

LANGUAGE-DRIVEN 3D HUMAN POSE ESTIMATION: GROUNDING MOTION FROM TEXT DESCRIPTIONS —— SUPPLEMENTARY MATERIAL ——

Anonymous authors

Paper under double-blind review

CONTENTS

1	Significance of L3DHPE Task	2
1.1	Practical Applications in Real Life	3
1.2	Comparison with Similar Tasks	3
2	More Details about the CPP Model	3
2.1	Loss Function of CPP	3
2.2	Detailed Structure of the CPP Model	4
3	More Details of the Panoptic-L3D Dataset	5
3.1	Details of Video Collection	5
3.2	Annotation Tools	6
3.3	Data License and Access Details	7
4	More Details of Experiments and Visualization	7
4.1	Evaluation on Different Backbone Networks	7
4.2	Sensitivity Analysis on Hyperparameters	8
4.3	Discussion of Different Metrics	8
4.4	Visualizations of Motion Grounding.	8
5	Datasheet of the Panoptic-L3D Dataset	10
5.1	Motivation	10
5.2	Composition	11
5.3	Collection Process	12
5.4	Preprocessing/cleaning/labeling	14
5.5	Uses	15
5.6	Distribution	15
5.7	Maintenance	15

1 SIGNIFICANCE OF L3DHPE TASK

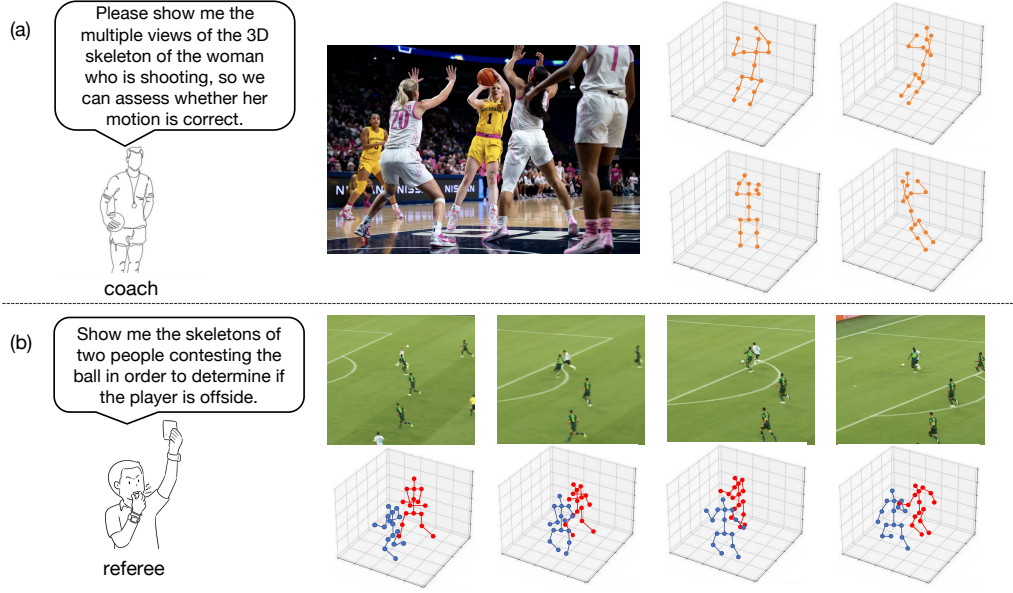


Figure 1: Practical applications examples of L3DHPE task in real life. (a) Example of sports coaching. (b) Examples of sports officiating.

Language-Driven 3D Human Pose Estimation (L3DHPE), alternatively termed *3D Motion Grounding*, extends the conventional 3D human pose estimation (3DHPE) problem to encompass complex multi-person scenes. Compared with traditional 3DHPE tasks that can only estimate all individuals appearing in the video, L3DHPE emphasizes more on the ability to select specific individuals via full explorations of semantic interactions between human poses and linguistic expressions. In this case, the L3DHPE targets a more demanding yet practical problem of reconstructing 3D pose sequences for individuals, guided by comprehensive language descriptions that include appearance, behavior, and body movements. As a novel research topic, our proposed L3DHPE task serves as a fundamental problem to be solved in various downstream applications, including:

- **Human-Computer Interaction (HCI):** Facilitating systems that can interpret human actions through verbal commands for more intuitive and natural interactions.
- **Robotics:** Enhancing robots' capabilities to comprehend and simulate human movements as articulated through language, thus improving adaptability in complex real-world contexts.
- **Animation and Gaming:** Automating the creation of 3D character animations derived from textual descriptions, thereby streamlining content generation for games and virtual environments.
- **Surveillance and Security:** Advancing behavior recognition techniques by enabling systems to align language-based descriptions with 3D pose estimates in security applications.
- **Healthcare:** Supporting physical therapy by reconstructing and analyzing patient poses based on verbal instructions or reports, thereby aiding therapeutic interventions.

The introduction of L3DHPE creates new opportunities for advancing research in the domain of 3D human pose estimation, including the alignment of visual, pose, and textual features to enhance the utilization of both detailed and semantic information, the differentiation of similar human poses, and the integration of human pose with semantic motion. Our proposed L3DHPE task, along with the newly established Panpotic-L3D dataset and the CPP approach, provides a foundational framework for these explorations.

1.1 PRACTICAL APPLICATIONS IN REAL LIFE

In the following section, we will provide two case studies of L3DHPE as concrete examples of utilizing L3DHPE in real-world applications.

Sports Coaching. As illustrated in Fig. 1(a), coaches must often visualize multiple perspectives of individual players. Displaying each player independently based on natural language input offers considerable convenience. By combining language comprehension with 3D pose reconstruction, our task could offer potentially real-time assessment of player movements by coaches, trainers, and analysts, thereby facilitating precise feedback and enhancing athletic performance.

Sports Officiating. Fig. 1(b) presents an example of sports officiating. In this scenario, a referee requires the skeletal representations of two players to evaluate an offside decision. While current systems can automatically detect all players’ skeletons, identifying the relevant individuals remains time-consuming. The L3DHPE task leverages language-guided selection to estimate the target players, disregarding irrelevant ones, thereby streamlining the decision-making process and reducing delays during the game.

1.2 COMPARISON WITH SIMILAR TASKS

We provide a detailed comparison with three closely related tasks: 3D human pose estimation, human pose recognition, and human pose generation, to highlight the distinctive contributions and the necessity of our proposed L3DHPE task.

