

## Appendix for Pose Priors from Language Models

In this appendix, we provide additional details about our method (Section 7), details about metrics (Section 8.1), additional quantitative results (Section 8.2), examples of failure cases (Section 8.3), experiments with LLaVA (Section 8.4), and more qualitative comparisons (Section 8.5). We also provide a video overview of the method and qualitative results (video.mp4).

## 7. Additional Method Details

### 7.1. LMM Prompts

The box below contains our prompt for the two-person experiments.

You are a helpful assistant. You follow all directions correctly and precisely.

For each image, identify all pairs of body parts of Person 1 and Person 2 that are touching.

Write all of these in a Markdown table where the first column is "Person 1 Body Part" and the second column is "Person 2 Body Part".

You can pick which is Person 1 and which is Person 2.

The list of possible body parts is: head, neck, chest, stomach, waist (back), waist (front), back, shoulder (back), shoulder (front), arm, hand, leg, foot, butt. Do not include left/right.

List ALL pairs you are confident about.

If you are not confident about any pairs, output an empty table.

Carefully write your reasoning first, and then write the Markdown table.

The box below contains our prompt for the one-person experiment.

You are a helpful assistant. You answer all questions carefully and correctly.

Identify which body parts of the yogi are touching each other in this image (if any).

Write each pair in a Markdown table with two columns.

Each body part MUST be from this list: head, back, shoulder, arm, hand, leg, foot, stomach, butt, ground

Do not write "left" or "right".

Describe and name the yoga pose, and then write the Markdown table.

Note that the pose may differ from the standard version, so pay close attention.

Only list a part if you're certain about it.

In each setting, the prompt is given as the "system prompt" to the GPT-4 API, and the only other message given as input contains the input image with the "high" detail setting.

#### 7.1.1 Ablation Prompts

Below we give the alternative prompts evaluated in Table 5.

The box below contains the prompt that is like the default except that it asks for left/right labels.

You are a helpful assistant. You follow all directions correctly and precisely. For each image, identify all pairs of body parts of Person 1 and Person 2 that are touching. Write all of these in a Markdown table where the first column is "Person 1 Body Part" and the second column is "Person 2 Body Part". You can pick which is Person 1 and which is Person 2. The list of possible body parts is: head, neck, chest, stomach, waist (back), waist (front), back, shoulder (back), shoulder (front), arm, hand, leg, foot, butt. For arm/hand/leg/foot/shoulder, prepend "left" or "right". List each body part separately (don't use plural).

List ALL pairs you are confident about. If you are not certain about any pairs, output an empty table. Carefully write your reasoning first, and then write the Markdown table.

The box below contains the prompt that is like the default except that the image is labeled with Person 1 and Person 2 and the prompt asks for left/right labels. Figure 6 shows an example of an input image with labels for Person 1 and Person 2.



Figure 6. An image with labels for Person 1 and Person 2, from the FlickrCI3D validation set

You are a helpful assistant. You follow all directions correctly and precisely. For each image, identify all pairs of body parts of Person 1 and Person 2 that are touching. Write all of these in a Markdown table where the first column is "Person 1 Body Part" and the second column is "Person 2 Body Part". The list of possible body parts is: head, neck, chest, stomach, waist (back), waist (front), back, shoulder (back), shoulder (front), arm, hand, leg, foot, butt. For arm/hand/leg/foot/shoulder, prepend "left" or "right". List each body part separately (don't use plural). Only list pairs you are absolutely certain about. If you are not certain about any pairs, output an empty table. Carefully write your reasoning first, and then write the Markdown table.

The box below contains the prompt for obtaining a pose caption about a pair of people.

Describe the pose of the two people.

The box below contains the prompt for rewriting the caption so that it does not contain references to "left" and "right".

Rewrite the caption below so that it doesn't mention "left" or "right" to describe any hand, arm, foot, or leg. The revised caption should otherwise be identical. Write only the revised caption and no other text.

The box below contains the prompt for converting a caption into a table.

You are a helpful assistant. You will follow ALL rules and directions entirely and precisely. Given a description of Person 1 and Person 2 who are physically in contact with each other, create a Markdown table with the columns "Person 1 Body Part" and "Person 2 Body Part", listing the body parts of the two people that are guaranteed to be in contact with each other, from the following list. ALL body parts that you list must be from this list. You can choose which person is Person 1 and which is Person 2. Body parts: "chest", "stomach", "waist (front)", "waist (back)", "shoulder (front)", "shoulder (back)", "back", "hand", "arm", "foot", "leg", "head", "neck", "butt" Note that "back" includes the entire area of the back. Include all contact points that are directly implied by the description, not just those that are explicitly mentioned. If there are no contact points between these body parts that the description implicitly or explicitly implies, your table should contain only the column names and no other rows. First, write your reasoning. Then write the Markdown table.

For the last two prompts above, which do not involve image input, we use the gpt-4-0125-preview version of GPT4 via the OpenAI API.

## 7.2. Coarse Regions

Figure 7 illustrates the coarse regions referenced in the prompt in our two-person experiments. Figure 8 illustrates the coarse regions referenced in the prompt in our one-person experiments. In the one-person case, the prompt does not mention the "chest," "neck," or "waist" regions, since they tend to be less important for contacts in yoga poses, and the front/back shoulders are merged into one region, since the distinction tends to be less important for contacts in yoga poses.

As stated in § 3.2, the procedure converting LMM outputs to loss functions checks for region names other than those listed in the prompt. In particular, it checks for "waist" and the left/right variants of "hand"/"arm"/"foot"/"leg"/"shoulder"/"shoulder

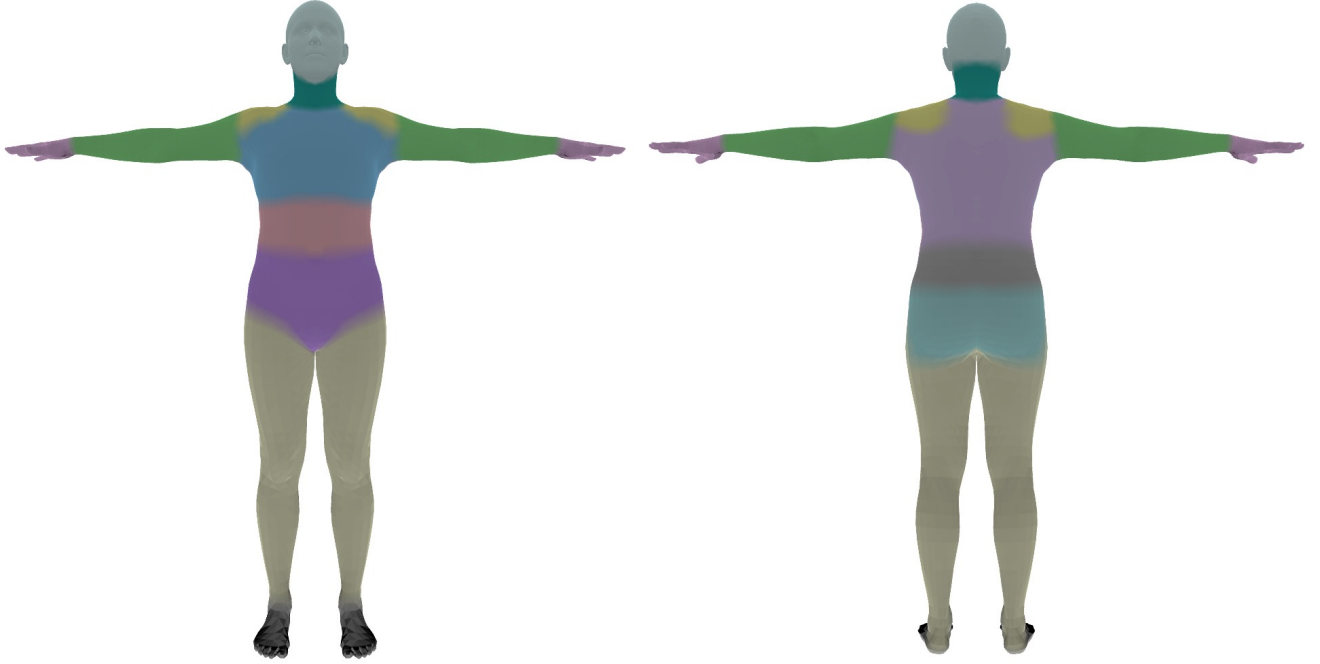


Figure 7. Color-coded coarse regions in the two-person prompt: head, neck, chest, stomach, waist (back), waist (front), back, shoulder (back), shoulder (front), arm, hand, leg, foot, butt. Note that some of these regions overlap. For instance, the “back” includes the “waist (back)” and “shoulder (back)” regions as a subset.

(front)”/“shoulder (back)”. “waist” corresponds to the union of “waist (front)” and “waist (back).” and “shoulder” corresponds to the union of “shoulder (front)” and shoulder (back).” In the one-person setting, the prompt also specifies “ground” as one of the regions, but this is not used in creating loss functions.

