Pose Priors from Language Models

Sanjay Subramanian[1] Evonne Ng[1] Lea Müller[1] Dan Klein[1] Shiry Ginosar[2,3] Trevor Darrell[1]
[1] University of California, Berkeley
[2] Google DeepMind
[3] Toyota Technological Institute at Chicago
{sanjayss,evonne_ng,mueller,klein,trevordarrell}@berkeley.edu
shiry@google.com

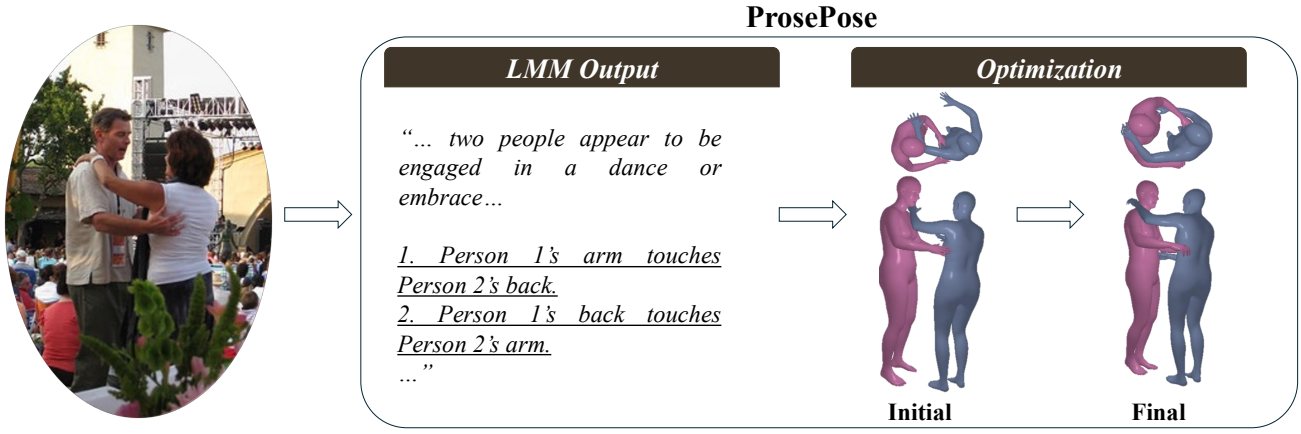

Figure 1. **Optimizing human-to-human contacts in 3D pose.** Our approach leverages the semantic priors of a Large Multimodal Model (LMM) to infer meaningful information about physical contact from images. Instead of relying on human annotations or motion capture data, we extract not only descriptive insights ("... engaged in a dance or embrace ...") but also structured constraints between body parts (underlined). By incorporating these LMM-derived constraints, we refine initial 3D human pose estimates, achieving realistic and semantically consistent reconstructions of contact. This scalable approach opens up new possibilities for contact-aware pose estimation without explicit contact annotations, making it a promising alternative to traditional methods.

Abstract

Language is often used to describe physical interaction, yet most 3D human pose estimation methods overlook this rich source of information. We bridge this gap by leveraging large multimodal models (LMMs) as priors for reconstructing contact poses, offering a scalable alternative to traditional methods that rely on human annotations or motion capture data. Our approach extracts contact-relevant descriptors from an LMM and translates them into tractable losses to constrain 3D human pose optimization. Despite its simplicity, our method produces compelling reconstructions for both two-person interactions and self-contact scenarios, accurately capturing the semantics of physical and social interactions. Our results demonstrate that LMMs can serve as powerful tools for contact prediction and pose estimation, offering an alternative to costly manual human annotations or motion capture data. Our code is publicly available at https://prosepose.github.io.

1. Introduction

Language, as a human artifact, encodes a rich set of social and physical interactions. Over centuries, our vocabulary has evolved to describe the nuances of touch, with words and phrases capturing contexts as varied as hugs, handshakes, or postures in sports and yoga. Perceiving physical contact is essential for understanding human behavior: e.g. several forms of parent-child contact are associated with affection [3, 43], and some forms of self-contact signal stress [16].

Since written language discusses our physical interactions at great length, can large multimodal models (LMMs) trained on images and text correctly perceive physical contact in human pose? This question has practical significance

because scenes with contact are challenging for pose estimation methods as some body parts are frequently occluded, This particularly holds for methods solely relying on 2D keypoints which do not convey contact information. Previously proposed approaches address these issues by curating task-specific datasets via motion capture or human-annotated points of contact between body parts [11, 33, 34]. However, collecting these datasets is expensive, and existing publicly available datasets include only tens of thousands of images [10, 22, 33, 50]. If LMMs can accurately identify contact points, they could decrease the cost of curating such datasets.

In this work, we study the efficacy of LMMs as tools for contact prediction in pose estimation. Since LMMs output language rather than pose parameters, answering this question requires a way of eliciting the required information from an LMM and operationalizing it in pose estimation. We introduce a framework, called ProsePose, which prompts an LMM for formatted constraints about physical contact in the image, converts the constraints into a loss function, and optimizes this loss function (jointly with losses from other cues, such as 2D keypoints) to refine initial pose estimates.

We use ProsePose to evaluate LMMs on both 2-person interaction datasets and a dataset of complex yoga poses. Our framework improves pose estimates compared to strong baselines that do not use contact supervision. In our extensive analysis, we show that several components are important: mitigating LMM hallucinations by aggregating predictions from several LMM samples, careful construction of the prompt and loss functions, and integrating losses from other cues. Finally, we conduct an extensive analysis of LMM predictions and their role in the optimization results. With respect to LMM failures, we find that identifying the chirality of limbs is a particular challenge for LMMs. While existing supervised methods excel by training on large amounts of supervised data with contact labels, we show that we can extract useful priors for contact prediction from pretrained LMMs *without fine tuning*.

In summary, our contributions are (1) we introduce a framework for applying LMMs as contact prediction tools in pose estimation, (2) we show that our framework can improve the quality of pose estimates in 2-person and 1-person settings, and (3) we provide an analysis of the components of our framework and LMM failure modes.