3D Human Pose Estimation (3DHPE) Pavlakos et al. (2017); Pavlo et al. (2019); Sun et al. (2017); Zheng et al. (2021); Sun et al. (2022); Su et al. (2022) aims at estimating the 3D pose of all individuals present in images or videos. However, it cannot selectively focus on specific individuals, limiting its applicability in real-world scenarios where targeted pose analysis is essential. In contrast, our L3DHPE task incorporates natural language guidance to identify and estimate the pose of referred individuals, providing a more context-aware understanding of human activity in complex environments.

Human Pose Recognition Zhang et al. (2020); Mazzia et al. (2022); Chi et al. (2022) involves categorizing human poses into predefined motion classes. However, as a classification task, these methods can only select from a fixed set of labels corresponding to known verbs or 3D pose sequences. This restricts their ability to comprehend the diverse and intricate actions encountered in real-world contexts. Moreover, such methods are unsuitable for interpreting the natural, everyday language used to describe actions and are limited in representing complex action combinations. In contrast, our L3DHPE task tackles a more advanced challenge of reconstructing 3D pose sequences based on detailed linguistic descriptions, enabling a broader and more flexible understanding of human behavior.

Human Pose Generation Delmas et al. (2024); Lin et al. (2024); Feng et al. (2024) focuses on generating human poses from textual descriptions or generating descriptions based on a given pose. These methods are inherently generative and differ fundamentally from our grounding-based approach in the L3DHPE task. Furthermore, they typically involve one or two individuals in static scenes, which does not reflect the complexities of real-world environments. In contrast, L3DHPE is designed to ground 3D poses based on textual descriptions in dynamic video contexts, often involving multiple interacting individuals. Additionally, while human pose generation tasks typically rely on virtual or synthetic environments, our L3DHPE task processes video input from intricate, real-world settings.

2 MORE DETAILS ABOUT THE CPP MODEL

In this section, we present more details of our proposed method CPP, including more details of our loss function and two modules, i.e., Body Fusion Block and Mask Fusion Block.

2.1 LOSS FUNCTION OF CPP

In stage 1, the CPP generates bounding box maps $M_b \in \mathbb{R}^{W \times H \times 4}$, root depth maps $M_r \in \mathbb{R}^{W \times H}$, and 2D joint maps $M_j \in \mathbb{R}^{W \times H \times N_j}$ for all individuals in the frame. Here, N_j denotes the number

of human joints, while W and H represent the width and height of the output maps, respectively. These maps are then compared with the corresponding ground-truth maps, denoted as \widetilde{M}_b , \widetilde{M}_r , and \widetilde{M}_j . The losses for bounding box maps, root depth maps, and 2D joint maps are defined as \mathcal{L}_{box} , \mathcal{L}_{depth} , and \mathcal{L}_{joint} , respectively, and are calculated as follows:

$$\mathcal{L}_{box} = \sum_{p \in \mathcal{P}} \|M_b^p - \widetilde{M}_b^p\|_1, \quad (1)$$

$$\mathcal{L}_{depth} = \sum_{p \in \mathcal{P}} |M_r^p - \widetilde{M}_r^p|, \quad (2)$$

$$\mathcal{L}_{joint} = \|M_j - \widetilde{M}_j\|_2^2, \quad (3)$$

where \mathcal{P} represents the set of ground-truth root joint locations. The total loss for stage 1, \mathcal{L}_1 , is a weighted sum of these individual losses:

$$\mathcal{L}_1 = \lambda_{box}\mathcal{L}_{box} + \lambda_{depth}\mathcal{L}_{depth} + \lambda_{joint}\mathcal{L}_{joint}, \quad (4)$$

where λ_{box} , λ_{depth} , and λ_{joint} are the respective hyperparameters.

In stage 2, the CPP generates a referring mask M , the 3D positions of the human root nodes H_r , and the 3D human skeleton H_s . These outputs are then compared with the ground-truth values, denoted as \widetilde{M} , \widetilde{H}_r , and \widetilde{H}_s . The losses for the referring mask, human root nodes, and 3D human skeleton are denoted as \mathcal{L}_{mask} , \mathcal{L}_{root} , and \mathcal{L}_{cord} , respectively. The loss \mathcal{L}_{mask} is a combination of the DICE loss Milletari et al. (2016) and the binary mask focal loss, while \mathcal{L}_{root} and \mathcal{L}_{cord} are defined as:

$$\mathcal{L}_{root} = \|H_r - \widetilde{H}_r\|_2^2, \quad (5)$$

$$\mathcal{L}_{cord} = \frac{1}{N_j} \sum_{k=1}^{N_j} \|H_s^k - \widetilde{H}_s^k\|_1, \quad (6)$$

The total loss for stage 2, \mathcal{L}_2 , is computed as follows:

$$\mathcal{L}_2 = \lambda_{mask}\mathcal{L}_{mask} + \lambda_{root}\mathcal{L}_{root} + \lambda_{cord}\mathcal{L}_{cord}, \quad (7)$$

where λ_{mask} , λ_{root} , and λ_{cord} are the corresponding hyperparameters.

2.2 DETAILED STRUCTURE OF THE CPP MODEL

Details of the Body Fusion Block. With the input of Body Fusion Block as a visual feature from image encoder $x_v \in \mathbb{R}^{N \times D}$, text feature from text encoder $x_t \in \mathbb{R}^{L \times D}$ and 2D joints map from pose encoder $m_j \in \mathbb{R}^{W \times H \times N_j}$, where $N = HW/P^2$ stands for the number of patches (tokens) in the Video Swin transformer Liu et al. (2022), N_j represents the number of human joints, and L stands for the total length of the text and embedding context. We show the details of the Body Fusion Block as follows. First, the 2D joints map is reshaped and then passed through a linear layer to align with visual and text features $m'_j \in \mathbb{R}^{N \times D}$. Then following Luo et al. (2024), we utilize the MHA module to interact with different kinds of features. Specifically, m'_j and x_v are put into the first MHA module and produce $x_{vj} \in \mathbb{R}^{N \times D}$. And then x_{vj} and x_t are put into the second MHA module and produce $x_{vjt} \in \mathbb{R}^{N \times D}$. This process can be formulated as follows:

$$m'_j = \text{Linear}(\text{reshape}(m_j)), m'_j \in \mathbb{R}^{N \times D} \quad (8)$$

