches the representation space of a close contact prior for HME, improving accuracy particularly for less common interaction scenarios in the NTU RGB+D 120 dataset.

## 2. Related Work

**Multi-person HME.** Monocular Human Mesh Estimation is an underspecified problem, particularly challenging

due to the difficulty of capturing paired 2D to 3D ground truth, especially for multiple people. To address this challenge, various datasets have adopted alternate supervision strategies (see Table 1). Some datasets, [4, 39] use synthetic data to bypass the difficulties of capturing real-world 3D ground truth, providing a large number of images and subjects. Others [36, 70] restrict their settings to lab environments, capturing high-quality 3D data but a trade-off on diversity. Since the work by [5], single person HME methods have relied on weak supervision to overcome the scarcity of paired 2D to 3D ground truth, using body part segmentations [24], 2D keypoints [29, 60], and priors based on mocap data [28, 34, 41, 58]. Our work extends this line of research by introducing a data generation method for paired 2D to 3D pseudo-ground truth.

Multi-person HME methods often process individuals independently, which can yield accurate predictions for isolated figures but fails to correctly position them relative to one another in world space. Recent advancements have addressed these issues by using a unified spatial framework for entire images [31], jointly modeling scene and camera dynamics [66], tracking people across time [16, 44, 52, 68], harnessing all available data [6], and managing occlusions [10, 25, 71]. In contrast, single-stage approaches [2, 23, 48–50, 59], which predict all subjects simultaneously, have demonstrated superior performance in terms of spatial accuracy and scale consistency. Despite their effectiveness, these methods depend heavily on extensive datasets, which are scarce for interactions involving close proximity. Our work aims to enrich the data available for these scenarios, potentially enhancing the effectiveness of existing models.

**Close interactions in HME.** Recently, studying close interactions in multi-person HME has become possible largely due to the introduction of new datasets (see Table 1). Lab-based datasets such as CHI3D [14], Hi4D [67], ExPI [17], and Harmony4D [26] provide 3D ground truth via capture systems with multiple calibrated cameras. However, these precise annotations come at the cost of the variety of scenes that can be captured. These works make the most of their scenes by defining a set of actions which reflect common or every-day interactions.

In this context, Fieraru et al. [14] expand on the interaction types by tackling close interactions in-the-wild via weak supervision. In particular, they formulate proximity as a contact problem where the objective is to minimize the distance between surfaces in contact. [14] introduce the FlickrCI3D dataset, a large collection of images from the internet where pairs of people in contact are manually annotated with contact maps, binary matrices that indicate which body parts are in contact. This approach has also been used to train diffused models as contact priors [38] and has also been expanded to include self-contact [37] and scene contact [3, 19, 22] scenarios. More recently, [12] explore contact between people via conditional motion generation. Despite the effectiveness of contact maps in lifting 2D information onto a 3D representation space, they are costly to annotate. By leveraging LVLMs, our work introduces an automatic method for predicting contact maps directly from 2D images. Our approach can also use LVLMs to extract other relevant contextual information such as the type of interaction or descriptions of the scene. This automated approach reduces the annotation burden, scales up data acquisition, and enriches a model's training data with examples from a new or target distribution.

Learning from contact maps to model interactions has been explored more extensively in the context of human-object interactions, either by predicting contact maps directly from images [9, 56] or by inferring object affordances from mesh estimates [11]. Concurrent to this work, [47] has very recently proposed a method to produce contact maps from LVLMs, but their approach is limited to the existing scope of close interaction datasets. In contrast, our method and generated dataset explicitly address the problem of data diversity by using LVLMs as part of a scalable data generation technique that improves HME on novel interactions.

Beyond human-object interactions, the focus has shifted towards understanding how individuals interact with their environments, such as improving motion realism through accurate ground-plane contact [46, 56, 65]. Some works combat data scarcity through the use of synthetic data [20, 35], pre-scanned scenes [45], and by leveraging expert models in object detection and mesh reconstruction [69]. Similarly, person-to-person interaction studies have also focused on predicting contact maps from images [8, 15], but they do not handle out-of-distribution or complex interactions well. In response, Müller et al. [38] propose a diffusion-based contact prior trained with pseudo-ground truth 3D meshes created by constraining the optimization with manually annotated contact maps. Our work builds on this line of research by proposing a method to make a contact prior more robust to new interactions.

**LVLMs for 3D understanding.** There is a growing line of research that employs LVLM's to obtain representations better aligned with real-world scenarios. Some works focus on 3D reasoning, utilizing LVLMs for the assessment of 3D reconstructions [61], motion generation [42], and enhancing the diversity of representation spaces [18, 62]. Others explore human-object contacts [27, 63], and directly reasoning about pose [13, 54]. Our work contributes to this line of research by employing LVLMs as weak annotators to scale data generation and improve the modeling of close human interactions in 3D.