2. Related Work

3D human pose reconstruction. Reconstructing 3D human poses from single images is an active area of research. Prior works have explored optimization-based approaches [13, 26, 35, 36, 39] or pure regression [2, 14, 20, 21, 25] to estimate the 3D body pose given a single image. HMR2 [12] is a recent state-of-the-art regression model in this line of work. Building on these approaches, some methods have looked into reconstructing multiple individuals jointly from a single image. These methods [19, 41, 51] use deep networks to reason about multiple people in a scene to directly output multi-person 3D pose predictions. BEV [42] accounts for the relative proximity of people explicitly using relative depth annotations to reason about proxemics when placing each individual in the scene (e.g. relative depth of people). However, approaches in both categories generally do not accurately capture physical contact between parts of a single person or between people [33, 34].

Contact inference in 3D pose reconstruction. 3D pose reconstruction is especially challenging when there is self-contact or inter-person contact. This has motivated a line of work on pose reconstruction approaches tailored for these settings. [33] focuses on predicting self contact regions for 3D pose estimation by leveraging a dataset with contact annotations to model complex poses such as crossed arms. [10] introduces the first dataset with hand-annotated ground-truth contact labels between two people. REMIPS [11] and BUDDI [34] train models on the person-to-person contact maps in this data in order to improve 3D pose estimation of multiple people from a single image. CloseInt [17] trains a physics-guided diffusion model on two-person motion capture data for this task. However, contact annotations, which are crucial for these approaches, are expensive to acquire. Our method does not require any training on such annotations. Instead, we leverage an LMM's implicit knowledge of pose to constrain pose optimization to capture both self- and person-to-person contact.

Language priors on human pose. There exists a plethora of text to 3D human pose and motion datasets [15, 37, 38], which have enabled work focused on generating 3D motion sequences of a single person performing a general action [18, 44, 52]. This line of work has been extended to generating the motion of two people conditioned on text [29, 40].

PoseScript [7] is a method for generating a single person's pose from fine-grained descriptions, which uses training data from motion capture annotated with detailed text. PoseFix [8] introduces a labeled dataset for the task of modifying a pose given a fine-grained description of the desired change and trains a model on this data. PoseGPT [9] is a pose regressor that uses language as part of its training data. However, PoseGPT does not produce better pose estimates than previous state-of-the-art regressors (i.e. regressors that do not use language) and applies only to the one-person setting. [46] uses a text-only LM to improve action-conditioned human-object pose estimation. This method relies on a limited database of action-pose pairs to classify an input pose, and uses an LM to improve pose estimates based on the action retrieved from the database.

Our work differs from previous work on language and pose in several ways. First, whereas all prior work relies on training data with pairs of language and pose, which

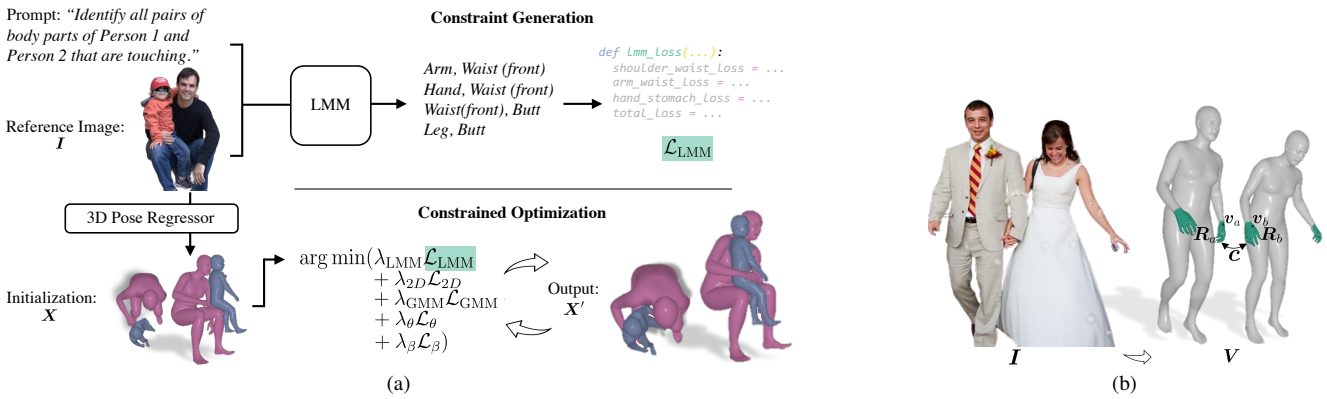

Figure 2. **LMM-guided Pose Estimation.** (a) Method overview: ProsePose takes as input an image of one or two people in contact. We first obtain initial pose estimates for each person from a pose regressor. Then we use an LMM to generate contact constraints, each of which is a pair of body parts that should be touching. This list of contacts is converted into a loss function $\mathcal{L}_{\text{LMM}}$. We optimize the pose estimates using $\mathcal{L}_{\text{LMM}}$ and other losses to produce a refined estimate of each person's pose that respects the predicted contacts. (b) Defining contact constraints: Given an image $\boldsymbol{I}$, we can lift each individual into corresponding 3D meshes $\boldsymbol{V}$. A contact constraint $\boldsymbol{c}$ is a pair of regions $(\boldsymbol{R}_a, \boldsymbol{R}_b)$ in contact. The loss is defined in terms of the distance between the vertices $(\boldsymbol{v}_a, \boldsymbol{v}_b)$ on the mesh.

is expensive to collect, our method leverages the existing knowledge in an LMM to reason about pose from a given image. Second, prior work in this area focuses on either the one-person or the two-person setting. In contrast, our work presents a single framework to reason about physical contacts within or between poses. Finally, in scenes with physical contact, we show that our method improves the pose estimates of state-of-the-art regressors.

3. Guiding Pose Optimization with an LMM

Given an image, our goal is to estimate the 3D body pose of individuals in the image while capturing the self and cross-person contact points. While we cannot trivially use natural language responses (hug, kiss) to directly optimize 3D body poses, we leverage the key insight that LMMs understand *how* to articulate a given pose (arms around waist, lips touching). We propose a method to structure these articulations into constraints and convert them into loss functions.