$$x_{vj} = \text{MHA}(m'_j, x_v), x_{vj} \in \mathbb{R}^{N \times D} \quad (9)$$

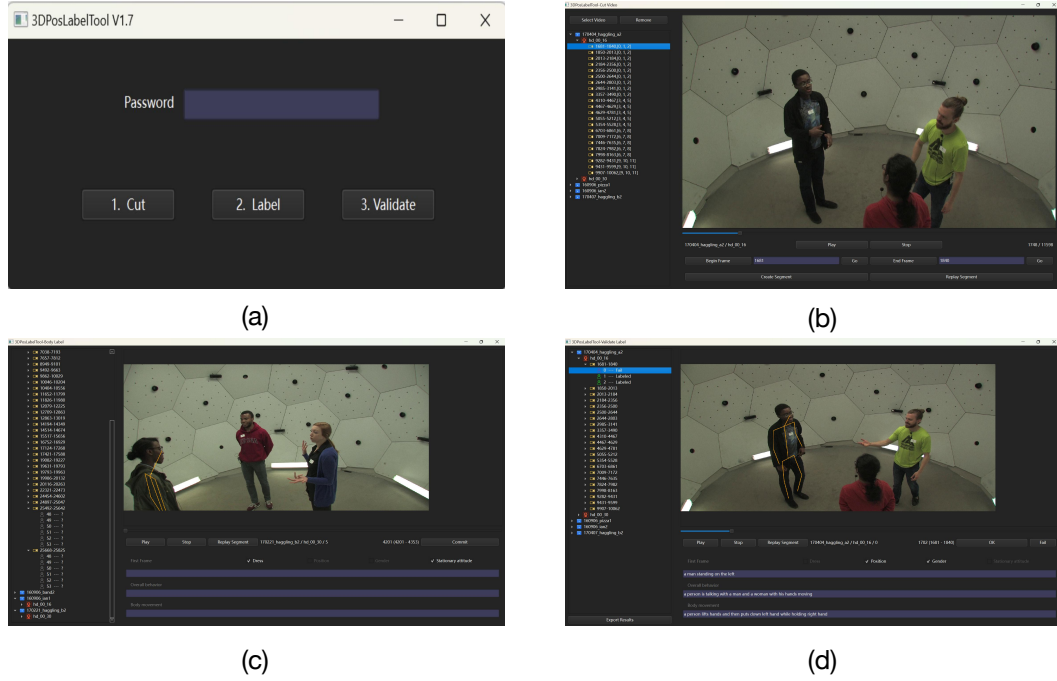


Figure 2: Screenshots of our sentence annotation system.

$$x_{vjt} = MHA(m_{vj}, x_t), x_{vjt} \in \mathbb{R}^{N \times D} \quad (10)$$

Details of the Mask Fusion Block. For Mask Fusion Block, the input features are 2D joints map $m_j \in \mathbb{R}^{W \times H \times N_j}$ and masks $m \in \mathbb{R}^{W \times H}$. We show the details of the Mask Fusion Block as follows. First, we concat the 2D joints map and masks together to $x_c \in \mathbb{R}^{W \times H \times (N_j+1)}$. Then, we utilize multiple downsample layers, including 2d convolutional layers, batchnorm, and Relu activation, to downsample the features, resulting in $X_d = \{x_d^i\}, x_d^i \in \mathbb{R}^{\frac{W}{2^i} \times \frac{H}{2^i} \times ((N_j+1) \times 2^i)}$. There are 3 downsample layers, so $i \in \{1, 2, 3\}$. Next, we utilize multiple upsample layers, including 2d transpose convolutionary layers, batchnorm, and Relu activation. For each upsampling layer, we first pass through upsampling layers and result in $X_u = \{x_u^i\}, x_u^i \in \mathbb{R}^{\frac{W}{2^i} \times \frac{H}{2^i} \times ((N_j+1) \times 2^i)}$ and then concat corresponding upsampling features. Finally, we utilize batchnorm and linear layers to further process the features to $x_{out} \in \mathbb{R}^{W \times H \times N_j}$. This process can be formulated as follows:

$$x_c = \text{concat}(m_j, m), x_c \in \mathbb{R}^{W \times H \times (N_j+1)} \quad (11)$$

$$x_d^i = \text{Relu}(\text{BN}(\text{Conv}(x_c))), x_d^i \in \mathbb{R}^{\frac{W}{2^i} \times \frac{H}{2^i} \times ((N_j+1) \times 2^i)}, i \in \{1, 2, 3\} \quad (12)$$

$$x_u^{i-1} = \text{concat}(\text{Relu}(\text{BN}(\text{Conv}(x_u^i))), x_d^i), x_u^{i-1} \in \mathbb{R}^{\frac{W}{2^{i-1}} \times \frac{H}{2^{i-1}} \times ((N_j+1) \times 2^{i-1})}, i \in \{1, 2, 3\} \quad (13)$$

$$x_{out} = \text{BN}(\text{Linear}(x_u^0)), x_{out} \in \mathbb{R}^{W \times H \times N_j} \quad (14)$$

3 MORE DETAILS OF THE PANOPTIC-L3D DATASET

3.1 DETAILS OF VIDEO COLLECTION

In Sec.3 of the main paper, we mentioned that we selected fourteen activity entries featuring more than one person. The specific entries are: 160422_haggling1, 160226_haggling1, 160224_haggling1, 170404_haggling_a2, 170407_haggling_b2, 170221_haggling_b2, 160422_ultimatum1, 160906_band1, 160906_band2, 160906_band3, 160906_pizza1, 160906_ian1, 160906_ian2, and 160906_ian3.

Mask Annotation System

Current Number of Images

1103 / 6029

Current Number of Person

3

Mask Selection

☐ 1
 ☒ 2
 ☐ 3

Last Image

Image

Ok

Fail

Image

Point prompt

☒ Positive
 ☐ Negative

Clear Points

Confirm Points

Figure 3: Screenshots of our mask annotation system.

3.2 ANNOTATION TOOLS

We designed a sentence annotation system for human description annotation. Fig. 2 shows the various stages of this system. Panel (a) shows the login page. We have three types of pages corresponding to our three annotation processes: video clip creation, sentence annotation, and sentence validation, as shown in panels (b), (c), and (d) respectively.

Annotators begin by cutting the videos into clips, as shown in (b). They can select the start and end frames of the clips and replay them to ensure accuracy. Next, annotators annotate the description of every individual who meets the requirements outlined in Sec.3 in the main paper, as shown in (c). Finally, the validation page (d) presents the original video and its corresponding descriptions to the first validator, who must identify the target person referred to by the statements. If the chosen target matches the annotator’s target, the annotated video is retained. If there is a discrepancy, the system sends the descriptions to a second validator for further review.