### 7.3. Chirality Condition

As mentioned briefly in § 3.2, when enumerating possible chiralities for each constraint, we enforce the following condition in the two-person setting: if the same body part (e.g. “hand”) is mentioned in at least two separate rows of the table output by the LMM (without any “left” or “right” prefix) or is mentioned in the plural form (e.g. “hands”), we enforce that both the left and right limbs of this type must participate in the loss. The motivation for this condition is that when the same constraint applies to both limbs of a given type (e.g. “right hand, back” and “left hand, back”) and the chirality is not specified in the constraint set, the two constraints will appear to be the same (e.g. “hand, back”). But often the LMM

will list the constraint twice, since there are two different contact points, which activates this condition.

### 7.4. Bounding Boxes and Cropping

As stated in Section 3 of the main paper, we take bounding boxes of the subjects of interest as input and use them to crop the image in order to isolate the person/people of interest when prompting the LMM. For FlickrCI3D, we use the ground-truth bounding boxes of the two subjects of interest. For the other datasets, we use keypoints detected by ViTPose/OpenPose to create the bounding boxes. We use Segment Anything [24] as the segmentation model, used to remove extraneous people in the image (we only apply this step for FlickrCI3D, since other datasets are from motion capture). For the single-person MOYO dataset, we manually check that the bounding boxes from the keypoints and the selected HMR2 outputs correspond to the correct person in the image. We note that the baseline HMR2+opt also benefits from this manual checking, since HMR2+opt also depends on the HMR2 outputs and accurate keypoints.

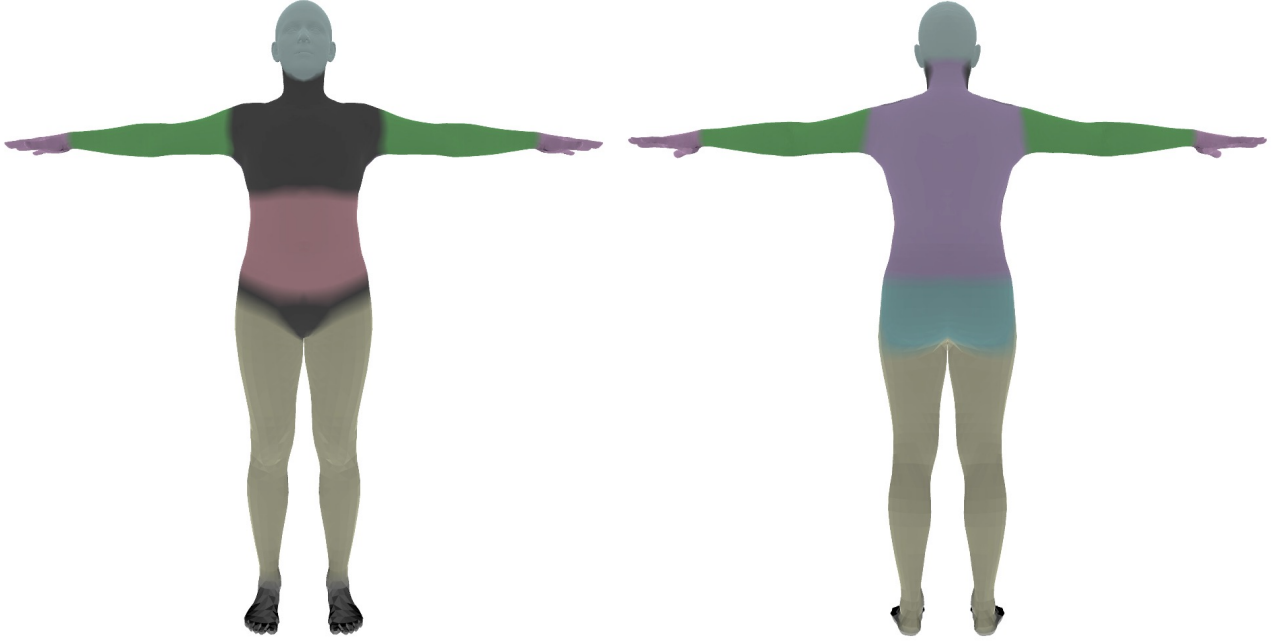


Figure 8. Color-coded coarse regions in the one-person prompt: head, stomach, back, shoulder, arm, hand, leg, foot, butt. Note that the “chest,” “neck,” and “waist (front)” regions are not covered by the regions in the prompt, since they tend to have less importance for contacts in yoga poses.

### 7.5. Loss Coefficients and Optimization Details

We set  $\lambda_{\text{LMM}} = 1000$  in the 2-person experiments, and  $\lambda_{\text{LMM}} = 10000$  in the 1-person setting. In the two-person case, all other loss coefficients are taken directly from [34]. In the one-person case, we find that removing the GMM pose prior and doubling the weight on the initial pose loss improves optimization dramatically, likely because the complex yoga poses are out of distribution for the GMM prior. These hyperparameters and our prompts were chosen based on experiments on the validation sets. Furthermore, following [34], we run both optimization stages for at most 1000 steps. We use the Adam optimizer [23] with learning rate 0.01.

### 7.6. MOYO Dataset Processing Details

The dataset provides views from multiple different cameras. We pick a single camera that shows the side view for evaluation. For each video, we take single frame from the middle as it generally shows the main pose. There is no official test

set, and the official validation set consists of only 16 poses. Therefore, we created our own split by picking 79 arbitrary examples from the training set to form our validation set. We then combine the remaining examples in the training set with the official validation set to form our test set.

## 8. Experiments

### 8.1. PCC Calculation

Figure 9 illustrates the 75 fine-grained regions used for PCC calculation, which are the same as those used in [10]. We opted to compute PCC on the fine-grained regions rather than on the coarse ones since prior work uses the fine-grained regions [34] and since we want to measure contact correctness at a finer granularity (e.g. upper vs. lower thigh vs. knee). Since the regressors BEV and HMR2 use the SMPL mesh while the fine-grained regions are defined on the SMPL-X mesh, we use a matrix  $M \in \mathbb{R}^{\text{num\_vertices\_smplx} \times \text{num\_vertices\_smpl}}$  to convert the SMPL meshes to SMPL-X in order to compute PCC.



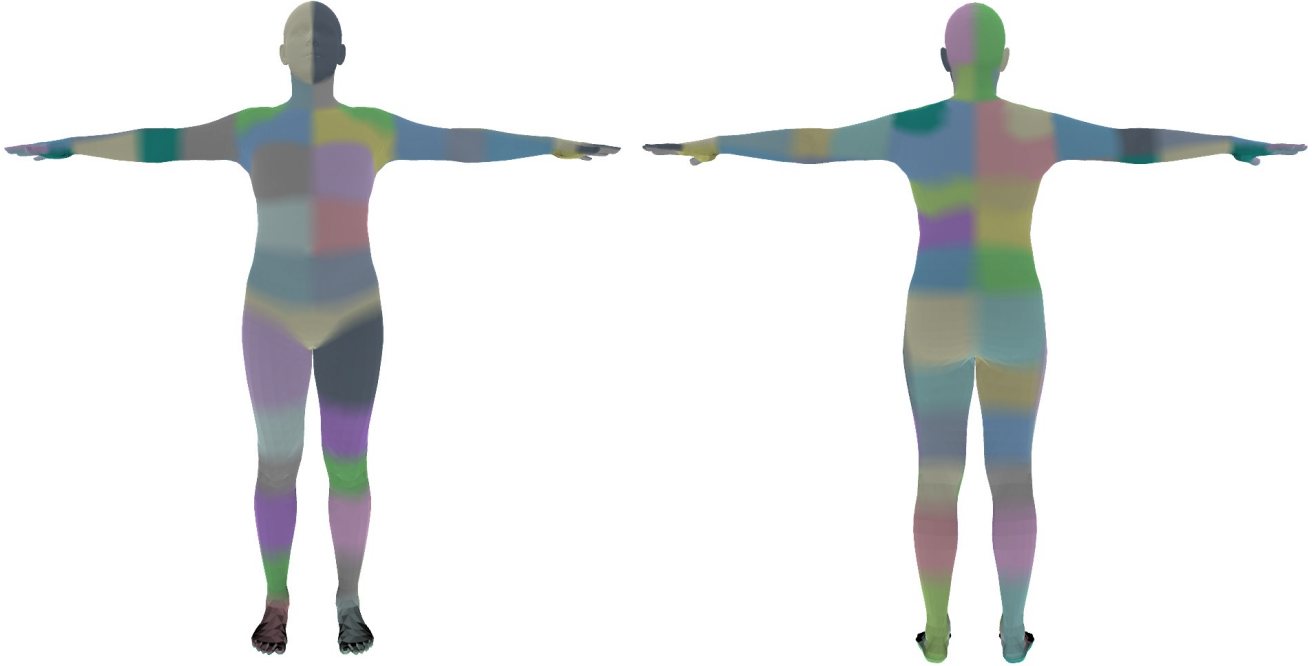


Figure 9. Color-coded 75 fine-grained regions used for PCC calculation

	Hi4D PA-MPJPE $\downarrow$	FlickrCI3D PA-MPJPE $\downarrow$	CHI3D PA-MPJPE $\downarrow$
<i>Without contact supervision</i>			
BEV [42]	76	71	51
Heuristic	65	31	48
ProsePose	65	31	49
<i>With contact supervision</i>			
BUDDI [34]	70	43	47

Table 4. **Two-person Results.** Per-person PA-MPJPE (lower is better). For FlickrCI3D, PA-MPJPE is computed using the pseudo-ground-truth fits.