More concretely, our framework, illustrated by Figure 2a, takes as input the image $\boldsymbol{I}$ and the bounding boxes $\boldsymbol{B}$ of the subjects of interest. In the first stage, a pose regressor takes the image and produces a rough estimate of the 3D pose $\boldsymbol{X}^p$ for each individual p in the image. In the second stage, an LMM takes the image and a set of instructions and generates a list of self- or inter-person contact constraints, which we then convert into a loss function (Sec. 3.2). Finally, in the third stage, we jointly optimize the generated loss function with several other pre-defined loss terms (Sec. 3.3). We refer to our framework as **ProsePose**.

3.1. Preliminaries

We focus our description on the two-person case to keep the exposition simple. We also demonstrate results on the one-person case, which is simply an extension of the two-person case. In particular, we apply our method to the one-person case by setting $X^0 = X^1$. Please see Appendix § 7 for details on the differences between the two cases.

Large Multimodal Models. An LMM is a model that takes as input an image and a text prompt and produces text output that answers the prompt based on the image. Our framework is agnostic to the architecture of the LMM. LMMs are typically trained to respond to a wide variety of instructions [6, 30]. However, LMMs are prone to hallucination [27, 28]. Handling cases of hallucination is a key challenge when using LMMs. We mitigate this issue by aggregating information across several samples from the LMM.

Pose representation. We use a human body model [36] to represent each person $p \in \{0, 1\}$. The body model is composed of a pose parameter that defines the joint rotations $\boldsymbol{\theta} \in \mathbb{R}^{d_\theta \times 3}$, where d_θ is the number of joints, and a shape parameter $\boldsymbol{\beta} \in \mathbb{R}^{d_\beta}$, where d_β is the dimensions of the shape parameter. We can apply a global rotation $\boldsymbol{\Phi} \in \mathbb{R}^3$ and translation $\boldsymbol{t} \in \mathbb{R}^3$ to place each person in the world coordinate space. The full set of parameters for each person is denoted by $\boldsymbol{X}^p = [\boldsymbol{\theta}^p, \boldsymbol{\beta}^p, \boldsymbol{\Phi}^p, \boldsymbol{t}^p]$. For simplicity, we refer to the parameter set $(\boldsymbol{X}^0, \boldsymbol{X}^1)$ as $\boldsymbol{X}$.

These parameters can be plugged into a differentiable function that maps to a mesh consisting of d_v vertices $\boldsymbol{V} \in \mathbb{R}^{d_v \times 3}$. From the mesh, we can obtain the 3D locations of the body's joints $\boldsymbol{J} \in \mathbb{R}^{d_j \times 3}$. From these joints, we can calculate the 2D keypoints $\boldsymbol{K}_{proj}$ by projecting the 3D joints to 2D

using the camera intrinsics Π predicted from [36].

$$\boldsymbol{K}_{proj} = \Pi\left(\boldsymbol{J}\right) \in \mathbb{R}^{d_j \times 2}. \quad (1)$$

Vertex regions. In order to define contact constraints between body parts, we define a set of *regions* of vertices. Prior work on contact has partitioned the body into fine-grained regions [10]. However, since our constraints are specified by a LMM trained on natural language, the referenced body parts are often coarser in granularity. We therefore update the set of regions to reflect this language bias by combining these fine-grained regions into larger, more commonly referenced body parts such as arm, shoulder (front&back), back, and waist (front&back). Some regions, e.g. back and waist (back), overlap. Please see Appendix § 7.2 for a visualization of the coarse regions. Formally, we write $\boldsymbol{R} \in \mathbb{R}^{d_r \times 3}$ to denote a region with d_r vertices, which is part of the full mesh ($\boldsymbol{R} \subset \boldsymbol{V}$).

Constraint definition. A contact constraint specifies which body parts from two meshes should touch. We define contact constraints as pairs of coarse regions $\boldsymbol{c} = (\boldsymbol{R}_a, \boldsymbol{R}_b)$ between a region $\boldsymbol{R}_a$ of one mesh and $\boldsymbol{R}_b$ of the other mesh, as shown in Figure 2b. For instance, ("hand", "arm") indicates a hand should touch an arm.

3.2. Constraint generation with a LMM

Our key insight is to leverage a LMM to identify regions of contact between different body parts on the human body surface. As shown in Figure 2a, we prompt the LMM with an image and ask it to output a list of all region pairs that are in contact. However, we cannot simply use its output natural language descriptions to directly optimize a 3D mesh. As such, we convert these constraints into a loss function.

LMM-based constraint generation. Given the image $\boldsymbol{I}$, we first use the bounding boxes $\boldsymbol{B}$ to crop the part containing the subjects. We then use an image segmentation model to mask any extraneous individuals. While cropping and masking the image may remove information, we find the LMMs are relatively robust to missing context, and more importantly, this allows us to indicate which individuals to focus on. Given the segmented image, we ask the LMM to generate a set $\boldsymbol{C} = \{\boldsymbol{c}_1, ...\boldsymbol{c}_m\}$ of all pairs of body parts that are touching, where m is the total number of constraints the LMM generates for the image.

In the prompt, we specify the full set of coarse regions to pick from. We find that LMMs fail to reliably reference the left and right limbs correctly or consistently, so the prompt instructs the LMM not to specify chirality for each limb (see Appendix 8.2 for prompt analysis). If the LMM uses "left" or "right" to reference a region, despite an instruction to not do so, we directly use the part of the region with the specified chirality rather than considering both possibilities.

Motivated by the chain-of-thought technique, which has been shown to improve language model performance on reasoning tasks [48], we ask the LMM to write its reasoning or describe the pose before listing the constraints. For the full prompt used in each setting, please refer to Appendix § 7.

We sample N responses from the LMM. Below we describe (1) how we convert these natural language responses into N sets of constraints $\{\boldsymbol{C}_1, \boldsymbol{C}_2, ..., \boldsymbol{C}_N\}$ and (2) how we convert each constraint set $\boldsymbol{C}_j$ into a loss.