Table 1: Quantitative evaluation on different backbone networks.

id	Encoder Type			PICI			MPJPE-L		
	Text	Video	Image	App.	Beh.	Mov.	App.	Beh.	Mov.
1	Bert	-	-	41.5%	40.4%	40.3%	883.2	911.9	913.6
2	-	I3D	-	41.0%	40.0%	39.9%	895.5	926.2	929.4
3	-	-	Resnet-50	40.8%	39.7%	39.5%	902.3	938.2	941.5
Full	Roberta	Video-Swin-T	Resnet-152	42.6%	41.3%	41.2%	853.7	885.4	886.3

We also designed a mask annotation system for human semantic mask annotation. Fig. 3 shows a screenshot of our mask annotation system. This system displays images along with the masks and 2D skeleton for each individual. Validators are required to check whether each person’s mask is complete and if the mask and skeleton correspond to the same person.

If an error in the mask is found, validators can select the person (in the top right corner of the screen) and click on several points on the person (in the bottom left corner of the screen) in the image using the mouse. These points are fed into SAM for segmentation. Through multiple-point selections, a corrected mask is eventually produced. If SAM is unable to accurately segment the person’s mask, validators have the authority to discard that person’s mask.

3.3 DATA LICENSE AND ACCESS DETAILS

The dataset is released under the CC BY-NC-SA 4.0 license. The authors of the dataset bear all responsibility in case of rights violations. The accompanying code is released under the CC0 license.

Our proposed Panoptic-L3D is based on the Panoptic dataset Joo et al. (2015), and we have obtained consent from its authors. We appreciate the support of the Panoptic dataset authors for our work.

4 MORE DETAILS OF EXPERIMENTS AND VISUALIZATION

4.1 EVALUATION ON DIFFERENT BACKBONE NETWORKS

As shown in Table 1, we conduct a series of experiments to explore the influence of different backbone networks on CPP’s performance. We use RoBERTa Liu et al. (2019), Video-Swin-T Liu et al. (2022), and ResNet-152 He et al. (2016) as the text, video, and image backbones, respectively. Three different variants are designed, each changing one of these modules.

BERT Devlin et al. (2018) is a pre-trained transformer-based model that captures context from both directions (left-to-right and right-to-left) in all layers. RoBERTa is an enhanced version of BERT, improving upon it by training with more data, longer sequences, and dynamic masking. Using BERT as our text backbone reduces the PICI metric by 0.9 and leads to an average increase of 27.7 in MPJPE-L (id 1 vs. Full).

Video-Swin-T extends the Swin Transformer, which is based on hierarchical feature maps and shifted windows, to handle temporal information in videos. This allows the model to efficiently capture both spatial and temporal dependencies, making it suitable for tasks like video classification and action recognition. I3D (Inflated 3D ConvNet) Carreira & Zisserman (2017) is a deep learning model for video analysis that extends 2D convolutional neural networks to 3D. Using I3D as our video backbone reduces the PICI metric by 1.4 and leads to an average increase of 41.9 in MPJPE-L (id 2 vs. Full).

ResNet-50 He et al. (2016) has 50 layers, while ResNet-152 He et al. (2016) has 152 layers. The increased depth in ResNet-152 allows for learning more complex features. Using ResNet-50 as our image backbone reduces the PICI metric by 1.6 and leads to an average increase of 52.2 in MPJPE-L (id 3 vs. Full).

Table 2: Quantitative evaluation on different hyperparameters.

id	Hyperparameters				PICI			MPJPE-L		
	λ_{joint}	$\lambda_{box} \& \lambda_{depth}$	λ_{mask}	$\lambda_{cord} \& \lambda_{root}$	App.	Beh.	Mov.	App.	Beh.	Mov.
1	1000	-	-	-	41.0%	40.7%	40.6%	903.4	929.2	930.1
2	100	-	-	-	36.3%	36.0%	36.1%	1178.5	1225.0	1213.5
3	-	0.01	-	-	42.5%	41.2%	41.0%	861.7	891.7	892.3
4	-	1.0	-	-	42.6%	41.3%	41.1%	854.1	885.3	889.2
5	-	-	0.1	-	42.1%	40.7%	40.9%	871.0	901.6	903.3
6	-	-	10.0	-	42.1%	40.7%	40.9%	870.8	900.9	897.4
7	-	-	-	0.01	42.5%	41.3%	41.1%	855.1	885.5	887.8
8	-	-	-	1.0	42.6%	41.3%	41.1%	853.8	885.9	887.0
Ori.	10000	0.1	1.0	0.1	42.6%	41.3%	41.2%	853.7	885.4	886.3

Table 3: Quantitative evaluation of MPJPE-success-only

Methods	Pub.	MPJPE-success-only		
		App.	Beh.	Body Move.
URVOS+VP	ECCV	170.4	171.2	170.5
MTTR+VP	CVPR	170.8	170.3	170.6
Referformer+VP	CVPR	170.6	171.2	170.2
SOC+VP	NeurIPS	170.3	170.9	170.1
CPP (Ours)	-	158.6	159.4	158.3

4.2 SENSITIVITY ANALYSIS ON HYPERPARAMETERS

As shown in Table 2, we conduct a series of experiments to explore the influence of different hyperparameters on CPP’s performance. We test six different hyperparameters, each with two different values. Among these parameters, λ_{joint} has the most significant impact on the model.

When λ_{joint} is relatively small (100 or 1000), the model has difficulty converging, leading to poor skeletal performance during training, with MPJPE-L increasing by 45.7 and 330.5, respectively. Our 2D joint map displays the positions where different joints appear. These joints are sparsely distributed, and the background without the skeleton occupies a large portion of the image. This significant distribution difference between the joints and the background might explain why the 2D joint map struggles to converge.

λ_{mask} also influences the model’s performance. When λ_{mask} is set to 0.1 or 10, the PICI metric decreases by 0.4, and the average MPJPE-L increases by 16.8 and 14.5, respectively. In contrast, the λ_{box} , λ_{depth} , λ_{cord} , and λ_{root} parameters have relatively little effect on the model’s performance.