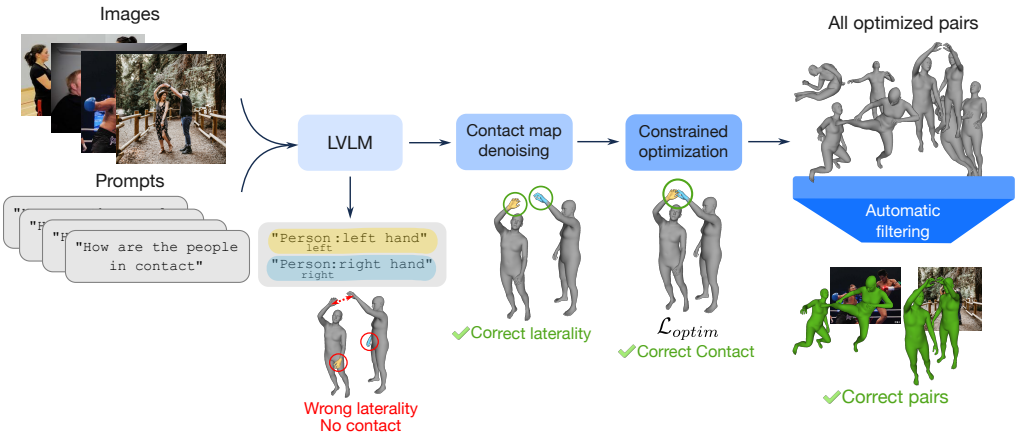


Figure 2. **Overview of our data generation method**. From any set of images we obtain pairs of people in contact and their pseudo-ground truth meshes. For candidate pairs of people in contact we query an LVLM for their contact maps, then denoise the laterality of the contact maps via predicted 2D keypoint chirality and confidence-based soft contact maps. We use the contacts to constrain the optimization of the mesh parameters and automatically filter out failure cases to produce a set of image and correctly reconstructed mesh pairs.

# 3. Ask Pose Unite

## 3.1. Data generation method

**Problem formulation.** We aim to curate images depicting pairs of people closely interacting with well-reconstructed pseudo-ground truth meshes from any set of in-the-wild images. To achieve this, we propose a data generation method (Figure 2 outlines the main steps of our approach). Specifically, our goal is to locate pairs of closely interacting people within any set of images and produce mesh estimates for each pair. Since we only rely on weak supervision in the form of predicted contact maps, 2D keypoints, and interaction labels, we also aim to automatically select the well-reconstructed meshes.

In the context of a single image capturing a scene of close interactions between individuals, our objective is to fit a SMPL-X [41] parametric 3D human mesh model for each individual $p$ to recover their pose $\theta_p \in \mathbb{R}^{21\times3}$ and shape $\beta_p \in \mathbb{R}^{10}$ parameters. We position each mesh in world coordinates by also estimating the root translation $\gamma_p \in \mathbb{R}^3$ and global body rotation $\phi_p \in \mathbb{R}^3$. Following previous work [39, 50], we support the prediction of multiple ages including children with the SMPL-XA model which adds an interpolation parameter $\sigma_p$ between the shape space of SMPL-X and SMIL [21]. In practice $\sigma_p$ is concatenated to the shape parameters such that $\beta_p \in \mathbb{R}^{11}$.

Given an unannotated image $I$ of two people closely interacting we aim to recover their meshes $M^a$ and $M^b$ by following an optimization of the parameters $\{\theta_p, \beta_p, \gamma_p, \phi_p, \sigma_p\}_{p=a,b}$ under the constraint of a contact map $C$. Where $C \in \{0,1\}^{R\times R}$ is defined as a guidance of which body surface regions are in contact. In particular, $C_{i,j} = 1$ indicates that region $r_i$ of $M^a$ is in contact with region $r_j$ of $M^b$.

**Candidate proposal.** For a set of images we obtain 2D keypoints and initial mesh predictions from off-the shelf estimators. We propose as candidates all pairs of people with $k$ valid keypoints within a distance $d$ of each other and with mesh predictions aligned with the keypoints.

**LVLM contact map querying.** We employ a LVLM to automatically generate $C$. The inherent challenge of using LVLMs for this task lies in their low performance when grounding complex spatial relationships depicted in 2D images [55]. Naively querying the LVLM often results in hallucinated or missing contacts leading to degenerate mesh predictions. We tackle this limitation by in-context prompting and denoising the contact maps (explained in the following section).

To query the LVLM for pairs of body parts that are touching in $I$ we first regroup the 75 body regions introduced in [14] into coarser semantically meaningful sets. These sets correspond to the body parts: hand, arm, leg, thigh, chest, stomach, back, neck, face, head, foot, shoulder, elbow, knee, forearm, upper arm, and waist. In practice we re-map a body part $B_i$ to a list of corresponding body regions such that $B_i = \{r_1, r_2, ..., r_n\}$. When available, we ground the LVLM by incorporating a low-cost soft label $A$, which indicates the type of interaction depicted in the image. $A$ is a fundamental element of existing close interaction datasets and serves in this setting as a contextual prior.

**Contact map denoising.** We observe that even with contextual clues, LVLMs are unreliable when predicting the laterality of the body parts. We hypothesize that this problem arises from the model needing to reconcile two conflicting frames of reference: the visual perspective of the image and the anatomical orientation of the human body. For instance, a person's right hand may appear on the left side of an im-

Figure 3. **Examples of mesh pairs and images from our APU dataset** obtained with our data generation method. Note that even with imperfect 3D meshes from the weak-supervision, our method can be used on a wide variety of subjects, ages, interactions, and settings.

age. To correct these mistakes, we exploit the commonalities between estimated 2D pose keypoints and surface body regions. In particular, 2D keypoint estimation methods have been trained on larger sets of manual annotations that explicitly address the conflicting frames of reference.