## 8.2. Additional Quantitative Results

### 8.2.1 Per-person PA-MPJPE

Table 4 shows the per-person PA-MPJPE for each of the datasets used in our two-person experiments.

### 8.2.2 LMM Analysis

In Table 5, we show that ProsePose works with other LMMs—performance is comparable with GPT-4o and worse with LLaVA, which is in line with the general capabilities of these LMMs. For details on how we use LLaVA and other LLaVA results, see § 8.4. We also compare the default prompt with three other prompt types: (1) asking for left/right labels on each limb, (2) labeling each person on the image and asking for left/right labels on each limb, and (3) first generating a pose-focused caption and converting the caption to a set of constraints. Prompt 3 resembles the classify-then-constrain

pipeline of [46]. When using GPT4-V, the default prompt outperforms Prompt 3 on both datasets and Prompts 1/2 on at least one (see § 4.3 for further discussion).

### 8.2.3 Running Time

We compare BUDDI and BUDDI+ProsePose on Hi4D val. The average time per example is 64 sec. for BUDDI vs. 89 sec. for BUDDI+ProsePose. The time to sample 20 programs from GPT-4o, averaged over 30 examples, is 16 sec.

### 8.2.4 Variance across Camera Angles

We quantify the impact of the camera angle on LMM predictions by running GPT-4o (the gpt-4o-08-06 version) with each camera angle. The F1 ranges from 0.31 to 0.42, but the only cameras with F1 less than 0.37 are the front

LMM	Prompt Style	Ask for Left/right	P1/P2 Labeled?	Hi4D $\downarrow$	Flickr $\downarrow$
Heuristic	–	–	–	124	67
GPT4-V	Direct	No	No	83	58
GPT4-V	Direct	Yes	No	80	62
GPT4-V	Direct	Yes	Yes	82	65
GPT4-V	Caption	No	No	84	60
GPT4-o	Direct	No	No	84	57
LLaVA	Direct	No	No	86	67
LLaVA	Caption	No	No	89	61

Table 5. **LMM Analysis:** We compare the default LMM and prompt (line 2) with several variants. We consider two other LMMs: GPT4-o (line 6), which refers to the `gpt-4o-08-06` version, and LLaVA v1.6 34B [31] (lines 7-8), and three other prompt variants: asking for left/right labels on limbs (line 3); asking for left/right labels on limbs and labeling the people in the image (line 4); and first generating a pose caption, removing mentions of left/right; and then converting the caption to constraints with a text-only LM (lines 6, 8). We take  $N = 5$  samples from each LMM. Metric is PA-MPJPE on Hi4D/FlickrCI3D validation sets.

and back cameras. Some body parts are often occluded from these angles, which may explain the lower scores.

### 8.3. Failure cases

Figure 10 shows examples of two types of ProsePose failures: (1) incorrect chirality (example a) and (2) hallucination (examples b and c). In example (a), the top constraints are correct but without the chirality specified. The optimization then brings both hands of one person to roughly the same point on the other person’s waist, rather than positioning one hand on each hip. Similarly, both hands of the other person are positioned on the same shoulder of the first person. Examples (b) and (c) both show cases of hallucination. In example (b), the hand is predicted to touch the back rather than the hand. In example (c), the hand is predicted to touch the foot rather than the leg. Interestingly, in the yoga example, GPT4-V correctly predicts the name of the yoga pose in all 20 samples (“Parivrtta Janu Sirsasana”). However, it outputs a constraint between a hand and a foot, which is true in the standard form of this pose but not in the displayed form of the pose. Consequently, the optimization brings the left hand closer to the right foot than to the right knee.

Figure 11 shows an example in which the camera view affects GPT4-V’s predictions substantially.

Given that the chirality is an important issue, it is natural to consider a prompt that asks the LMM to specify a chirality for each limb. Table 5 shows that such a prompt does not outperform our default prompt. Figure 12 shows an example of a failure of the prompt that requests chirality. With the prompt that asks for chirality, GPT4-V incorrectly predicts that the right leg of one person is touching the left leg of the other person. With the default prompt, GPT4-V predicts

in one constraint set that one person’s legs and chest are touching the other’s waist and back, respectively (and empty constraint sets otherwise). In the prompt ablation study, we also consider a prompt in which the image is labeled with Person 1 and Person 2 and the prompt asks for left/right labels, and this prompt also does not outperform the default one. Figure 13 shows an example in which the alternative prompt leads to incorrect predictions from GPT4-V.

### 8.4. LLaVA Results

In this section, we provide test set results when LLaVA-NeXT 34B (i.e. LLaVA v1.6) [30] is used as the LMM. We use the caption-to-table prompting approach described in the prompt ablation study (§ 4.1). That is, we generate a caption from the LMM, and we feed the caption alone to GPT4 in order to convert it into a table of constraints. For the two-person case, the prompts are given in § 7.1.1. We use a temperature of 0.3 and top-p of 0.7 when sampling from LLaVA.

For the one-person case, we use the following prompt for LLaVA:

Describe the person’s pose.

We use the same prompt as above to rewrite the caption. We then use the following prompt to create the formatted table:

You are a helpful assistant. You will follow ALL rules and directions entirely and precisely.

Given a description of a yoga pose, create a Mark-down table with the columns “Body Part 1” and “Body Part 2”, listing the body parts of the person that are guaranteed to be in contact with each other, from the following list. ALL body parts that you list must be from this list. Body parts: “head”, “back”, “shoulder”, “arm”, “hand”, “leg”, “foot”, “stomach”, “butt”, “ground” Note that “back” includes the entire area of the back.

Include all contact points that are directly implied by the description, not just those that are explicitly mentioned. If there are no contact points between these body parts that the description implicitly or explicitly implies, your table should contain only the column names and no other rows.

First, write your reasoning. Then write the Mark-down table.

We set  $N = 5$  for these experiments. Since we change  $N$ , we also need to select appropriate thresholds  $f$  and  $t$ . As in the experiments with GPT4-V, we set  $t = N$  for all datasets except CHI3D. For CHI3D, we find on the validation set that



Figure 10. **Failure cases** We show examples in which ProsePose fails to output a semantically correct pose. The constraints shown are the top 3 constraints (or the total number of constraints, whichever is smaller) that meet the threshold  $f$  along with their counts ( $f = 1$  for two-person experiments and  $f = 10$  for the one-person experiment).

Original Image

Predicted Constraints



*Hand, Back*  $\times 20$



*Arm, Hand*  $\times 4$   
*Shoulder (front), Hand*  $\times 9$   
*Back, Hand*  $\times 5$   
*Shoulder (back), Hand*  $\times 1$

Figure 11. **Variation due to camera angle** Example from CHI3D in which GPT4-V outputs substantially different constraint sets for different views of the same pose.

$t = 2$  works better than  $t = 1$ , so we set  $t = 2$ . As in the experiments with GPT4-V, we set  $f = 1$  for the 2-person datasets, and we set  $f = 3$  for MOYO, to approximate the ratio  $f/N$  used in the GPT4-V experiments. Finally, when

converting the constraint pairs to loss functions, we found that on a small number of examples, the pipeline produced a large number of constraints, leading to very slow loss functions. Therefore, we discarded loss functions that are

Original Image



Default Prompt



*Legs, waist  $\times 1$   
Back, chest  $\times 1$*

Prompt asking for  
Left/right



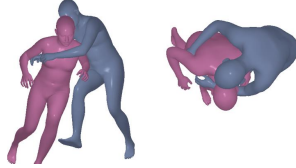
*Right leg, left leg  $\times 2$   
Back, stomach  $\times 1$*

Figure 12. **Failure of left/right prompt** In this example, GPT4-V responds with an incorrect chirality when asked to specify the chirality of the limbs in the constraints. The last column shows the unordered valid region pairs occurring in the 5 samples from GPT4-V along with the number of samples in which the pair occurs.

Original Image



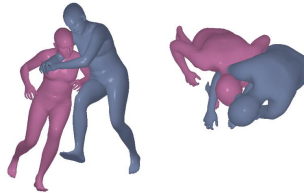
Default Prompt



**P1/P2 not pre-assigned**

1. (Arm, Waist (back)), (Shoulder (front), Chest), (Head, Head)
2. (Waist (back), Arm), (Arm, Shoulder (front)), (Back, Chest)
3. (Arm, Arm), (Waist (front), Chest)
4. (Chest, Shoulder (front)), (Stomach, Chest), (Arm, Back)
5. (Chest, Arm), (Stomach, Arm), (Chest, Hand), (Stomach, Hand)

Labeled Image w/  
Prompt asking for  
Left/right



**P1/P2 pre-assigned**

1. (R arm, R Shoulder (front)), (R shoulder (front), L Shoulder (front)), (R Leg, L Leg), (L hand, L hand)
2. (R Arm, L Shoulder (front)), (L Shoulder (front), R Shoulder (front)), (R Hand, L Arm), (Waist (back), R Arm)
3. (R Arm, Waist (front)), (L Arm, Back)
4. (R Arm, L Shoulder (front)), (L Arm, R Shoulder (front)), (Waist (front), R Arm)
5. (R Arm, L Shoulder (front)), (Waist (back), R Hand)

Figure 13. **Failure of labeled image with left/right prompt** In this example, the LMM responds with incorrect constraints when given an image in which persons 1 and 2 are labeled and asked to specify the chirality of the limbs in the constraints. The last column shows the constraint sets in the 5 LMM samples. For the default prompt, since persons 1 and 2 are not labeled in the image, the minimum loss from the two labelings is used, while for the other prompt, only the loss corresponding to the given labeling is considered.

longer than 10000 characters.