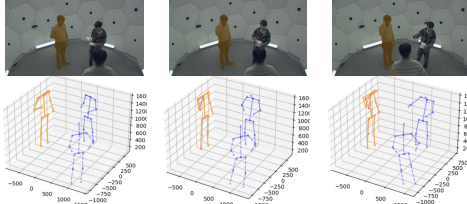
4.3 DISCUSSION OF DIFFERENT METRICS

There are two main goals in our L3DHPE setting: the first is to precisely identify the person referred to by the given text description, and the second is to minimize the deviation between the predicted pose and the ground truth pose. In our main manuscript, we adopt the PICI metric for the first goal and the MPJPE-L for both person identification and pose prediction. However, there is an intuitive solution that the metrics should evaluate these two goals separately. PICI evaluates the first goal and there should be an another metric for evaluating the second goals only. In this context, we could solely estimate the precision of pose estimation after performing successfully language-guided person identification. We perform this evaluation via an extra metric named MPJPE-success-only, as reported in Tab. 3. Although the comparison results consistently demonstrate the superiority of our proposed CPP, however, the MPJPE-success-only ignores the impact of language, which violates the key concentrations of our proposed L3DHPE task. In alignment with our goals of understanding natural language and interacting with human pose information, we believe that the MPJPE-L is more suitable for evaluating the L3DHPE task than the MPJPE-success-only.

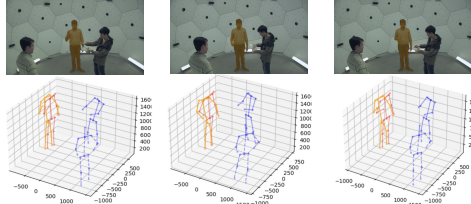
4.4 VISUALIZATIONS OF MOTION GROUNDING.

We provide detailed visualizations of successful cases of CPP in Fig. 4. Samples (a) and (b) illustrate cases of appearance description. CPP effectively generates clear boundary masks and accurate 3D

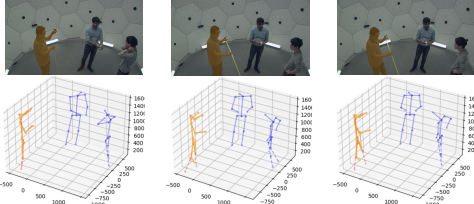
(a). A person on the left wearing blue shirt and jeans



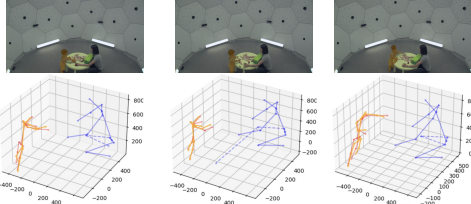
(b). The person in blue shirt and jeans stands with both hands resting on right ear at the beginning of the video



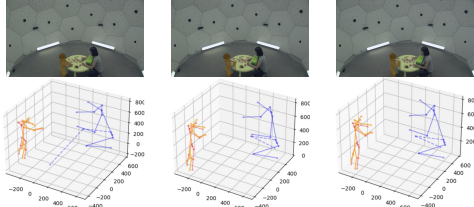
(c). The person pinches the ruler with hand, extending it forward



(d). The person walks up to the table, then picks up two blocks



(e). The person sways arms and rotates wrists



(f). The person leaned backward and swung right arm to the right

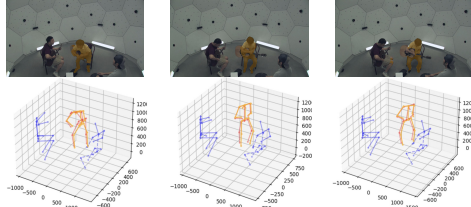


Figure 4: Successful samples of our CPP.

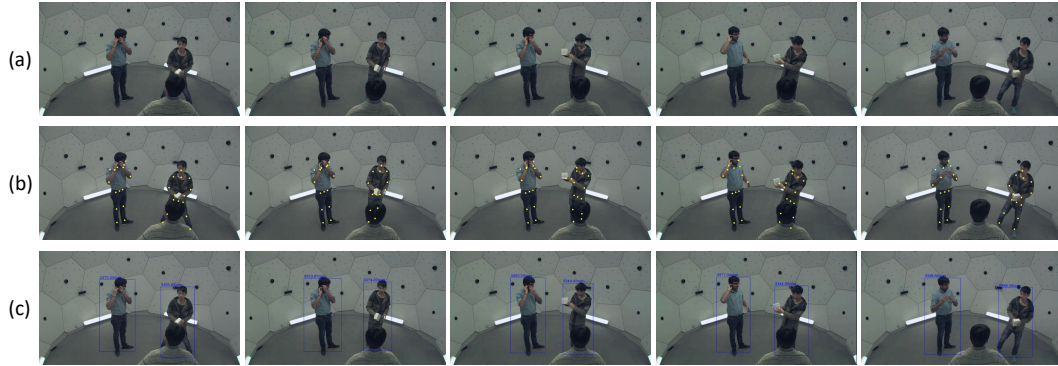


Figure 5: Visualizations of (a) the original video frames, (b) the predicted 2D joint maps, and (c) the output 2D box maps of CPP, respectively.

poses based on the given expressions, demonstrating its ability to understand terms like "blue shirt" and "jeans". Samples (c) and (d) represent behavior description cases. CPP accurately detects movement directions indicated by words such as "forward" and "up" and recognizes common objects like "ruler" and "blocks". Samples (e) and (f) showcase of body movement description. They strongly demonstrate CPP's capability to understand verbs like "sway" and "swing" and recognize joint-related nouns like "wrists" and "arm".

Additionally, we provide visualizations of 2D joint maps and 2D box maps in Fig. 5. CPP's outputs (b) 2D joint maps and (c) 2D box maps show that CPP accurately detects joints and human boxes.