Given two predicted body parts in contact with their body sides (either left, right, or both) $B_{i,side}$ and $B_{j,side}$, we use the normalized distance between the set of corresponding 2D keypoints for each body part to determine the chirality. Due to the difference in appearance between same-side and opposite-side contacts, we only compare the combinations that match the description. For example, if the prediction is $B_{i,left}$ and $B_{j,right}$, we evaluate two possible combinations: ($B_{i,left}$, $B_{j,right}$) and ($B_{i,right}$, $B_{j,left}$). We then select the combination with the closest normalized distance between the corresponding 2D keypoints.

**Constrained optimization.** Following prior work [5, 38], we obtain pseudo-ground truth meshes $M^a$ and $M^b$ for a pair of people in contact using a two-stage optimization which takes as input estimated 2D keypoints, an initial estimate of the parameters $\{\tilde{\theta}_p, \tilde{\beta}_p, \tilde{\gamma}_p\}_{p=a,b}$, and $C$. In the first stage we optimize $\{\theta_p, \beta_p, \gamma_p\}_{p=a,b}$ given a contact loss $\mathcal{L}_C$ and other priors as guidance. We propose a soft version of the contact loss from [38] to account for the uncertainty in the predicted contacts. $\mathcal{L}_C = \sum_{i,j} W_{ij} C_{ij} \min_{v \in r_i, u \in r_j} \|v - u\|^2$, where $u$ and $v$ are vertices, and $W_{ij} =$ is the normalized keypoint distance from the denoising step scaled by the LVLM's contact confidence.

As additional guidance for the optimization we use a pose prior based on a Gaussian Mixture Model $\mathcal{L}_{GMM}$ [5], an $L_2$ shape prior $\mathcal{L}_\beta$ that penalizes deviation from the SMPL-X mean shape, $\mathcal{L}_{\bar{\theta}}$ an $L_2$ loss that penalizes deviation from the initial pose $\tilde{\theta}_p$, and a 2D keypoint reprojection loss $\mathcal{L}_J$. In the second stage we fix $\beta_p$ and add $\mathcal{L}_P$ to resolve interpenetration between meshes [38].

The complete loss for the constrained optimization with values that re-weigh each term is: $\mathcal{L}_{optim} = \lambda_J \mathcal{L}_J + \lambda_C \mathcal{L}_C + \lambda_{GMM} \mathcal{L}_{GMM} + \lambda_\beta \mathcal{L}_\beta + \lambda_P \mathcal{L}_P + \lambda_{\bar{\theta}} \mathcal{L}_{\bar{\theta}}$

**Automatic filtering.** As a last step we implement a filtering strategy to remove incorrect mesh products from the optimization by thresholding the 2D keypoint reprojection loss of $M^a$ and $M^b$. We keep all instances with error less than 20 for both subjects.

**Implementation details.** We use GPT-4V [1] from 2024-04-09 as the LVLM and ViTPose [64] as the keypoint estimator. For the constrained optimization with the generated contact maps we include Openpose [7] as an additional keypoint estimator. We set $\lambda_C = 1.0, \lambda_J = 0.02$ for both stages and follow [38] for all other hyperparameters. The optimization was processed on an internal Slurm Linux cluster with Nvidia A600, A100, and L40 GPUs. Due to mismatches between the assigned person identity between the LVLM and initial mesh predictions, for guidance during training we consider the minimum of $\mathcal{L}_C$ for both configurations of people.

### 3.2. Diverse data with APU

We use our data generation method, APU, to compile a dataset with diverse person-to-person interactions. We build on a key insight from prior works [14, 38]: using 2D images with weak labels to target interaction diversity in 3D meshes. We have gathered more than 6,000 meshes paired with images, contact annotations, and natural language descriptions of the interactions from both laboratory and in-the-wild scenes, encompassing a variety of ages, subjects, and interactions (see Table 1). Figure 3 shows examples of the images and mesh pairs obtained with our method from

Table 2. Results on close interaction NTU RGB+D 120 test set. PA-MPJPE: Joint two-person Procrustes aligned MPJPE. Auto CM: contact maps generated by our method. Best values in **bold**.

| Method | Mean | Pat on back | Handshake | Knock over | Grab stuff | Step on foot | High-five | Whisper | Support |
|---|---|---|---|---|---|---|---|---|---|
| BEV | 111.9 | 93.5 | 117.8 | 113.0 | 109.0 | 107.7 | 108.8 | 129.9 | 115.7 |
| Ours (Auto CM) | 100.8 | 100.3 | **100.6** | 96.2 | 94.8 | 95.2 | 103.4 | 106.2 | 109.5 |
| BUDDI [38] | 98.7 | 89.7 | 112.7 | 89.3 | 96.8 | 100.4 | 105.6 | 96.5 | 98.6 |
| Ours (Contact prior) | **92.5** | **86.7** | 101.2 | **87.8** | **89.7** | **90.9** | **94.1** | **94.6** | **94.8** |

2D images.