Table 6 shows the results. On the 2-person datasets, the LLaVA+GPT4 approach performs better than the contact heuristic but not as well as GPT4-V. This is in line with holis-

tic multimodal evaluations that indicate that GPT4-V performs better than LLaVA [32]. On the 1-person yoga dataset, the performance of LLaVA+GPT4 is comparable with that of the baseline (HMR2+opt). The reason that LLaVA per-

	Hi4D		FlickrCI3D			CHI3D			MOYO		
	Err $\downarrow$	F1 $\uparrow$	Err $\downarrow$	PCC $\uparrow$	F1 $\uparrow$	Err $\downarrow$	PCC $\uparrow$	F1 $\uparrow$	Err $\downarrow$	PCC $\uparrow$	F1 $\uparrow$
Heuristic	116	–	67	77.8	–	105	74.1	–	–	–	–
HMR2+opt	–	–	–	–	–	–	–	–	81	85.2	–
GPT4-V	93	24	58	79.9	13	100	75.8	23	82	87.8	25
LLaVA+GPT4	95	22	60	79.7	7	101	75.2	13	82	85.2	14

Table 6. **LLaVA Results.** Err denotes Joint PA-MPJPE for the two-person datasets (Hi4D, FlickrCI3D, CHI3D) and PA-MPJPE for MOYO. Lower is better for Err, and higher is better for Avg. PCC. Note that the GPT4-V results use 20 samples from the LMM, while the LLaVA results use 5 samples from the LMM.

forms worse than GPT4-V in this setting may be that LLaVA does not have enough training data on yoga to provide useful constraints.

### 8.5. Additional Qualitative Results

Figures 14, 15, 16, and 17 show additional, randomly selected examples from the multi-person FlickrCI3D test set. Figures 18, 19, 20, and 21 show the same examples comparing ProsePose with the pseudo-ground truth fits. Figures 22, 23, and 24 show additional, randomly selected examples from the Hi4D test set. Figures 25 and 26 show additional, randomly selected examples from the CHI3D validation set (which we use as the test set following [34]). Figures 27 and 28 show additional, randomly selected examples from the 1-person yoga MOYO test set.



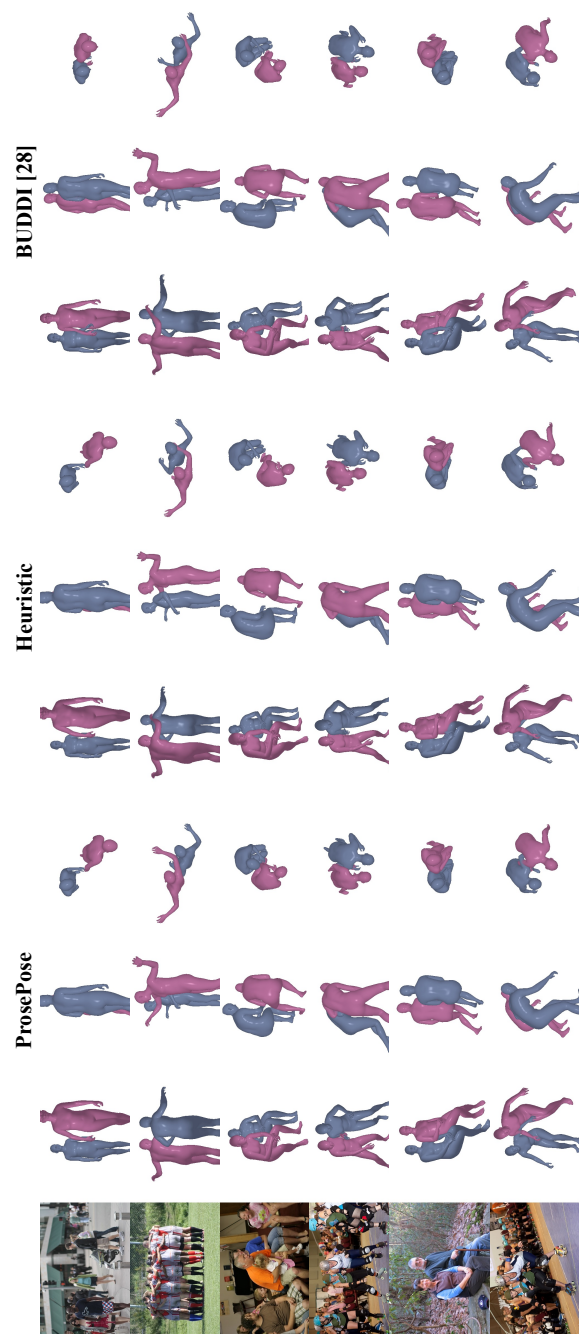


Figure 14. Non-curated examples from the FlickrCI3D test set. They are randomly selected from the examples for which there is at least one non-empty constraint set.

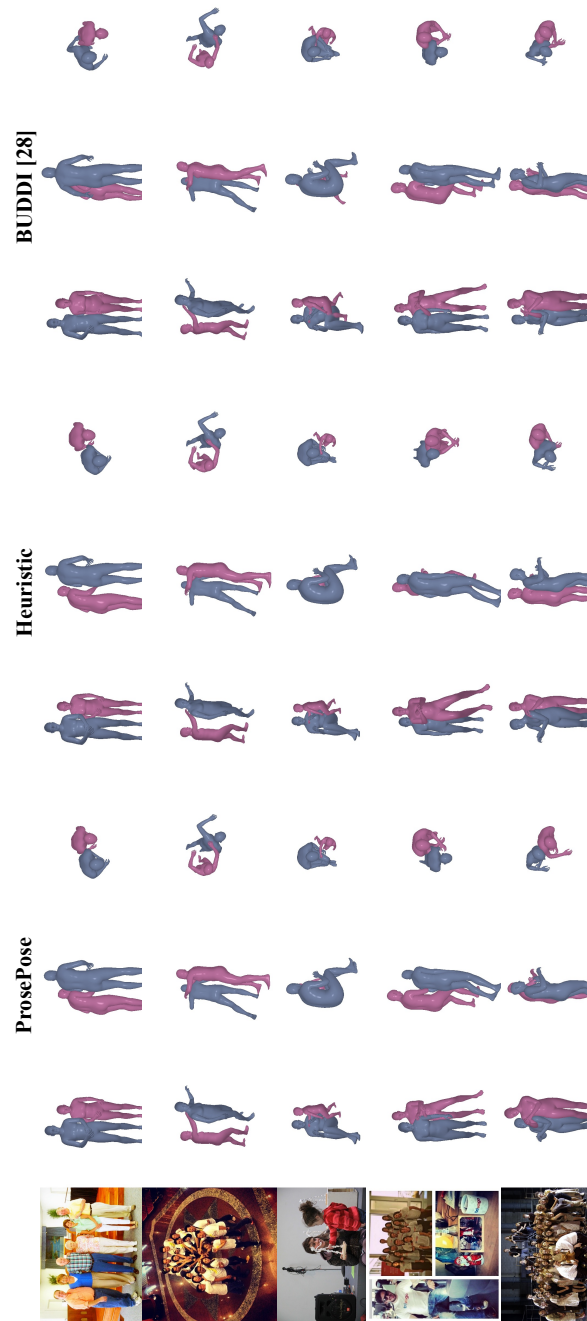


Figure 15. Non-curated examples from the FlickrCI3D test set. They are randomly selected from the examples for which there is at least one non-empty constraint set.