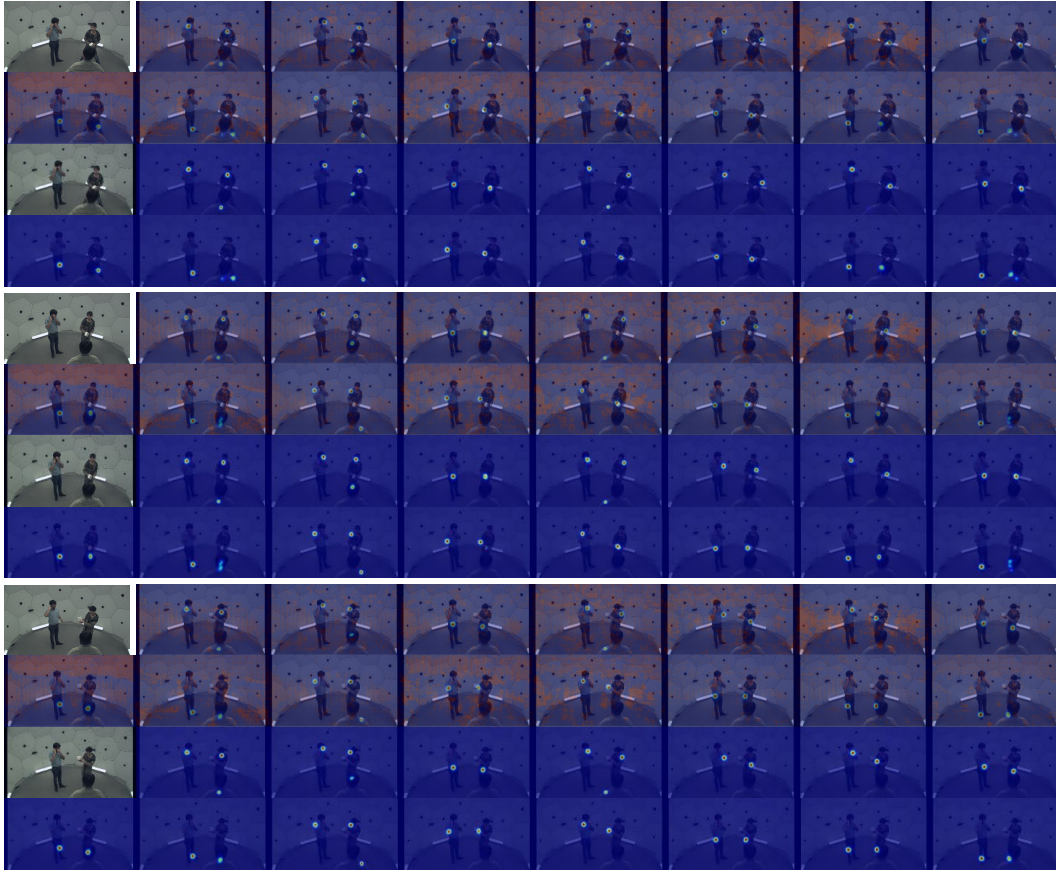


Figure 6: Visualizations of three samples for CPP’s generated 2D joint maps along with their GT maps. The former two rows of each sample are the generated joints sequence, while the latter two rows are their ground-truth joints’ positions.

This reliable 2D information helps the model recognize different individuals and segment the one referred to by the text.

Moreover, we provide visualizations of 2D joint maps for each type of joint along with their ground-truth (GT) 2D joint maps in Fig. 6. Comparing the CPP outputs with the GT 2D joint maps, we observe that CPP accurately recognizes different joints for all individuals. These 2D joint maps enhance the accuracy of detecting the 3D positions of various joints.

5 DATASHEET OF THE PANOPTIC-L3D DATASET

5.1 MOTIVATION

For what purpose was the dataset created? In reality, human motion is characterized by diverse and detailed descriptions, highlighting an emerging need for 3D human pose estimation to address more complex, real-world, and multi-person scenarios. This necessitates a shift towards incorporating motion grounding, a concept that connects dynamic human movements with rich semantic context, enabling a more comprehensive understanding of human activities in natural and multifaceted environments. Given the scarcity of pose estimation datasets with precise and richly contextual linguistic annotations, we developed Panoptic-L3D, the first L3DHPE dataset, to propel advancements in this field.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? The creators of the Panoptic-L3D datasets comprise nine researchers from a single research group. We represent only a laboratory within a university.

Table 4: Labels of the Panoptic-L3D dataset

Key	Value type	Example
video_id	str	160224_haggling1
camera_id	int	30
begin_frame	int	3249
end_frame	int	3392
body_id	int	4
category	list	[4]
caption	str	The person is holding a tape measuring with both hands

Any other comments? Through the release of Panoptic-L3D, we extend an invitation to the research community to build upon our foundation, exploring and improving language-guided 3D pose estimation.

5.2 COMPOSITION

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? The Panoptic-L3D dataset consists of several videos along with their corresponding annotations, including language descriptions, mask images, and frame-level 3D skeleton annotations for all individuals in the videos.

How many instances are there in total (of each type, if appropriate)? There are 588 videos featuring 1,476 targeted individuals included in the Panoptic-L3D dataset. The annotations are of 3,838 language descriptions, 6035 mask images, and corresponding frame-level 3D skeleton annotations for all individuals.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? Panoptic-L3D builds upon the Panoptic dataset Joo et al. (2015), and we have obtained consent from its authors, for which we are grateful for their support. From the Panoptic dataset, we have selected fourteen activity entries that involve more than one person, spanning five social interaction categories: Haggling, Ultimatum, Toddler, Musical Instruments, and Special Events. For each activity entry, we have obtained two variants from different camera viewpoints (cameras No.16 and No.30) to ensure the robustness of 3D human pose estimation. Specifically, the selected entries are: 160422_haggling1, 160226_haggling1, 160224_haggling1, 170404_haggling_a2, 170407_haggling_b2, 170221_haggling_b2, 160422_ultimatum1, 160906_band1, 160906_band2, 160906_band3, 160906_pizza1, 160906_ian1, 160906_ian2, and 160906_ian3.

What data does each instance consist of? Each video consists of several individuals that are in motion. Each annotated individual has its own language descriptions, mask images, and frame-level 3D skeleton annotations.

Is there a label or target associated with each instance? Yes, each instance is associated with its annotations. The details of annotations are shown in the Table 4. ‘Video_id’ represents the entry of the video clips. ‘Camera_id’ indicates the camera number used in the Panoptic dataset. We use only cameras No.16 and No.30, so ‘Camera_id’ has only two possible values: 16 and 30. Since redistribution of the Panoptic dataset is not permitted, we require the Panoptic-L3D dataset users to download the original videos from the Panoptic official website. ‘Begin_frame’ and ‘end_frame’ represent the starting and ending frame numbers of the original videos. ‘Caption’ provides a language description of individuals. ‘Body_id’ represents the ID of the person referred to by the caption. ‘Category’ represents the type of caption, ranging from 0 to 5. Categories 0-3 represent appearance descriptions such as clothes, posture, gender, and initial pose, respectively. Categories 4 and 5 represent behavior and body motion descriptions. Since two types of captions can be described in one appearance description sentence, we use a type list to store two categories that appear simultaneously.