Canonicalizing Region Names and Assigning Chirality. Given the LMM's output, we must map the mentioned region names to our fixed set of coarse regions. Since the LMM may deviate from the names in the prompt, we check for some additional names (see Appendix 7.2). We then filter out contact pairs that occur fewer than f times across constraint sets, where f is a hyperparameter.

Next, we assign a chirality (left/right) to each hand/arm/foot/leg/shoulder region in cases when the LMM does not itself specify the chirality. We enumerate all possible assignments of left/right to these regions and take the one resulting in the minimum loss. In the two-person setting, we only consider assignments satisfying a condition designed for the case in which a region type occurs in multiple constraint pairs (see Appendix 7.3).

Loss function generation. We compute a loss for each constraint by mapping the relevant regions to sets of vertices and calculating the minimum distance between vertices in the two sets. In particular, we first specify a mapping between each of the coarse region names and the fine-grained regions from [10]. We then use a mapping from fine-grained regions to SMPL-X vertices (provided by [34]) to obtain a set of vertices for each coarse region. Then for each contact pair of coarse regions $\boldsymbol{c} = (\boldsymbol{R}_a, \boldsymbol{R}_b)$ in $\boldsymbol{C}_j$, we define $dist(\boldsymbol{c})$ as the minimum distance between the two regions:

$$dist(\boldsymbol{c}) = \min \|\boldsymbol{v}_a - \boldsymbol{v}_b\|_2 \quad \forall \boldsymbol{v}_a \in \boldsymbol{R}_a, \forall \boldsymbol{v}_b \in \boldsymbol{R}_b \quad (2)$$

where $\{\boldsymbol{v}_a, \boldsymbol{v}_b\} \in \mathbb{R}^3$. In practice, the number of vertices in each region can be very large. To make this computation tractable, we first take a random sample of vertices from $\boldsymbol{R}_a$ and from $\boldsymbol{R}_b$ before computing distances between pairs of vertices in these samples.

Furthermore, since the ordering of the people in the LMM constraints is unknown (i.e. does $\boldsymbol{R}_a$ come from the mesh defined by parameter $\boldsymbol{X}^0$ or $\boldsymbol{X}^1$), we compute the overall loss for both possibilities and take the minimum. We use $\boldsymbol{c}^\top = (\boldsymbol{R}_b, \boldsymbol{R}_a)$ to denote the flipped ordering. We then sum over all constraints in the list $\boldsymbol{C}_j$:

$$dist_{\text{sum}}(\boldsymbol{C}_j) = \min\left(\sum_{\boldsymbol{c} \in \boldsymbol{C}_j} dist(\boldsymbol{c}), \sum_{\boldsymbol{c} \in \boldsymbol{C}_j} dist(\boldsymbol{c}^\top)\right) \quad (3)$$

Each constraint set sampled from the LMM is likely to contain noise or hallucination. To mitigate this issue, we average over all N losses corresponding to each constraint set to obtain the overall LMM loss. This technique is similar to self-consistency [47], which is commononly used for code generation. Concretely, the overall LMM loss is defined as

$$\mathcal{L}_{\text{LMM}} = \frac{1}{N}\sum_{j=1}^{N} dist_{sum}(\boldsymbol{C}_j) \tag{4}$$

If a constraint set $\boldsymbol{C}_j$ is empty (i.e. the LMM does not suggest any contact pairs), then we set $dist_{\text{sum}}(\boldsymbol{C}_j) = 0$. If there are several such constraint sets, we infer that the LMM has low confidence about the contact points (if any) in the image. Consequently, we set a threshold t and if the number of empty constraint sets is at least t, we gracefully backoff to the appropriate baseline optimization procedure (described in Sections 4.1 and 4.2 for each setting). We also backoff to the baseline if the LMM-based optimization diverges.

3.3. Constrained pose optimization

Drawing from previous optimization-based approaches [4, 34, 35], we employ several additional losses in the optimization. We then minimize the joint loss to obtain a refined subset of the body model parameters $\boldsymbol{X}' = [\boldsymbol{\theta}', \boldsymbol{\beta}', \boldsymbol{t}']$:

$$[\boldsymbol{\theta}', \boldsymbol{\beta}', \boldsymbol{t}'] = \arg\min(\lambda_{\text{LMM}}\mathcal{L}_{\text{LMM}} + \lambda_{\text{GMM}}\mathcal{L}_{\text{GMM}} + \lambda_{\beta}\mathcal{L}_{\beta} + \lambda_{\theta}\mathcal{L}_{\theta} + \lambda_{2D}\mathcal{L}_{2D} + \lambda_{P}\mathcal{L}_{P})$$

Following [34], we divide the optimization into two stages. In the first stage, we optimize all three parameters. In the second stage, we optimize only $\boldsymbol{\theta}$ and $\boldsymbol{t}$, keeping the shape $\boldsymbol{\beta}$ fixed. We detail the remaining losses below.

Pose and shape priors. We compute a loss $\mathcal{L}_{\text{GMM}}$ based on the Gaussian Mixture pose prior of [4] and a shape loss $\mathcal{L}_{\beta} = \|\boldsymbol{\beta}\|_2^2$, which penalizes extreme deviations from the body model's mean shape.

Initial pose loss. To ensure we do not stray too far from the initialization, we penalize large deviations from the initial pose $\mathcal{L}_{\theta} = ||\boldsymbol{\theta}' - \boldsymbol{\theta}||_2^2$.