To address the skewed distribution of interaction types in existing datasets, we curated the APU dataset from two primary sources of images: those with and without action classes. Below we detail each data source and its attributes.

*TV Interactions* [40]. This dataset was collected from 300 video clips from 20 TV shows, containing 4 interactions: handshakes, hugs, high fives, and kisses, and clips without or with other interactions.

*Human Interaction Images* [53]. This dataset comprises images of facial expressions of people interacting. We selected 7 uncrowded action types: boxing-punching, handshaking, high-five, hugging, kicking, kissing, and talking.

*Relative Human* [50]. This dataset focuses on multi-person scenes with people of all ages including young children.

*NTU RGB+D 120 train* [32]. A dataset for human action recognition with 3D joint annotations from Kinect sensors. We selected 11 of the 26 two-person interaction classes that involve close interactions and contact: punch/slap, kicking, pat on back, hugging, handshake, knock over, grab stuff, step on foot, high-five, whisper in ear, and support somebody. Then, we randomly selected a subset of 3000 images from frames where the subjects are in contact, determined by the 2D keypoints and 3D joint distance between people.

### 3.3. Dataset interaction type analysis

We hypothesize that existing datasets that feature closely interacting humans often suffer from a lack of diversity and imbalance in their interaction types. We perform an analysis of the interactions in HME close interaction datasets using the representation space of CLIP [43] text embeddings as a proxy for analyzing the variety of interaction types across all datasets. For each dataset, including our APU dataset, we curate a list of all unique interaction types as well as their respective frequencies. Then, we extract the CLIP text embeddings for all unique interaction types and visualize the principal components after PCA. Because FlickrCI3D lacks explicit classes we obtain per image descriptions with the BLIP-2 [30] captioning model and group similar actions by pattern matching on the action phrases. For APU we use the interaction predicted by the LVLM.

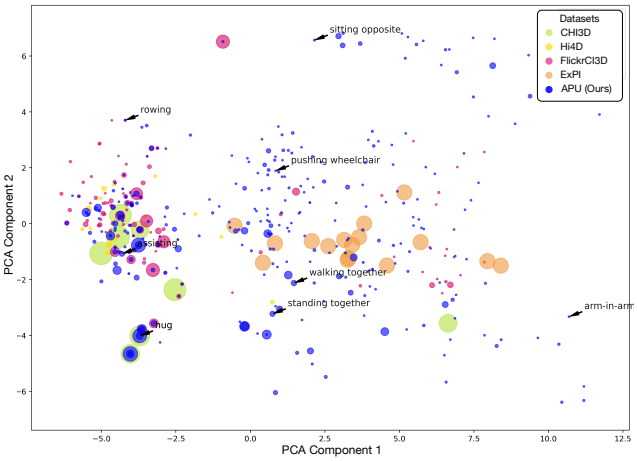


Figure 4. **Distribution of interaction types.** First two principal components of CLIP text embeddings on interaction names and grouped descriptions for existing datasets—CHI3D, Hi4D, FlickrCI3D, and ExPi— and our dataset. Size of points indicate quantity of examples. Our APU dataset contributes a wide range of interactions compared to existing datasets, increasing the diversity of both examples and types of interactions captured.

Figure 4 shows our method's ability to increase data collection on interaction types that are typically underrepresented in prior datasets. Our APU dataset extends beyond the clusters formed by the other datasets, indicating that it includes novel interaction classes. We highlight some example interactions where the increase in diversity or points are noteworthy, such as "assisting", "rowing", "arm-in-arm", among others. This straightforward experiment confirms our hypothesis that both our data generation method and dataset are viable solutions for responding to the data scarcity problem in close interactions for HME.

## 4. Experiments: Using APU to improve estimation for novel interactions

Our data generation method and dataset offer the key advantage of introducing a broader variety of interaction scenarios to enhance training for downstream HME models. However, evaluating data scarcity for out-of-domain interactions poses a challenge due to the limited availability of

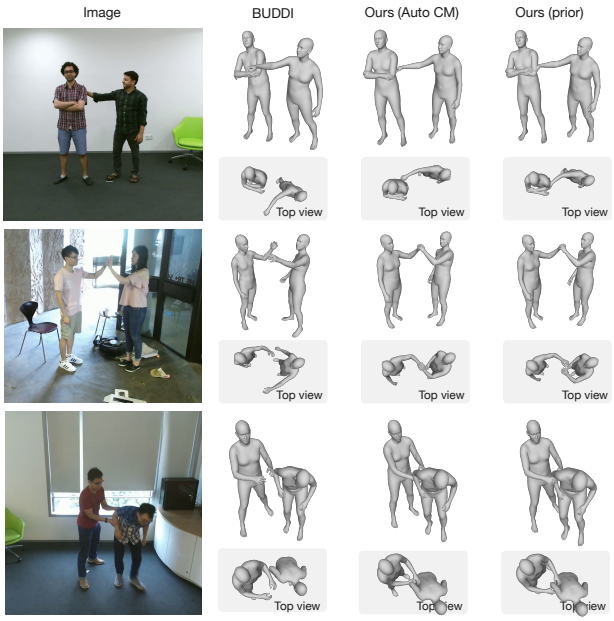


Figure 5. Examples of posed meshes on the close interactions NTU+RGBD 120 test set with BUDDI, Ours Auto CM, and Ours contact prior. Note the improved contact with our contact maps (Auto CM) and trained prior. *Top row* pat on back. *Middle row* high-five. *Bottom row* Support someone.

datasets with 3D ground truth that are not already employed by existing HME methods. To assess the impact of our approach, we repurpose NTU RGB+D 120—a general action recognition dataset. We show how training with our data improves the performance of a state-of-the-art contact interaction prior, BUDDI [38].