Is any information missing from individual instances? Yes, only individuals visible from the waist up should be considered valid subjects for description. All others will be discarded.

Are there recommended data splits (e.g., training, development/validation, testing)? Yes, we provide data splits for the Panoptic-L3D dataset. The dataset is split according to the textual descriptions, where the training/validation/testing set comprises 3,005/312/521 sentences, respectively, along with their corresponding videos, masks, and 3D skeletons.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? Yes, our proposed Panoptic-L3D relies on the Panoptic Joo et al. (2015) dataset, and we have obtained consent from its authors. We collect the annotations our-self.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? Since the dataset targeting our proposed *Language-Driven 3D Human Pose Estimation* (L3DHPE) task, aims to reconstruct 3D pose sequences for individuals based on detailed language descriptions that capture aspects such as appearance, behavior, and body movements, it is possible to identify individuals directly.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? While the PanopticL3D dataset provides valuable advancements in video analysis, it is important to consider the potential risks associated with its misuse. There is a possibility that the dataset could be used in ways that might infringe on privacy or be employed in systems that do not fully respect individual rights. For example, if not properly regulated, such systems could be misapplied in scenarios that raise concerns about personal privacy or lead to unintended consequences. To address these concerns, we believe it is essential to ensure that developments based on PanopticL3D adhere to ethical standards and are guided by appropriate regulatory frameworks.

Any other comments? We appreciate the support of the Panoptic dataset authors for our work. We require all users of Panoptic-L3D to cite the papers of Panoptic Joo et al. (2015) dataset.

5.3 COLLECTION PROCESS

How was the data associated with each instance acquired? To construct the language-driven 3D human pose estimation dataset, we utilize videos from the Panoptic Joo et al. (2015) dataset. To simulate complex multi-person interactions found in real life, we select fourteen activity entries that feature more than one person and span five social interaction categories: Haggling, Ultimatum, Toddler, Musical Instruments, and Special Events. For each activity entry, we obtain two variants from different camera viewpoints (cameras No.16 and No.30) to ensure the robustness of 3D human pose estimation. Subsequently, these videos are separated into clips and filtered. Then, we employ over 10 annotators to label individuals in the videos according to our developed linguistic expression annotation system and guidelines. Specifically, each annotator is required to provide three sentences describing the attributes (i.e., appearance, behavior, and body movements) of each person visible from the waist up.

To ensure concise language descriptions, we establish the following guidelines for annotators:

- **Shared Rules:** 1) Only individuals visible from the waist up should be considered valid subjects for description. 2) Each description must uniquely refer to only one person in the given video. 3) Unobtrusive attributes should be omitted to avoid ambiguity if they are difficult to describe properly.
- **Rules for Describing Appearance:** Annotators describe the appearance based on the first frame of the video, focusing on four specified elements: clothes, posture, gender, and initial pose. To simulate variability in individual perception, two of these four elements are uniformly sampled for each individual, and annotators are required to base their descriptions on the selected attribute pair. This procedure enhances the diversity and informativeness of appearance descriptions, aiding in video 3D pose analysis.
- **Rules for Describing Behaviors & Body Movements:** Linguistic expressions can describe behaviors across multiple frames, encompassing both fleeting and prolonged actions. Annotators must describe an individual’s behavior and body movements after watching the entire video, ensuring comprehensive coverage. Descriptions of behaviors must exclude

appearance information, while descriptions of body movements should focus solely on the motion itself, excluding interactions with environmental elements (e.g., cups, food, pens).

Finally, the mask annotations for each referred individual in our video data are generated using the advanced 2D segmentation model SAM Kirillov et al. (2023). We utilize the projected joint points of the upper body (i.e., neck and head top) from the 3D skeleton provided by the original Panoptic dataset as coarse point prompts for SAM. For each video, we generate a coarse mask every 15 consecutive frames.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?

We design a sentence annotation system and a mask annotation system for human description annotation and human semantic mask annotation, respectively.

Fig. 2 shows the various stages of the sentence annotation system. Panel (a) shows the login page. We have three types of pages corresponding to our three annotation processes: video clip creation, sentence annotation, and sentence validation, as shown in panels (b), (c), and (d) respectively. Annotators begin by cutting the videos into clips, as shown in (b). They can select the start and end frames of the clips and replay them to ensure accuracy. Next, annotators annotate the description of every individual, as shown in (c). Finally, the validation page (d) presents the original video and its corresponding descriptions to the first validator, who must identify the target person referred to by the statements. If the chosen target matches the annotator’s target, the annotated video is retained. If there is a discrepancy, the system sends the descriptions to a second validator for further review.

Fig. 3 shows a screenshot of our mask annotation system. This system displays images along with the masks and 2D skeleton for each individual. Validators are required to check whether each person’s mask is complete and if the mask and skeleton correspond to the same person. If an error in the mask is found, validators can select the person (in the top right corner of the screen) and click on several points on the person (in the bottom left corner of the screen) in the image using the mouse. These points are fed into SAM for segmentation. Through multiple-point selections, a corrected mask is eventually produced. If SAM is unable to segment the person’s mask accurately, validators have the authority to discard that person’s mask.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? The videos in our Panoptic-L3D dataset are a subset of the Panoptic dataset. We have our strategy for sampling high-quality video clips. To simulate complex multi-person interactions found in real life, we select fourteen activity entries from the Panoptic dataset that feature more than one person and span five social interaction categories: Haggling, Ultimatum, Toddler, Musical Instruments, and Special Events. For each activity entry, we obtain two variants from different camera viewpoints (cameras No.16 and No.30) to ensure the robustness of 3D human pose estimation.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors)?

We employ 16 students for our annotation. Specifically, 2 students select the entries of the Panoptic dataset and cut them into video clips. Then 4 students annotate descriptions of individuals. Then another 4 students validate the description to make sure our Panoptic-L3D is high-quality. If these 4 students find some descriptions have an ambiguity, another 2 students will make a final check and decide whether to modify or discard the ambiguous data.

Over what timeframe was the data collected? The annotation of our Panoptic-L3D dataset lasted for three months from March to May.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? We collect the data from the individuals from the Panoptic dataset.

Did the individuals in question consent to the collection and use of their data? Yes, as shown in figure 7, we have the consent of the authors of the Panoptic Dataset.