2D keypoint loss. Similar to BUDDI [34], for each person in the image, we obtain pseudo ground truth 2D keypoints and their confidences from OpenPose [5] and ViTPose [49]. Given this pseudo ground truth, we merge all the keypoints into $\boldsymbol{K} \in \mathbb{R}^{d_j \times 2}$, and their corresponding confidences into $\boldsymbol{\gamma} \in \mathbb{R}^{d_j}$. From the predicted $\boldsymbol{X}'$, we can compute the 2D projection of each 3D joint location using Equation 3.1. Then, the 2D keypoint loss is defined as:

$$\mathcal{L}_{2D} = \sum_{j=1}^{d_j} \boldsymbol{\gamma}(\boldsymbol{K}_{proj} - \boldsymbol{K})^2 \tag{5}$$

Interpenetration loss. To prevent parts of one mesh from being in the interior of the other, we add an interpenetration loss. Generically, given two sets of vertices $\boldsymbol{V}_0$ and $\boldsymbol{V}_1$, we use winding numbers to compute the subset of $\boldsymbol{V}_0$ that intersects $\boldsymbol{V}_1$, which we denote as $\boldsymbol{V}_{0,1}$. Similarly, $\boldsymbol{V}_{1,0}$ is the subset of $\boldsymbol{V}_1$ that intersects $\boldsymbol{V}_0$. The interpenetration loss is then defined as

$$\mathcal{L}_P = \sum_{x \in \boldsymbol{V}_{0,1}} \min_{v_1 \in \boldsymbol{V}_1} \|x - v_1\|_2^2 + \sum_{y \in \boldsymbol{V}_{1,0}} \min_{v_0 \in \boldsymbol{V}_0} \|y - v_0\|_2^2 \tag{6}$$

For efficiency, this loss is computed on low-resolution versions of the two meshes (roughly 1000 vertices per mesh).

4. Experiments

Implementation details. Following prior work on two-person pose estimation [34], we use BEV [42] to initialize the poses since it was trained to predict both the body pose parameters and the placement of each person in the scene. However, on the single person yoga poses, we find that the pose parameter estimates of HMR2 [12] are much higher quality, so we initialize the body pose using HMR2.
We use the SMPL-X [36] body model and (unless specified otherwise) GPT4-V [1] as the LMM with temperature $= 0.7$ when sampling from it. GPT4-V refers to the `gpt-4-vision-preview` model in the OpenAI API: platform.openai.com. In the OpenAI API, we use the "high" detail setting for image input. Appendix 8.2 provides results using other prompts and other LMMs (LLaVA [31], GPT-4o) and a running time analysis. Unless otherwise specified, we set $N = 20$ samples. For all of our 2-person experiments, $f = 1$, while $f = 10$ in the 1-person setting. We set $t = 2$ for the experiment on the CHI3D dataset and $t = N$ for all other experiments. The hyperparameters and our prompts were chosen based on experiments on the validation sets. For other implementation details refer to Appendix § 7.

Metrics. As is standard in the pose estimation literature, we report Procrustes-aligned Mean Per Joint Position Error (PA-MPJPE) in millimeters. This metric finds the best alignment between the estimated and ground-truth pose before computing the joint error. In the two-person setting, we focus on the *joint* PA-MPJPE, as this evaluation incorporates the relative translation and orientation of the two people. See Appendix § 8.2 for the per-person PA-MPJPE.

We also include the percentage of correct contact points (PCC) metric introduced by [34]. This metric captures the fraction of ground-truth contact pairs that are accurately predicted. For a given radius r, a pair is classified as "in contact" if the two regions are both within the specified radius. We use the set of fine-grained regions defined in [10] to compute PCC. The metric is averaged over $r \in$

	Hi4D		FlickrCI3D			CHI3D		
	PM$_\downarrow$	F1$_\uparrow$	PM$_\downarrow$	PCC$_\uparrow$	F1$_\uparrow$	PM$_\downarrow$	PCC$_\uparrow$	F1$_\uparrow$
w/o contact sup.								
BEV [42]	144	–	106	64.8	–	**96**	71.4	–
Heuristic	116	–	67	77.8	–	105	74.1	–
ProsePose	**93**	24	**58**	**79.9**	13	100	**75.8**	23
w. contact sup.								
BUDDI [34]	89	–	66	81.9	–	**68**	78.0	–
BUDDI+ProsePose	**88**	–	**65**	**83.2**	–	69	**78.8**	–
Coarse GT contacts								
Oracle	81	100	43	86	100	83	83.8	100

Table 1. **Two-person Results.** Joint PA-MPJPE (abbreviated PM) (lower is better), Avg. PCC (higher is better), and F1 (higher is better). For FlickrCI3D, PA-MPJPE is computed using the pseudo-ground-truth fits. F1 measures the accuracy of coarse contacts predicted by the LMM, while the other two metrics evaluate the quality of the estimated 3D pose. Last line shows results using ground-truth contact pairs of coarse regions (heuristic is still used as the backoff method when needed). **Bold** indicates best method without/with contact supervision in each column.

	FlickrCI3D PCC$_\uparrow$			CHI3D PCC$_\uparrow$		
@ $radius[mm]$	5	10	15	5	10	15
W/o contact sup.						
BEV [42]	3.6	6.3	10.8	5.8	17.4	32.5
Heuristic	14.6	33.9	49.3	11.1	28.0	45.3
ProsePose	**15.6**	**39.9**	**57.1**	**13.5**	**35.2**	**52.5**
W/ contact sup.						
BUDDI [34]	18.5	44.2	61.8	15.5	39.0	56.6
BUDDI+ProsePose	**21.8**	**49.3**	**66.4**	**19.5**	**43.9**	**58.8**

Table 2. **Two-person PCC.** Percent of correct contact points (PCC) for three different radii r in mm. **Bold** indicates the best score without/with contact supervision in each column. At the ground-truth contact points, our method brings the meshes closer together than the baselines.

$0, 5, 10, 15, ..., 95$ mm. Since these regions are defined on the SMPL-X mesh topology, we convert the regression baselines– BEV and HMR2– from the SMPL mesh topology to SMPL-X to compute this metric. Please see the Appendix § 8.1 for more details on the regions and on the mesh conversion.

Finally, we report the F1 of the LMM's predicted coarse region pairs. Specifically, we compute precision as the proportion of predicted pairs $(\mathbf{R}_a, \mathbf{R}_b)$ such that there is some ground-truth pair $(\mathbf{R}_a^*, \mathbf{R}_b^*)$ for which $\mathbf{R}_a$ overlaps with $\mathbf{R}_a^*$ and $\mathbf{R}_b$ overlaps with $\mathbf{R}_b^*$. We compute recall similarly as a proportion of ground-truth pairs. For datasets that do not provide contact maps, we define the ground-truth pairs from the ground-truth meshes, using a threshold on the minimum distance between regions (0.01 and 0.02 for Hi4D and MOYO, respectively). We ignore chirality when computing these. For a pair of people we take the maximum F1 of the two possible orderings. This metric serves to evaluate the raw output of the LMMs without 3D optimization.