### 4.1. Implementation details

**Contact prior model description.** The contact prior BUDDI [38] is a diffusion model conditioned on initial mesh estimates. During training the diffusion model gradually noises data samples to a point of randomness and then learns to reverse this process by denoising samples step-by-step until reaching a coherent structure. In particular, at each time step the noise level $t$ is uniformly sampled with $\epsilon_t \sim \mathcal{N}(0, \mathbf{I})$ to obtain from a ground-truth sample $\mathbf{x}_0$ the noisy sample $\mathbf{x}_t = \sqrt{\sigma'_t}\mathbf{x}_0 + \sqrt{1-\sigma'_t}\epsilon_t$ with $\sigma'_t = \prod_{i=1}^{t}(1-\sigma_t)$. BUDDI is trained to minimize $\mathbb{E}_{\mathbf{x}_0 \sim p_{\text{data}}} \mathbb{E}_{t \sim \mathcal{U}\{0,T\}, \mathbf{x}_t \sim q(\cdot|\mathbf{x}_0)} ||BUDDI(\mathbf{x}_t; t, \varnothing) - \mathbf{x}_0||$. Specifically, an input sample $\mathbf{x}_0$ corresponds to the input SMPL-XA parameters $\phi_p, \theta_p, \beta_p, \gamma_p$ for each person $p$. The loss to train the contact prior is $\mathcal{L}_{prior} = \lambda_\theta \mathcal{L}_\theta + \lambda_\beta \mathcal{L}_\beta + \lambda_\gamma \mathcal{L}_\gamma + \lambda_{v2v} \mathcal{L}_{v2v}$, where all terms are $L_2$ losses w.r.t the parameters and $\mathcal{L}_{v2v}$ is a squared $L_2$ loss on the vertices.

**Inference with the contact prior.** At test time, we perform a two-stage optimization process to obtain the mesh estimates $M^a$ and $M^b$ for the pair of people in an image similarly to section 3.1. However, we replace the contact map guidance with the trained contact prior. At each iteration we diffuse and denoise the current estimate $\mathbf{x}_0$ with a noise level at $t = 10$. The denoised estimate $\hat{\mathbf{x}}_0$ regularizes the current estimate $\mathbf{x}_0$ with an $L_2$ loss $L_{\text{diffusion}} = ||\hat{\mathbf{x}}_0 - \mathbf{x}_0||$. In practice, the decoded parameters are penalized directly by $L_{\text{diffusion}} = \lambda_{\hat{\phi}}||\hat{\phi}_0 - \tilde{\phi}|| + \lambda_{\hat{\theta}}||\hat{\theta}_0 - \tilde{\theta}|| + \lambda_{\hat{\beta}}||\hat{\beta}_0 - \tilde{\beta}|| + \lambda_{\hat{\gamma}}||\hat{\gamma}_0 - \tilde{\gamma}||$. The contact prior offers enough guidance that the GMM pose prior is not needed. Thus, the complete loss function for the optimization is $\mathcal{L}_{optim} = \lambda_J \mathcal{L}_J + \lambda_{\tilde{\theta}} \mathcal{L}_{\tilde{\theta}} + \lambda_P \mathcal{L}_P + L_{\text{diffusion}}$. Where $\mathcal{L}_{\tilde{\theta}}$ is a prior to encourage the solution to be close to the denoised initialization.

**Data preparation.** NTU RGB+D 120 contains 120 actions of which 11 involve 2 people and contact. For every sequence in the dataset's original test set we label the contact frames with a combination of the distance between the annotated 3D joints, 2D keypoints from an off-the shelf estimator, and manual frame-level annotation. Then, we ensure the quality of the 3D joints both visually and by calculating the error between the 2D keypoints and reprojected joints. The final test set comprises 309 frames across 8 classes with a mean of 38.6 (SD: 16.3) frames per class. We follow the data preparation of [38] and train the contact prior on the ground truth meshes of Hi4D and CHI3D, FlickrFits [38] (the pseudo-ground truth derived from FlickrCI3D), and our APU dataset. We train for 3k epochs on a batch composed of 40% for FlickrFits and 20% for the remaining datasets. We set the same hyperparameters as [38] for training and inference with the contact prior and use an internal Slurm Linux cluster with Nvidia A6000, A100, and L40 GPUs.