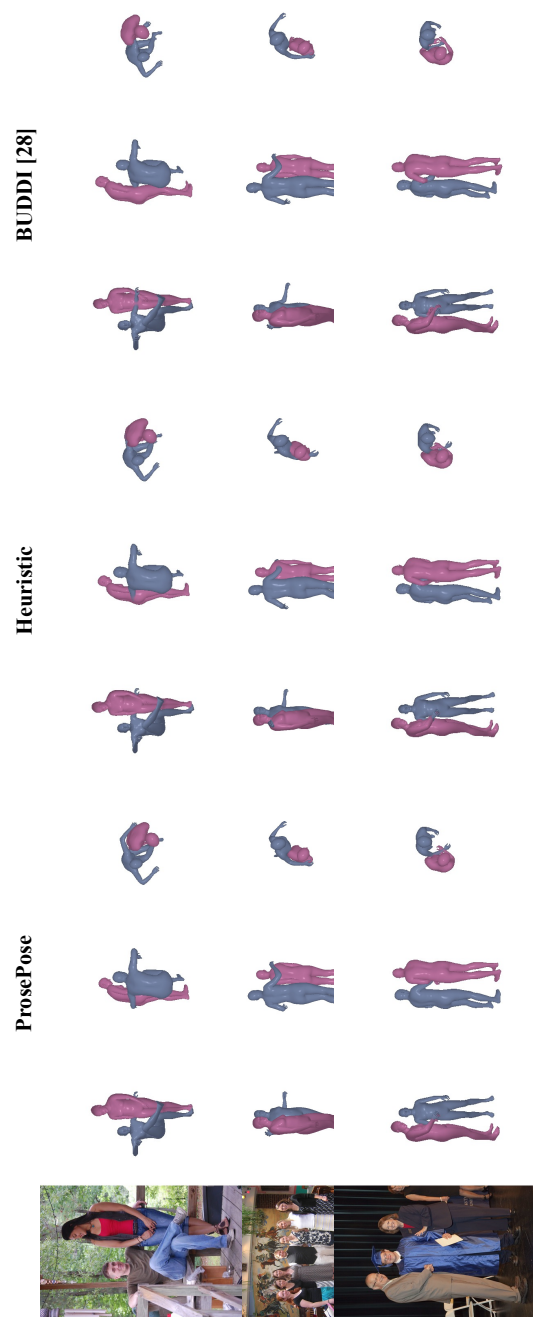


Figure 16. Non-curated examples from the FlickrCI3D test set. They are randomly selected from the examples for which there is at least one non-empty constraint set.

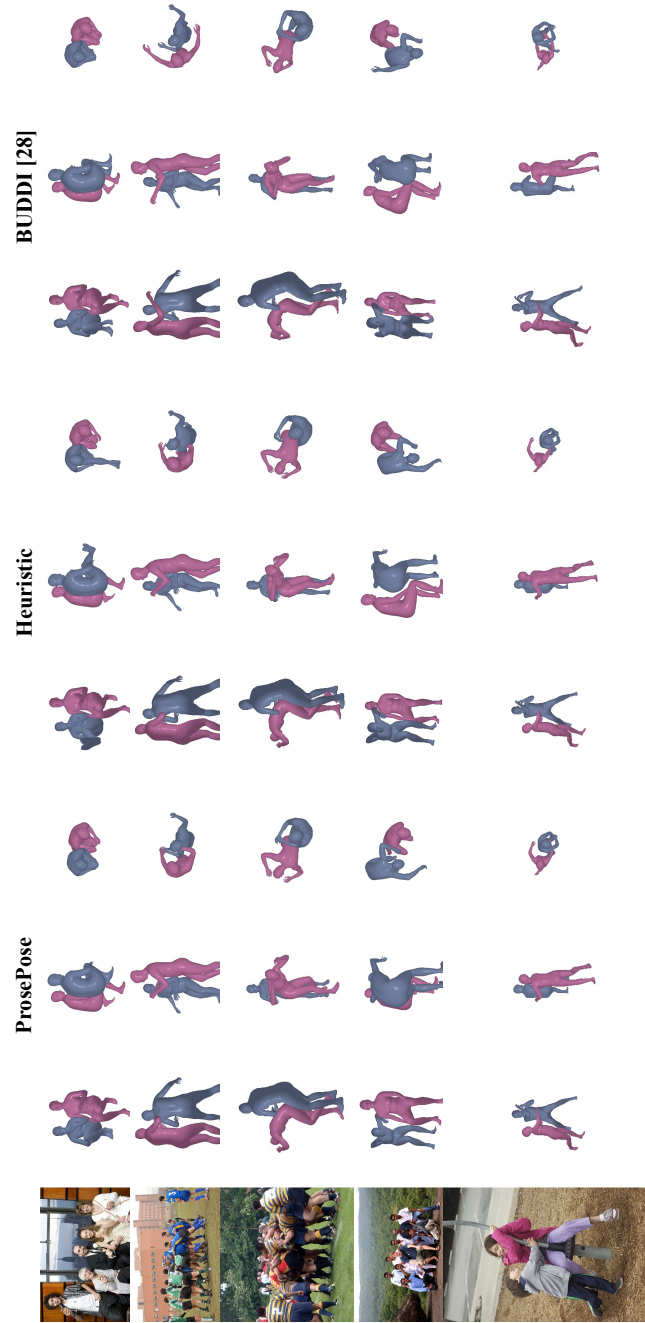


Figure 17. Non-curated examples from the FlickrCI3D test set. They are randomly selected from the examples for which there is at least one non-empty constraint set.

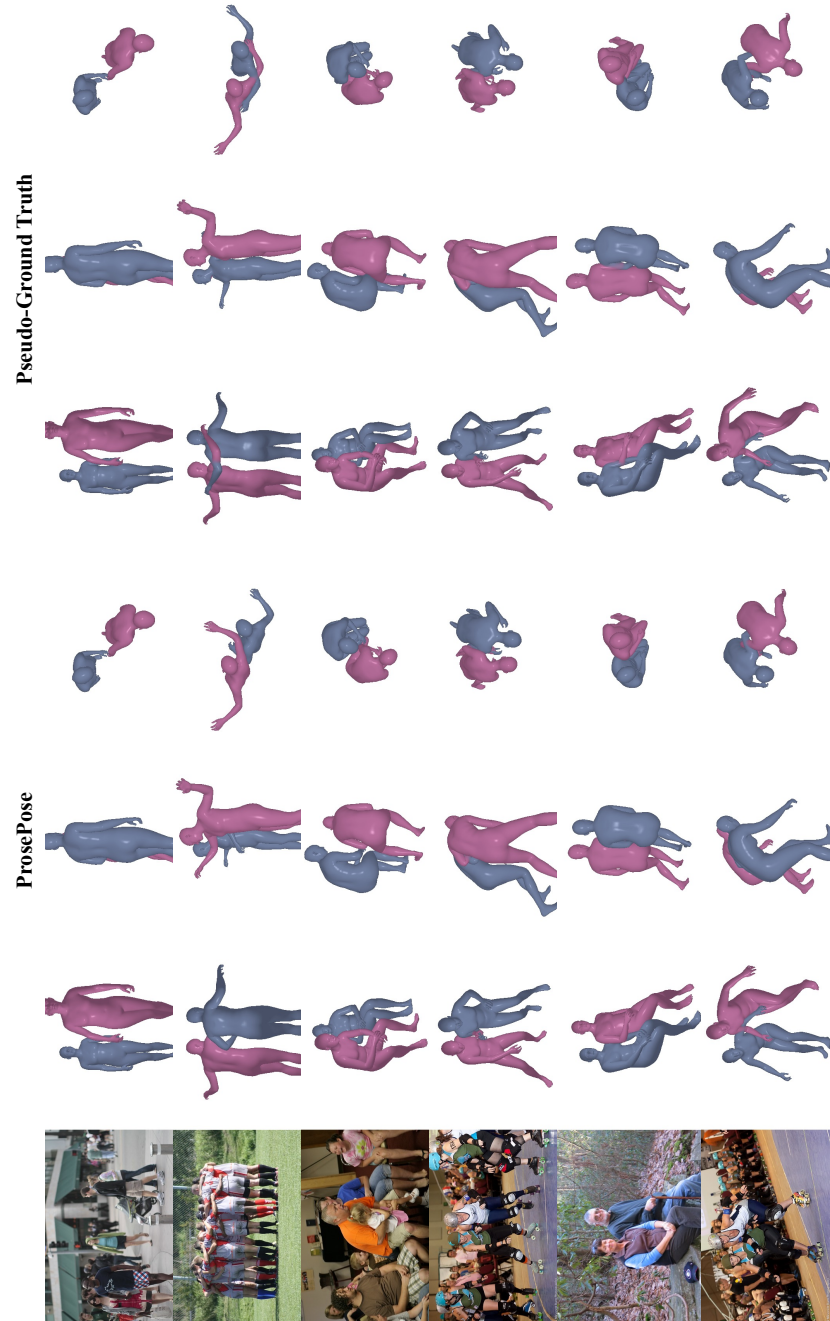


Figure 18. Non-curated examples from the FlickrCI3D test set, comparing ProsePose with the pseudo-ground truth fits. They are randomly selected from the examples for which there is at least one non-empty constraint set.