4.1. Two-person Pose Refinement

Datasets. We evaluate on three datasets, and our dataset processing largely follows [34]. **Hi4D** [50] is a motion capture dataset of pairs of people interacting. Each sequence has a subset of frames marked as contact frames, and we take every fifth contact frame. We use the images from a single camera, resulting in 241 images. **Flickr Close Interactions 3D (FlickrCI3D)** [10] is a collection of Flickr images of multiple people in close interaction. The dataset includes manual annotations of the contact maps between pairs of people. [34] used these contact maps to create pseudo-ground truth 3D meshes and curated a version of the test set to exclude noisy annotations, which has 1403 images. **CHI3D** [10] is a motion capture dataset of pairs of people interacting. We present results on the validation set. There are 431 images, distributed across 4 cameras. The images come from 126 video sequences, each of which has a single "contact frame."

To develop our method, we experimented on the validation sets of FlickrCI3D and Hi4D, and a sample of the training set from CHI3D. For our experiments, we can compute the PCC on FlickrCI3D and CHI3D, which have annotated ground-truth contact maps. Following [34], we exclude from evaluation images where BEV or the keypoint detectors, which are used by the baselines as well, fail to detect one of the subjects in the interaction pair.

Baselines We compare our estimated poses to the following:

- **BEV [42]** Multi-person 3D pose estimation method. Uses relative depth to reason about spatial placement of individuals in the scene. ProsePose , Heuristic, and BUDDI use BEV to initialize pose estimates.
- **Heuristic** A contact heuristic which includes the auxiliary losses in Section 3.3 as well as a term that minimizes the minimum distance between the two meshes. Introduced by [34]. We use their hyperparameters for this heuristic. This baseline is also used as the backoff method for ProsePose when the number of empty constraint sets is at least the threshold t or when the optimization diverges.
- **BUDDI [34]** This method uses a learned diffusion prior to constrain the optimization. We stress that BUDDI requires a large amount of annotated training data on pairs of interacting bodies, which is not used in our method.

Quantitative Results Table 1 provides quantitative results on the three datasets. Across datasets, ProsePose consistently improves over the strongest baseline, **Heuristic**. On the Hi4D dataset, ProsePose reduces 85% of the gap in PA-MPJPE between **Heuristic** and the fully supervised **BUDDI**. On the FlickrCI3D and CHI3D datasets, ProsePose narrows the gap in the average PCC between **Heuristic** and **BUDDI** by more than one-third. (While ProsePose achieves a better PA-MPJPE than **BUDDI** on FlickrCI3D, for this dataset,

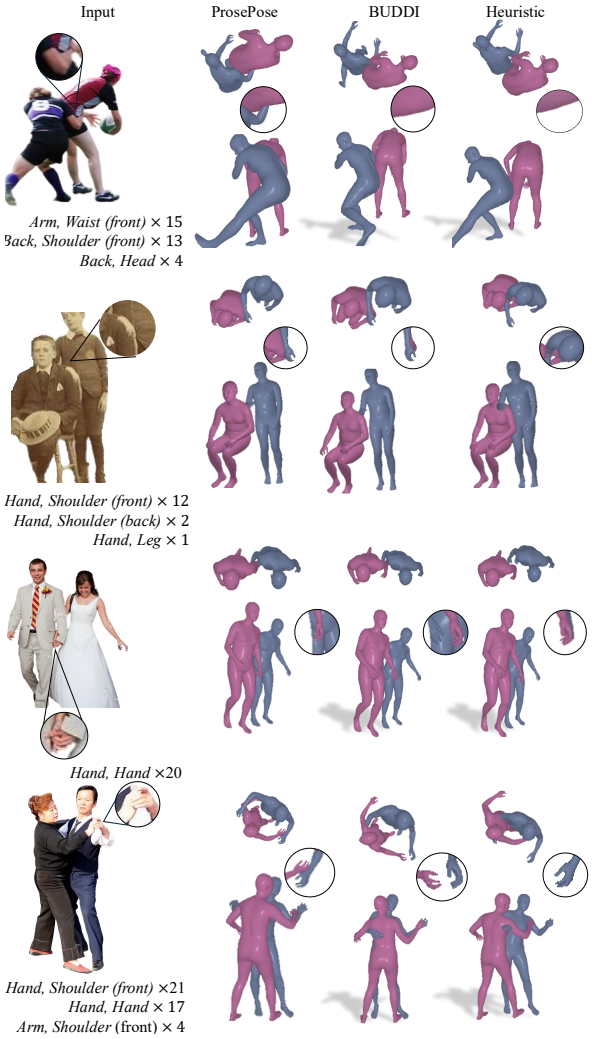

Figure 3. **Two-person examples** We show qualitative results from ProsePose , BUDDI [34], and the contact heuristic. For each example, we show GPT4-V's top 3 constraints and the number of times each constraint was predicted across all 20 samples. Our method correctly reconstructs people in a variety of interactions, and the predicted constraints generally align with each interaction type.

we rely primarily on PCC since PA-MPJPE is computed on *pseudo*-ground-truth fits.)

On CHI3D, ProsePose outperforms **Heuristic** but underperforms **BEV** in terms of PA-MPJPE. On the subset of images where we do not default to the heuristic (i.e. on images where GPT4-V predicts enough non-empty constraint sets), the PA-MPJPE for ProsePose and BEV is 86 and 87, respectively. In other words, in the cases where our method is actually used, the joint error is slightly less than that of BEV. As a result, we can attribute the worse overall error to the poorer performance of the heuristic. The backoff method (which is the heuristic) is used in 13/241 Hi4D examples, 106/1403 Flickr examples, and 224/431 CHI3D examples.