**Baselines and metrics.** To validate the effect of training the contact prior with our APU dataset, we compare the performance to several methods: BEV [50], a multi-human HME method that produces the initial mesh estimates input to the optimization; BUDDI [38], the state-of-the-art HME method for close interactions (the contact prior trained without our dataset); and a baseline that uses contact matrices automatically generated from our method using the soft action labels (Auto CM). We evaluate errors between the predicted and ground-truth 3D joints using Mean Per Joint Position Error after aligning both people jointly with Procrustes Alignment (PA-MPJPE).

### 4.2. Results

Table 2 shows the results for the close interaction categories of the NTU RGB+D 120 test set. The automatic contact map baseline (Auto CM) improves on most classes over the initial meshes from BEV. The contact prior benefits from the in-domain training data, performing better than BUDDI

Table 3. Quality of zero-shot mesh generation on the Hi4D test. PA-MPJPE: Joint two-person Procrustes aligned MPJPE. Auto CM: contact maps generated by our method. Best values in **bold**. * no use of groundtruth meshes from dataset for training. ** no automatic filtering. BUDDI trained on Hi4D is also shown for reference.

| Method | Mean | Backhug | Basketball | Cheers | Dance | Fight | Highfive | Hug | Kiss | Pose | Sidehug | Talk |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BUDDI | 95.6 | 95.3 | 94.1 | 110.4 | 96.4 | 109.9 | 64.7 | 120.3 | 82.2 | 103.3 | 94.9 | 79.6 |
| BEV | 138.3 | 219.8 | 104.3 | 90.6 | 153.8 | 145.1 | 109.6 | **112.2** | 155.1 | 167.0 | 150.5 | 113.6 |
| Optimization | 125.7 | 155.2 | 95.5 | **79.8** | 148.7 | 137.1 | 72.9 | 154.6 | 145.1 | 151.7 | 136.9 | 105.0 |
| Heuristic | 119.6 | **136.8** | 98.5 | 109.7 | 135.5 | **105.4** | 67.5 | 157.7 | 152.0 | 123.1 | 125.4 | 103.5 |
| BUDDI* | 111.5 | 153.3 | 99.0 | 103.4 | 106.3 | 112.7 | 58.9 | 147.9 | 129.7 | 118.3 | 108.2 | **88.9** |
| Ours (Auto CM)** | 106.8 | 160.2 | **72.0** | 101.0 | 121.4 | 115.8 | **52.2** | 118.5 | 92.4 | 128.5 | 107.1 | 105.9 |
| Ours (Auto CM) | **100.2** | 158.5 | **72.0** | 93.9 | **104.8** | 113.3 | **52.2** | 116.4 | **91.8** | **87.4** | **100.1** | 112.1 |


Figure 6. Example renderings of our method from the user study.

on all classes, and showing significant improvements on uncommon interactions such as step on foot, grab stuff, and support. Common actions, such as handshake and high-five, also benefit from a larger diversity of training examples.

Figure 5 shows examples of posed meshes from the close interactions subset of the NTU RGB+D 120 test set. Training the contact prior with in-domain data generated from APU improves the alignment of contact maps with image evidence, ensuring more accurate contact between interacting individuals.

**Measuring interaction diversity beyond lab settings.** One of the key strengths of our method is its applicability to in-the-wild images. To evaluate its effectiveness, we conduct a user study assessing the quality of meshes generated through our constrained optimization process using predicted contact maps. We compile a set of 90 images from Pexels across 30 diverse interaction categories, reflecting real-life scenarios such as couples yoga, judo, ice skating, and wedding dances. To further challenge our method, we include interactions involving individuals of varying sizes, such as being carried on shoulders and children playing. Figure 6 shows sample reconstructions from our method.

We recruit 30 participants which select the best reconstruction from BEV, BUDDI and our method following strict guidelines on realness and correct contact. Each user is shown 35 images of which 10 are shared among users, for ensuring evaluator agreement of at least 75%. Our method is the most preferred with a mean of 45.5% (SD 3.2%), while BUDDI and BEV have 34.1% (SD 5.8%) and 20.3% (SD 3.5%), respectively.

### 4.3. Quantitative comparison of zero-shot methods.

Additionally, we evaluate the quality of the generated meshes on the test set of Hi4D (on camera 4), a challenging lab-based dataset. We compare against other methods not trained with groundtruth meshes from Hi4D: BEV and BUDDI* (retrained with the source code). We also show results of the optimization without contact maps and our method without automatic filtering. Table 3 shows how overall our method is capable of predicting meshes in a new domain, improving on both the initial estimates (BEV) and optimization without contact maps, even without automatic filtering. Note the improvements on challenging actions like basketball (23 points) and dance (1.5 points). APU also shows improvements on every day actions like kiss (37.9 points) and pose (30.4 points).

## 5. Conclusion

In this paper, we address a key challenge in HME: data scarcity for new domains. We introduce APU a novel data generation method for close interactions, leveraging automatic annotations to produce pseudo-ground truth meshes from in-the-wild images. We curated the APU dataset, which consists of paired images and pseudo-ground truth meshes, covering a wide range of close interaction types. We demonstrate that our data can be used to improve HME methods for close interactions, particularly for interaction scenarios that are out-of-distribution from existing lab-based datasets.