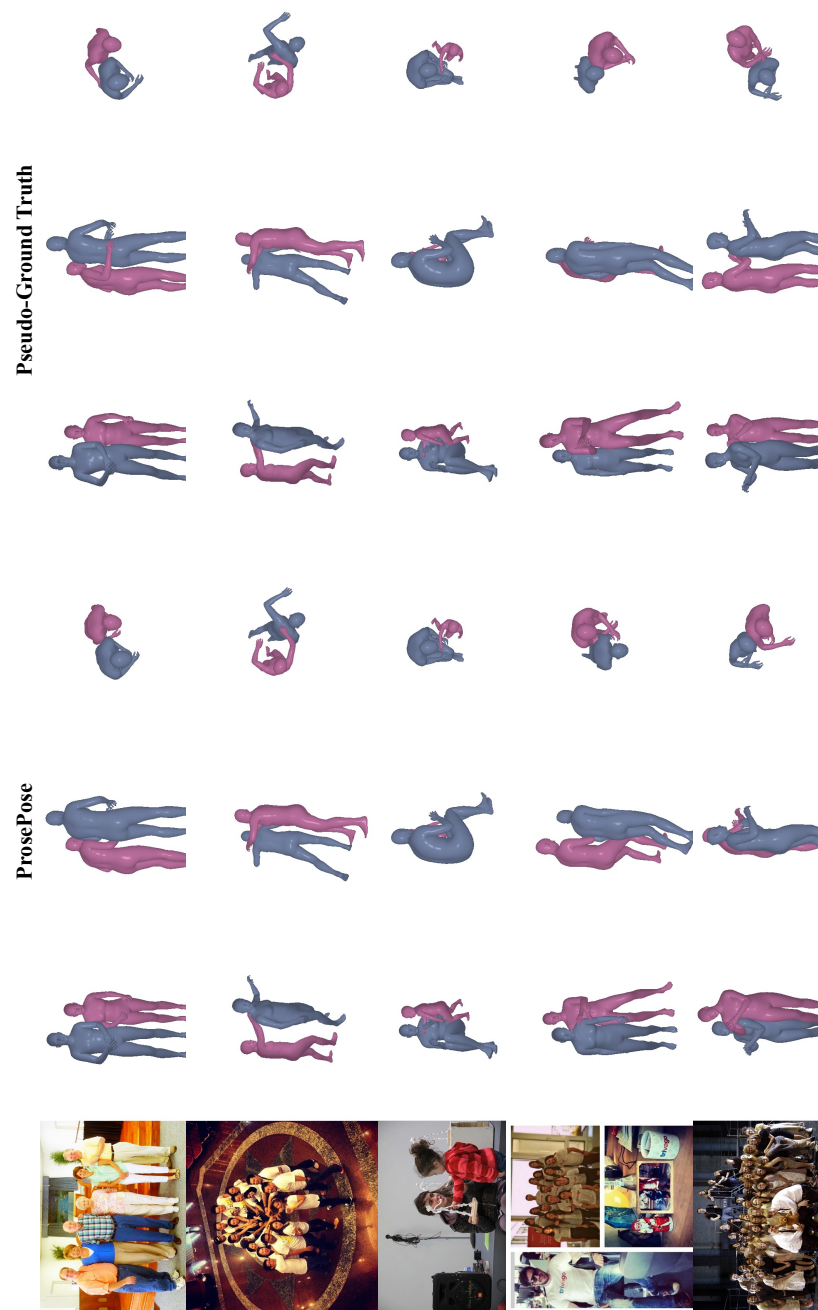


Figure 19. Non-curated examples from the FlickrCI3D test set, comparing ProsePose with the pseudo-ground truth fits. They are randomly selected from the examples for which there is at least one non-empty constraint set.

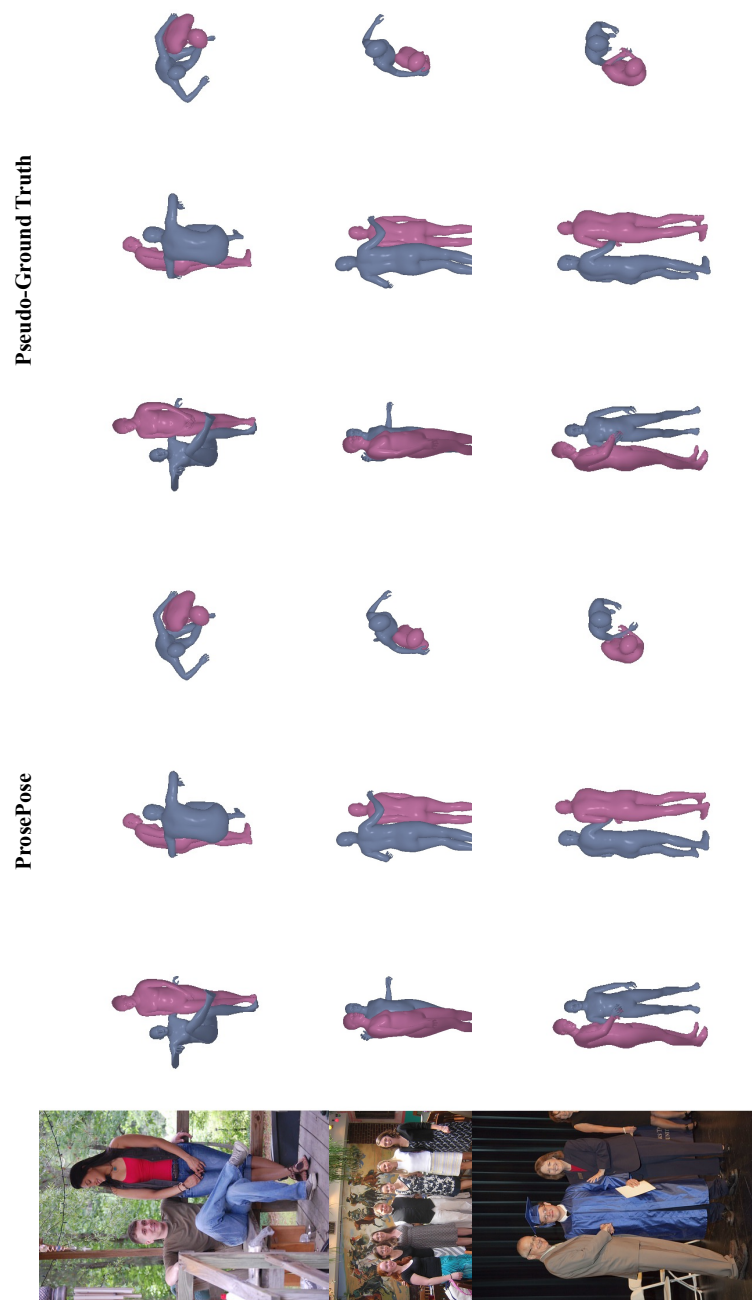


Figure 20. Non-curated examples from the FlickrCI3D test set, comparing ProsePose with the pseudo-ground truth fits. They are randomly selected from the examples for which there is at least one non-empty constraint set.

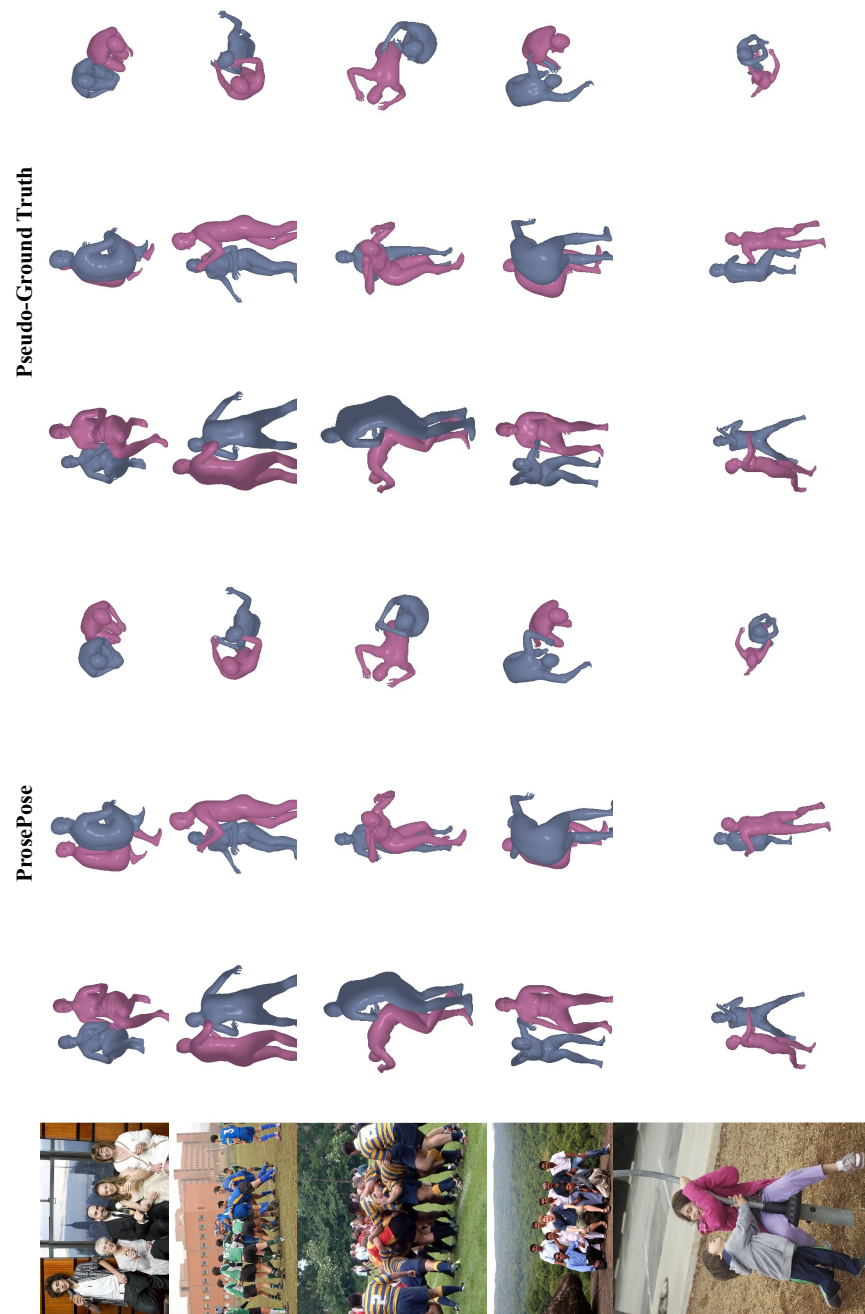


Figure 21. Non-curated examples from the FlickrCI3D test set, comparing ProsePose with the pseudo-ground truth fits. They are randomly selected from the examples for which there is at least one non-empty constraint set.