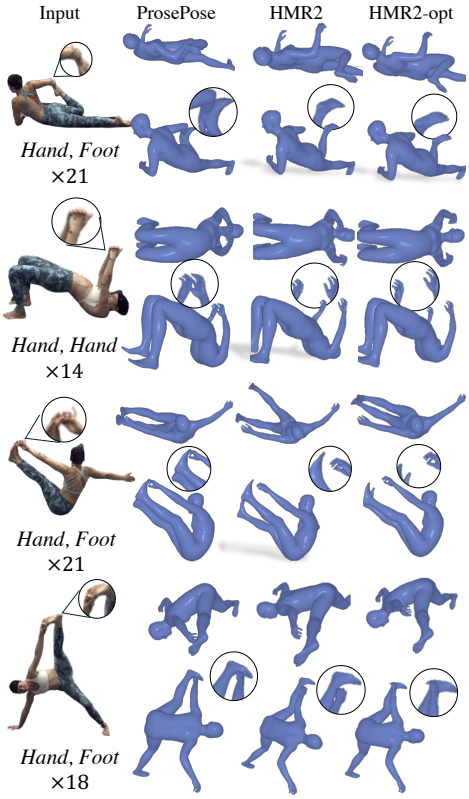

Figure 4. **Single-person examples** We show qualitative results from ProsePose , HMR2 [12], and HMR2-optim on complex yoga poses. Each example also shows the constraints that are predicted by the LMM at least $f = 10$ times (and are thus used to compute $\mathcal{L}_{\text{LMM}}$) with their counts. ProsePose correctly identifies self-contact points and optimizes the poses to respect these contacts.

Overall, our method improves over the other methods that do not use contact supervision in terms of both joint error and PCC. While not the focus of this work, Table 1 also shows that adding ProsePose , specifically $\mathcal{L}_{\text{LMM}}$, to BUDDI leads to improved PCC. The F1 scores show that LMMs' raw output is often flawed/incomplete, but our approach mitigates hallucinations in various ways (see the ablation study below). The last row in Table 1 shows the performance when the ground-truth coarse contacts (with correct left/right labels) are used in optimization. These results show the benefit of correct coarse contacts and the clear room for improvement from better LMM predictions.

Table 2 shows the PCC for each method at various radii. The results show that ProsePose brings the meshes closer together at the correct contact points. On both the FlickrCI3D and CHI3D datasets, ProsePose outperforms the other baselines that do not use contact supervision. Next, we ablate important aspects of ProsePose . In Figure 5, we show that averaging the loss over several samples from the LMM improves performance, mitigating the effect of LMM hallucination. Finally, ablating the various losses in optimization indicates

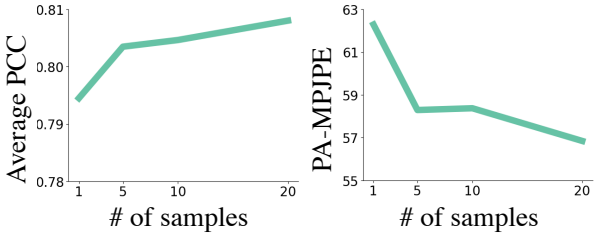

Figure 5. **More samples improve pose estimation.** On the FlickrCI3D validation set, taking more samples from the LMM and averaging the resulting loss functions improves joint PA-MPJPE (left) and average PCC (right).

	PA-MPJPE$_\downarrow$	PCC$_\uparrow$	PCC$_\uparrow$ @ r 5	10	15	F1$_\uparrow$
HMR2 [12]	84	83.0	34.2	55.2	69.5	–
HMR2+opt	**81**	85.2	47.7	65.5	74.6	–
ProsePose	82	**87.8**	**54.2**	**73.8**	**81.4**	25

Table 3. **One-person Results.** PA-MPJPE (lower is better) and Avg. PCC and F1 (higher is better). ProsePose captures ground-truth contacts better than the baselines, as shown by the PCC.

that our LMM-based loss and the 2D keypoint loss have the greatest impact on joint error: Using all losses results in a PA-MPJPE of 81. Removing $\mathcal{L}_{\text{LMM}}/\mathcal{L}_{\text{GMM}}/\mathcal{L}_\beta/\mathcal{L}_\theta/\mathcal{L}_{2D}/\mathcal{L}_P$ results in a PA-MPJPE of 138/85/91/84/130/78.

Qualitative Results Figure 3 shows examples of reconstructions from ProsePose, **Heuristic**, and **BUDDI**. Below each of our predictions, we list the most common constraints predicted by GPT4-V for the image. The predicted constraints correctly capture the semantics of each interaction. For instance, in tango, one person's arm should touch the other's back. In a rugby tackle, a player's arms are usually wrapped around the other player. Using these constraints, ProsePose correctly reconstructs a variety of interactions, such as tackling, dancing, and holding hands. In contrast, **Heuristic** struggles to accurately position individuals and/or predict limb placements, often resulting in awkward distances.

4.2. One-person pose refinement

Datasets Next, we evaluate ProsePose on a single-person setting. For this setting, we evaluate on MOYO [45], a motion capture dataset with videos of a single person performing various yoga poses. In total, our test set is composed of 76 examples from a single camera angle (side view). See Appendix 7.6 for further dataset details. Since this dataset does not have annotated region contact pairs, we compute the pesudo-ground-truth contact maps using the Euclidean and geodesic distance following [33] and report PCC for the subset of 67 examples where the ground-truth has self-contact.

Baselines We compare against the following baselines:

- **HMR2** [12] State-of-the-art pose regression method. We use HMR2 to initialize our pose estimates for optimization.
- **HMR2+opt** Optimization procedure that is identical to ours without $\mathcal{L}_{\text{LMM}}$. It is the default method when the number of empty constraint sets is at least the threshold t.

Both the quantitative and qualitative results echo the trends discussed in the 2-person setting. Table 3 provides the quantitative results. The PCC metrics show that our LMM loss improves the predicted self-contact in complex yoga poses relative to the two baselines. The backoff method is used in 43/76 examples. Figure 4 provides a qualitative comparison of poses predicted by ProsePose versus the two baselines. Below each of our predictions, we list the corresponding constraints predicted by GPT4-V. In each case, the predicted constraint captures the correct self-contact, which is reflected in the final pose estimates. Using the semantically guided loss, ProsePose effectively refines the pose to ensure proper contact between hand-foot or hand-hand, an important detail consistently overlooked by the baselines.