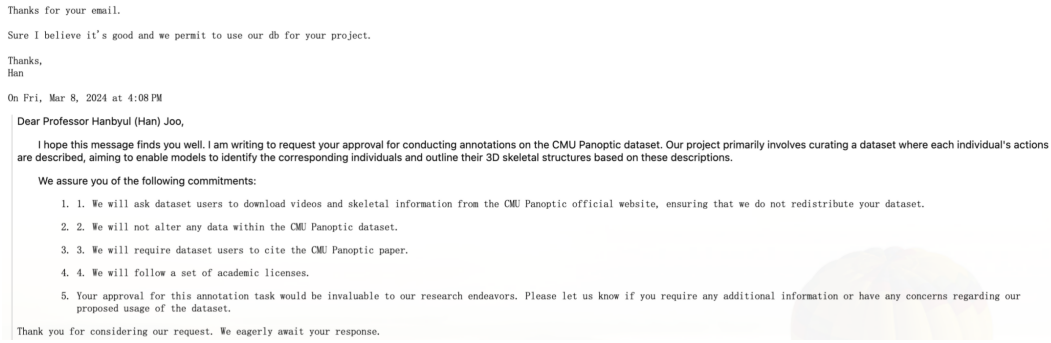


Figure 7: Screenshots of our consent from authors of Panoptic Dataset.

5.4 PREPROCESSING/CLEANING/LABELING

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

Yes, we filter and cut the video into video clips. Some low-quality data are discarded. Here are the guidelines for our filter process:

- **R1:** The length of the video clips is standardized to ensure an average length of 5 seconds. This duration allows individuals to perform a series of actions without being overly complex, making it suitable for describing characters' behavior using sentences.
- **R2:** To ensure the accuracy of language descriptions referring to the target individuals, each video clip must include at least two participants with fully exposed upper bodies, and at least one individual must be undergoing significant movements. This requirement focuses on perceptible limb movements, providing necessary motion information that matches the descriptions in 3D pose recognition and estimation.

For description annotation, we also make validation to filter the error data. Specifically, we follow an interactive game-like approach inspired by ReferIt Kazemzadeh et al. (2014), involving two participants in an alternating validation process. Initially, the original video and its corresponding descriptions are presented to the first validator, who must identify the target person referred to by the statements. If the chosen target matches the annotator's target, the annotated video is retained. If there is a discrepancy, the sample is forwarded to a second validator for further review. If the second validator finds any ambiguity, the description is revised. If ambiguity persists, the sample is discarded to maintain the accuracy of descriptions.

For mask annotation, we also make validation to filter the error data. We use SAM model to segment individuals introduced in section 5.3. However, some misprojected 3D joints may occur during mask generation due to overlapping between individuals. Additionally, SAM may focus only on the upper part of the referred person since the joints are mostly located in the upper body. To validate and correct the mask annotations, we perform a multi-round check and re-annotation process, ensuring completeness and consistency between the masked individual and the corresponding description.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? Some of the "raw" data are saved. Specifically, videos and corresponding frame-level 3D skeleton information are saved, since we utilize the Panoptic dataset which can be accessed by every user. Users can find these "raw" data on the official website of Panoptic dataset. However, the "raw" data of description, and masks are not saved.

Is the software that was used to preprocess/clean/label the data available? Yes, we design 2 annotation (validation) systems for our annotation process. Please refer to section 5.3 for more details.

5.5 USES

Has the dataset been used for any tasks already? The Panoptic-L3D dataset has already been used for our proposed language-driven 3D pose estimation task, which is introduced in the main paper.

Is there a repository that links to any or all papers or systems that use the dataset? Yes, please see in <https://languagedriven3dposeestimation.github.io/>.

Any other comments? The users should cite our paper as well as the Panoptic dataset paper Joo et al. (2015).

5.6 DISTRIBUTION

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? The dataset will be distributed through the website (<https://languagedriven3dposeestimation.github.io/>).

When will the dataset be distributed? We have already distributed the Panoptic-L3D dataset.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? The Panoptic-L3D dataset is released under the CC BY-NC-SA 4.0 license. The authors of the dataset bear all responsibility in case of rights violations.

5.7 MAINTENANCE

Who will be supporting/hosting/maintaining the dataset? The first author of the Panoptic-L3D dataset will be supporting/hosting/maintaining the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? The owner/curator/manager of the dataset can be contacted through Email addresses. We will release the email addresses of all the authors.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? We plan to update our dataset every month. We will collect the questions from users and dynamically modify our dataset. The updated information will be posted on <https://languagedriven3dposeestimation.github.io/>.

Will older versions of the dataset continue to be supported/hosted/maintained? Yes. The older versions will be stored in our private database. If any dataset users want the older version of our Panoptic-L3D dataset, please contact us.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? Yes, if others want to extend/augment/build on/contribute to the dataset, please contact any author of our papers by email.

REFERENCES

- Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.
- Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *CVPR*, pp. 20186–20196, 2022.
- Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. Posescript: Linking 3d human poses and natural language. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

- Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J Black. Chatpose: Chatting about 3d human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2093–2103, 2024.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.
- Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, pp. 3334–3342, 2015.
- Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pp. 787–798, 2014.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pp. 4015–4026, 2023.
- Jing Lin, Yao Feng, Weiyang Liu, and Michael J Black. Chathuman: Language-driven 3d human understanding with retrieval-augmented tool reasoning. *arXiv preprint arXiv:2405.04533*, 2024.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, pp. 3202–3211, 2022.
- Zhuoyan Luo, Yicheng Xiao, Yong Liu, Shuyan Li, Yitong Wang, Yansong Tang, Xiu Li, and Yujiu Yang. Soc: Semantic-assisted object cluster for referring video object segmentation. *NeurIPS*, 36, 2024.
- Vittorio Mazzia, Simone Angarano, Francesco Salvetti, Federico Angelini, and Marcello Chiaberge. Action transformer: A self-attention model for short-time pose-based human action recognition. *Pattern Recognition*, 124:108487, 2022.
- Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571. Ieee, 2016.
- Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *CVPR*, pp. 7025–7034, 2017.
- Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, pp. 7753–7762, 2019.
- Jiajun Su, Chunyu Wang, Xiaoxuan Ma, Wenjun Zeng, and Yizhou Wang. Virtualpose: Learning generalizable 3d human pose models from virtual data. In *ECCV*, pp. 55–71. Springer, 2022.
- Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *ICCV*, pp. 2602–2611, 2017.
- 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*, pp. 13243–13252, 2022.
- Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *CVPR*, pp. 1112–1121, 2020.
- Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *ICCV*, pp. 11656–11665, 2021.