# References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5, 2

[2] Fabien Baradel*, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas*. Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In *arXiv*, 2024. 3

[3] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022. 3

[4] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023. 2, 3

[5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 561–578. Springer, 2016. 3, 5

[6] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[7] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 5

[8] Junuk Cha, Hansol Lee, Jaewon Kim, Nhat Nguyen Bao Truong, Jaeshin Yoon, and Seungryul Baek. 3d reconstruction of interacting multi-person in clothing from a single image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5303–5312, 2024. 3

[9] Yixin Chen, Sai Kumar Dwivedi, Michael J Black, and Dimitrios Tzionas. Detecting human-object contact in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17100–17110, 2023. 3

[10] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3d human mesh from in-the-wild crowded scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1475–1484, 2022. 3

[11] Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. *arXiv preprint arXiv:2311.16097*, 2023. 3

[12] Qi Fang, Yinghui Fan, Yanjun Li, Junting Dong, Dingwei Wu, Weidong Zhang, and Kang Chen. Capturing closely interacted two-person motions with reaction priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 655–665, 2024. 3

[13] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J. Black. Chatpose: Chatting about 3d human pose. In *CVPR*, 2024. 3

[14] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3, 4, 5

[15] Mihai Fieraru, Mihai Zanfir, Teodor Szente, Eduard Bazavan, Vlad Olaru, and Cristian Sminchisescu. Remips: Physically consistent 3d reconstruction of multiple interacting people under weak supervision. *Advances in Neural Information Processing Systems*, 34:19385–19397, 2021. 3

[16] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 3

[17] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13053–13064, 2022. 2, 3

[18] Sookwan Han and Hanbyul Joo. Chorus: Learning canonicalized 3d human-object spatial relations from unbounded synthesized images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15835–15846, 2023. 3

[19] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision*, 2019. 3

[20] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14708–14718, 2021. 3

[21] Nikolas Hesse, Sergi Pujades, Javier Romero, Michael J Black, Christoph Bodensteiner, Michael Arens, Ulrich G Hofmann, Uta Tacke, Mijna Hadders-Algra, Raphael Weinberger, et al. Learning an infant body model from rgb-d data for accurate full body motion analysis. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 792–800. Springer, 2018. 4

[22] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 13274–13285, 2022. 3

[23] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proceedings of*

*the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2020. 3

[24] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 3

[25] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1715–1725, 2022. 3

[26] Rawal Khirodkar, Jyun-Ting Song, Jinkun Cao, Zhengyi Luo, and Kris Kitani. Harmony4d: A video dataset for in-the-wild close human interactions. *arXiv preprint arXiv:2410.20294*, 2024. 2, 3

[27] Hyeonwoo Kim, Sookwan Han, Patrick Kwon, and Hanbyul Joo. Zero-shot learning for the primitives of 3d affordance in general objects. *arXiv preprint arXiv:2401.12978*, 2024. 3

[28] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020. 3

[29] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 3

[30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 6

[31] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, pages 590–606. Springer, 2022. 3

[32] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2684–2701, 2020. 6, 2

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 2

[34] Junzhe Lu, Jing Lin, Hongkun Dou, Ailing Zeng, Yue Deng, Yulun Zhang, and Haoqian Wang. Dposer: Diffusion model as robust 3d human pose prior. *arxiv:2312.05541*, 2023. 3

[35] Zhengyi Luo, Shun Iwase, Ye Yuan, and Kris Kitani. Embodied scene-aware human pose estimation. *Advances in Neural Information Processing Systems*, 35:6815–6828, 2022. 3

[36] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018. 2, 3

[37] Lea Muller, Ahmed AA Osman, Siyu Tang, Chun-Hao P Huang, and Michael J Black. On self-contact and human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9990–9999, 2021. 3

[38] Lea Müller, Vickie Ye, Georgios Pavlakos, Michael Black, and Angjoo Kanazawa. Generative proxemics: A prior for 3d social interaction from images. *arXiv preprint arXiv:2306.09337*, 2023. 2, 3, 5, 6, 7

[39] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. Agora: Avatars in geography optimized for regression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13468–13478, 2021. 2, 3, 4

[40] Alonso Patron-Perez, Marcin Marszalek, Ian Reid, and Andrew Zisserman. Structured learning of human interactions in tv shows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(12):2441–2453, 2012. 6, 2

[41] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 3, 4

[42] Haoxuan Qu, Ziyan Guo, and Jun Liu. Gpt-connect: Interaction between text-driven human motion generator and 3d scenes in a training-free manner. *arXiv preprint arXiv:2403.14947*, 2024. 3

[43] 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. 6

[44] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people by predicting 3D appearance, location & pose. In *CVPR*, 2022. 3

[45] Zehong Shen, Zhi Cen, Sida Peng, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Learning human mesh recovery in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17038–17047, 2023. 3

[46] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. *arXiv preprint arXiv:2312.07531*, 2023. 3

[47] Sanjay Subramanian, Evonne Ng, Lea Müller, Dan Klein, Shiry Ginosar, and Trevor Darrell. Pose priors from language models. *arXiv preprint arXiv:2405.03689*, 2024. 3

[48] Qingping Sun, Yanjun Wang, Ailing Zeng, Wanqi Yin, Chen Wei, Wenjia Wang, Haiyi Mei, Chi Sing Leung, Ziwei Liu, Lei Yang, et al. Aios: All-in-one-stage expressive human pose and shape estimation. *arXiv preprint arXiv:2403.17934*, 2024. 3

[49] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, One-stage, Regression of Multiple 3D People. In *ICCV*, 2021.