4.3. Limitations

While ProsePose consistently improves contact across settings and datasets, it has limitations which are related to failures of LMMs. First, as shown in Table 5 (in Appendix), prompting the LMM for left/right labels sometimes leads to worse results, suggesting that LMMs struggle with disambiguating chirality. Improving this approach depends in large part on correctly identifying limbs as left/right. Another limitation is the use of coarse regions. Future work could improve by eliciting more fine-grained constraints from an LMM. Finally, LMM accuracy varies moderately across camera angles (quantified in Appendix 8.2). In Appendix § 8.3, we provide examples of LMM failures.

5. Conclusion

We present ProsePose, a framework for refining 3D pose estimates to capture touch accurately using the implicit semantic knowledge of poses in LMMs. Our key novelty is that we generate structured pose descriptions from LMMs and convert them into loss functions used to optimize the pose. Our experiments show that in both one-person and two-person settings, ProsePose improves over previous baselines that do not use contact supervision. These results suggest that LMMs may be useful in creating larger datasets with contact annotations, which are otherwise expensive but are crucial for training state-of-the-art priors for pose estimation in situations with physical contact. More broadly, this work provides evidence that LMMs are promising tools for 3D pose estimation, which likely has implications beyond touch.

6. Acknowledgements

SS, EN, and TD were supported in part by the NSF, DoD, and/or the Berkeley Artificial Intelligence Research (BAIR) industrial alliance program.

References

[1] OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Benjamin Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim'on Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Ryan Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Adeola Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P. Mossing, Tong Mu, Mira Murati, Oleg Murk, David M'ely, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang Long, Cullen O'Keefe, Jakub W. Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario D. Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin D. Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas A. Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cer'on Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report. 2023. 5

[2] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3395–3404, 2019. 2

[3] Ana Aznar and Harriet R Tenenbaum. Parent–child positive touch: Gender, age, and task differences. *Journal of nonverbal behavior*, 40:317–333, 2016. 1

[4] 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. 5

[5] 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

[6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500, 2023. 3

[7] Delmas, Ginger and Weinzaepfel, Philippe and Lucas, Thomas and Moreno-Noguer, Francesc and Rogez, Grégory. PoseScript: 3D Human Poses from Natural Language. In *ECCV*, 2022. 2

[8] Delmas, Ginger and Weinzaepfel, Philippe and Moreno-Noguer, Francesc and Rogez, Grégory. PoseFix: Correcting 3D Human Poses with Natural Language. In *ICCV*, 2023. 2

[9] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J. Black. Posegpt: Chatting about 3d human pose. *ArXiv*, abs/2311.18836, 2023. 2

[10] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7214–7223, 2020. 2, 4, 5, 6

[11] 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. 2

[12] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa*, and Jitendra Malik*. Humans in 4D: Reconstructing and tracking humans with transformers. In *International Conference on Computer Vision (ICCV)*, 2023. 2, 5, 7, 8

[13] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1381–1388. IEEE, 2009. 2

[14] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10884–10894, 2019. 2

[15] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, 2022. 2

[16] Jinni A Harrigan. Self-touching as an indicator of underlying affect and language processes. *Social science & medicine*, 20 (11):1161–1168, 1985. 1

[17] 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 (CVPR)*, pages 1011–1021, 2024. 2

[18] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *arXiv preprint arXiv:2306.14795*, 2023. 2

[19] 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. 2

[20] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *2021 International Conference on 3D Vision (3DV)*, pages 42–52. IEEE, 2021. 2

[21] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[22] Rawal Khirodkar, Jyun-Ting Song, Jinkun Cao, Zhengyi Luo, and Kris M. Kitani. Harmony4d: A video dataset for in-the-wild close human interactions. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 2

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 4

[24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 3

[25] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2252–2261, 2019. 2

[26] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6050–6059, 2017. 2

[27] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*, 2023. 3

[28] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 3

[29] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *arXiv preprint arXiv:2304.05684*, 2023. 2

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 3, 6

[31] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 5, 6

[32] Yujie Lu, Dongfu Jiang, Wenhu Chen, William Wang, Yejin Choi, and Bill Yuchen Lin. Wildvision arena: Benchmarking multimodal llms in the wild, 2024. 8

[33] 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. 2, 8

[34] 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, 4, 5, 6, 7, 9

[35] 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. 2, 5

[36] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and

Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2, 3, 4, 5

[37] Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. *Big Data*, 4(4):236–252, 2016. 2

[38] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with english labels. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 722–731, 2021. 2

[39] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11488–11499, 2021. 2

[40] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 2

[41] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11179–11188, 2021. 2

[42] 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, 5, 6

[43] Mika S Takeuchi, Hitoshi Miyaoka, Atsuko Tomoda, Masao Suzuki, Qingbo Liu, and Toshinori Kitamura. The effect of interpersonal touch during childhood on adult attachment and depression: A neglected area of family and developmental psychology? *Journal of Child and Family Studies*, 19:109–117, 2010. 1

[44] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[45] Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Taheri Omid, Michael J. Black, and Dimitrios Tzionas. 3D human pose estimation via intuitive physics. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4713–4725, 2023. 8

[46] Xi Wang, Gen Li, Yen-Ling Kuo, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Reconstructing action-conditioned human-object interactions using commonsense knowledge priors. In *2022 International Conference on 3D Vision (3DV)*, pages 353–362. IEEE, 2022. 2, 5

[47] X Wang, J Wei, D Schuurmans, Q Le, E Chi, S Narang, A Chowdhery, and D Zhou. Self-consistency improves chain of thought reasoning in language models. arxiv. *Preprint posted online March*, 21:10–48550, 2022. 5

[48] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903, 2022. 4

[49] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, 2022. 5

[50] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan José Zárate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17016–17027, 2023. 2, 6

[51] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2148–2157, 2018. 2

[52] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2