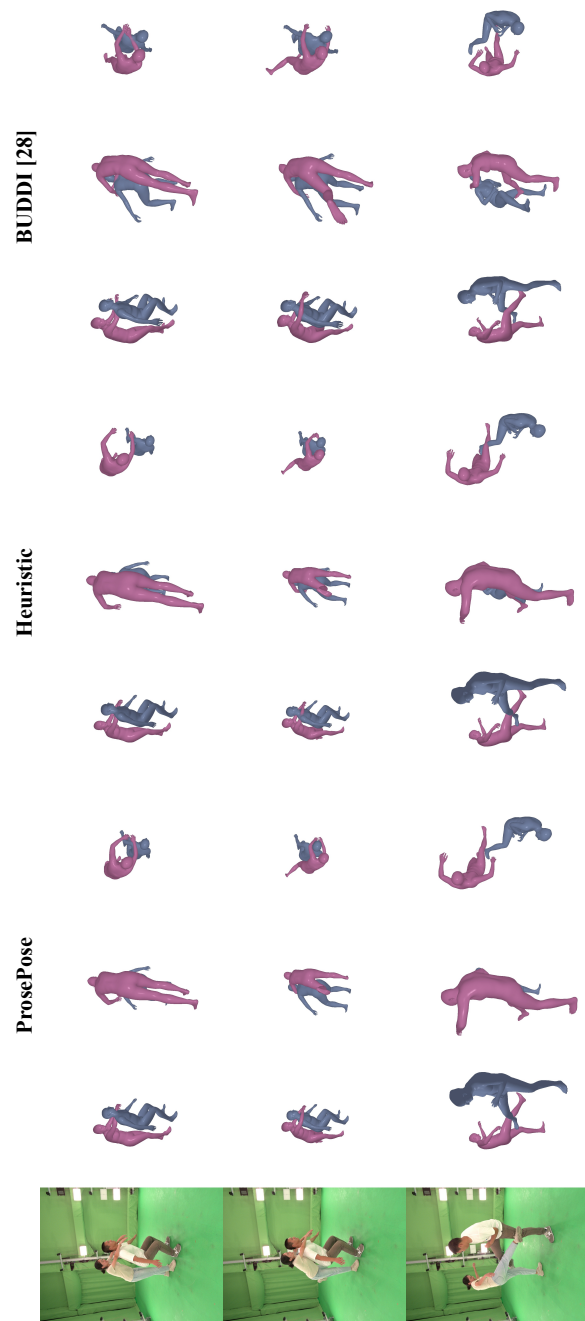


Figure 22. Non-curated examples from the Hi4D test set. They are randomly selected from the examples for which there is at least one non-empty constraint set.

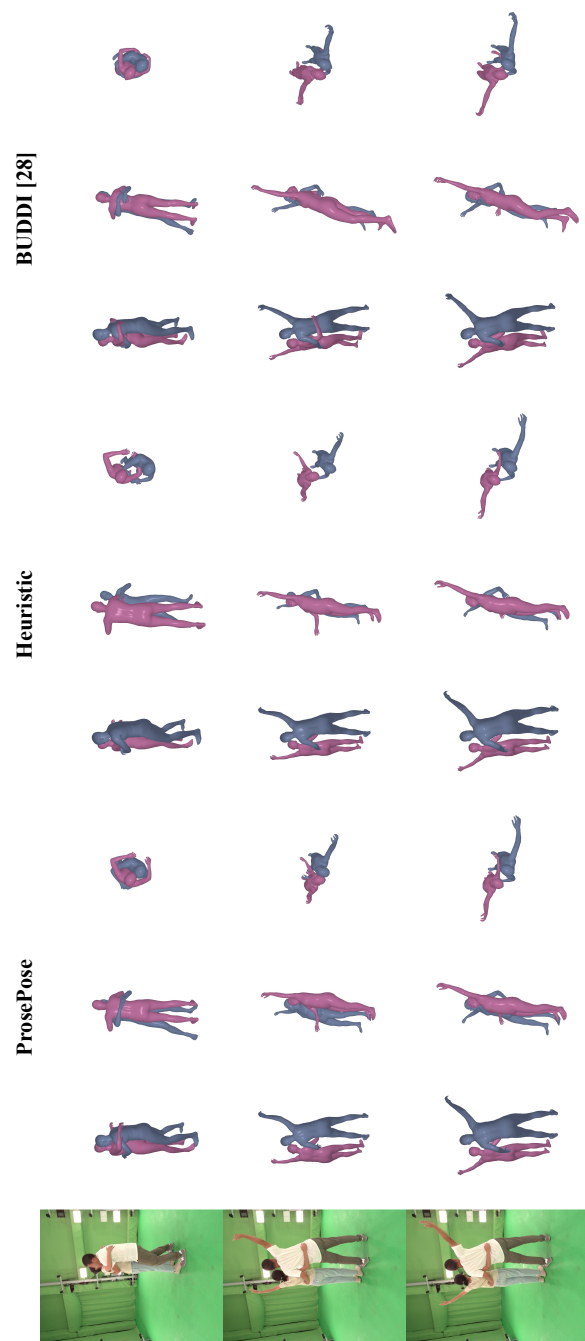


Figure 23. Non-curated examples from the Hi4D test set. They are randomly selected from the examples for which there is at least one non-empty constraint set.



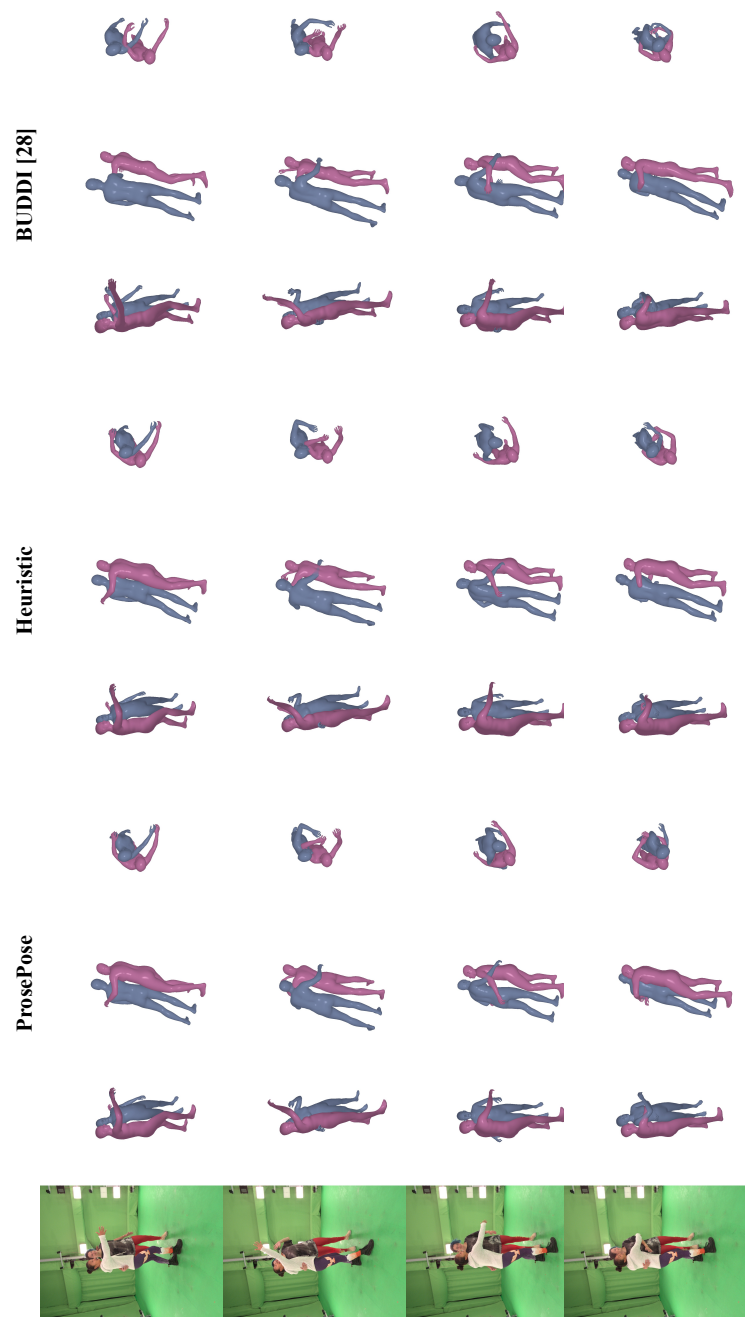


Figure 24. Non-curated examples from the Hi4D test set. They are randomly selected from the examples for which there is at least one non-empty constraint set.