[50] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting People in their Place: Monocular Regression of 3D People in Depth. In *CVPR*, 2022. 3, 4, 6, 7

[51] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13243–13252, 2022. 2

[52] Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J. Black. TRACE: 5D Temporal Regression of Avatars with Dynamic Cameras in 3D Environments. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[53] Gokhan Tanisik, Cemil Zalluhoglu, and Nazli Ikizler-Cinbis. Facial descriptors for human interaction recognition in still images. *Pattern Recognition Letters*, 73:44–51, 2016. 6, 2

[54] Yu Tian, Tianqi Shao, Tsukasa Demizu, Xuyang Wu, and Hsin-Tai Wu. Hpe-cogvlm: New head pose grounding task exploration on vision language model. *arXiv preprint arXiv:2406.01914*, 2024. 3

[55] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*, 2024. 4

[56] Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J Black. Deco: Dense estimation of 3d human-scene contact in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8001–8013, 2023. 3

[57] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018. 2

[58] Yufu Wang and Kostas Daniilidis. Refit: Recurrent fitting network for 3d human recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14644–14654, 2023. 3

[59] Zeyu Wang, Zhenzhen Weng, and Serena Yeung-Levy. Multi-human mesh recovery with transformers. *arXiv preprint arXiv:2402.16806*, 2024. 3

[60] Zhenzhen Weng, Kuan-Chieh Wang, Angjoo Kanazawa, and Serena Yeung. Domain adaptive 3d pose augmentation for in-the-wild human mesh recovery. In *International Conference on 3D Vision*, 2022. 3

[61] Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. Gpt-4v (ision) is a human-aligned evaluator for text-to-3d generation. *arXiv preprint arXiv:2401.04092*, 2024. 3

[62] Kevin Xie, Jonathan Lorraine, Tianshi Cao, Jun Gao, James Lucas, Antonio Torralba, Sanja Fidler, and Xiaohui Zeng. Latte3d: Large-scale amortized text-to-enhanced3d synthesis. *arXiv preprint arXiv:2403.15385*, 2024. 3

[63] Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. Interdreamer: Zero-shot text to 3d dynamic human-object interaction. *arXiv preprint arXiv:2403.19652*, 2024. 3

[64] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 5

[65] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. Lemon: Learning 3d human-object interaction relation from 2d images. *arXiv preprint arXiv:2312.08963*, 2023. 3

[66] Wei Yao, Hongwen Zhang, Yunlian Sun, and Jinhui Tang. W-hmr: Human mesh recovery in world space with weak-supervised camera calibration and orientation correction. *arXiv preprint arXiv:2311.17460*, 2023. 3

[67] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17016–17027, 2023. 2, 3

[68] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[69] Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 34–51. Springer, 2020. 3

[70] Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6239–6249, 2021. 2, 3

[71] Yixuan Zhu, Ao Li, Yansong Tang, Wenliang Zhao, Jie Zhou, and Jiwen Lu. Dpmesh: Exploiting diffusion prior for occluded human mesh recovery. *arXiv preprint arXiv:2404.01424*, 2024. 3

# 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

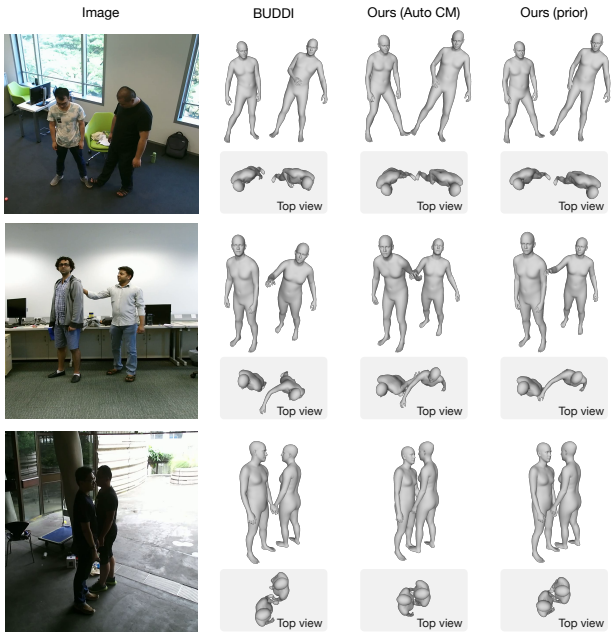


Figure S8. Examples of posed meshes on the close interactions NTU+RGBD 120 test set with BUDDI, Ours Auto CM, and Ours contact prior. *Top row* the contact maps (Auto CM) can enforce contacts when the prior-based methods do not (Ours prior and BUDDI). *Middle row* mistakes in the contact maps lead to 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 | $\geq 2$ | 5970 | 1679 |
| Human Interaction Images [53] | 1177 | all ages | ✓ | 7 | $\geq 2$ | 954 | 282 |
| Relative Human [51] | 8740 | all ages | ✗ | in-the-wild | $\geq 2$ | 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.