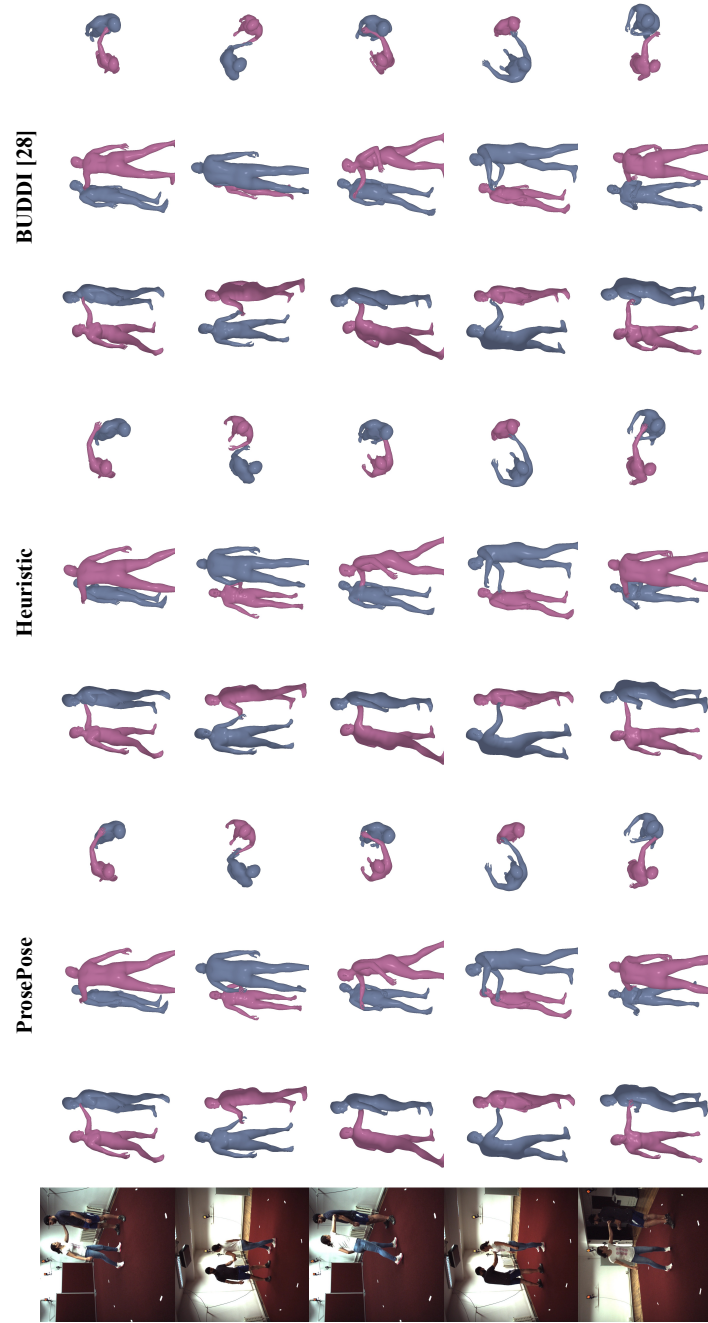


Figure 25. Non-curated examples from the CHI3D validation set (which we use as the test set). They are randomly selected from the examples for which there are at least nineteen non-empty constraint sets (since we set  $t = 2$  for CHI3D).

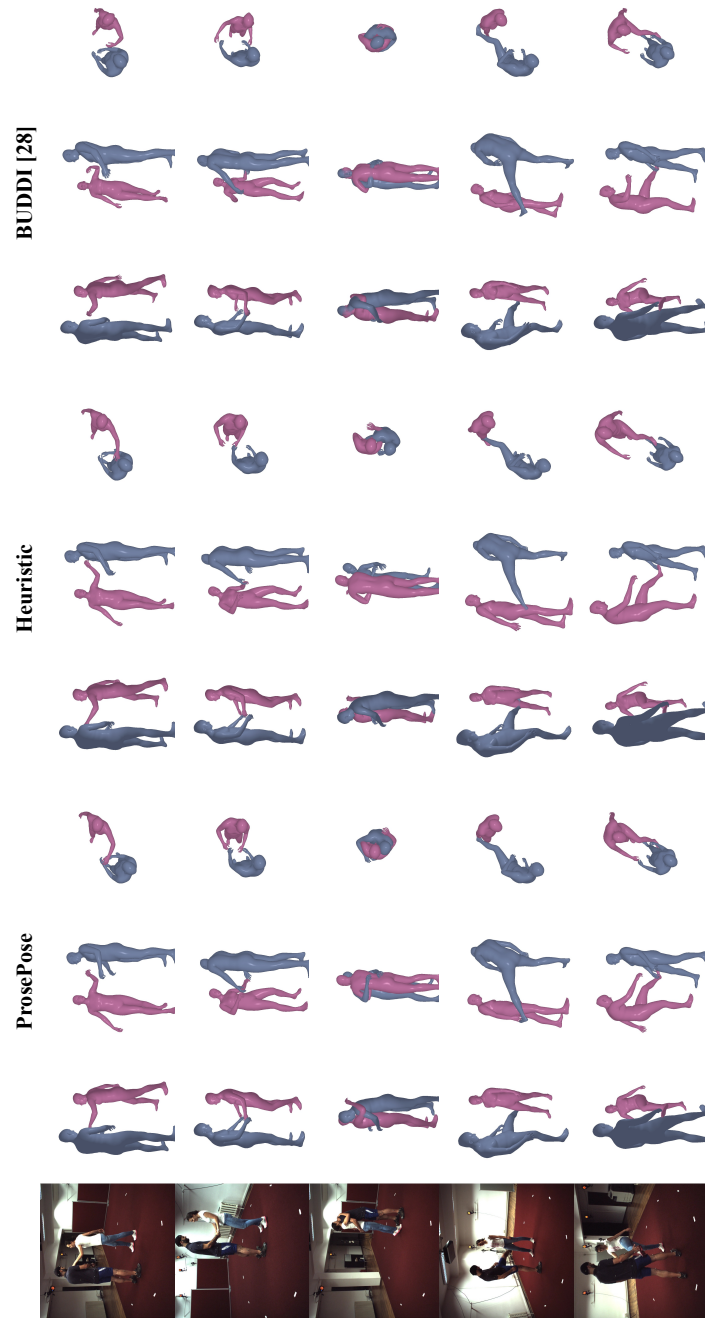


Figure 26. Non-curated examples from the CHI3D validation set (which we use as the test set). They are randomly selected from the examples for which there are at least nineteen non-empty constraint sets (since we set  $t = 2$  for CHI3D).



Figure 27. Non-curated examples from the MOYO test set. They are randomly selected from the examples for which there is at least one non-empty constraint set.

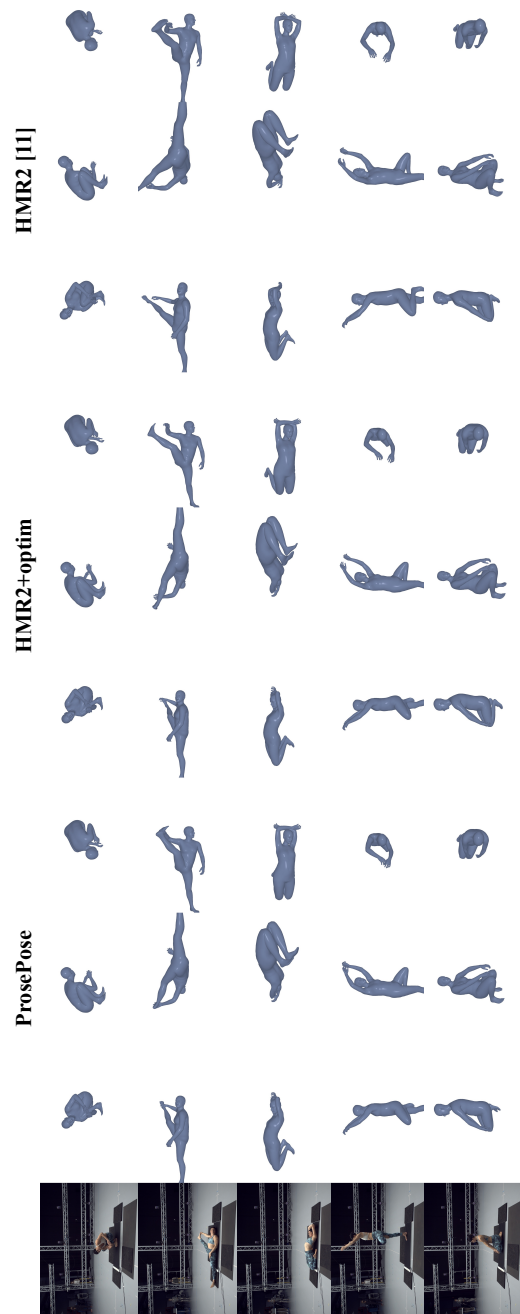


Figure 28. Non-curated examples from the MOYO test set. They are randomly selected from the examples for which there is at least one non-empty constraint set.