

Open-Vocabulary Semantic Part Segmentation of 3D Human

Supplementary Material

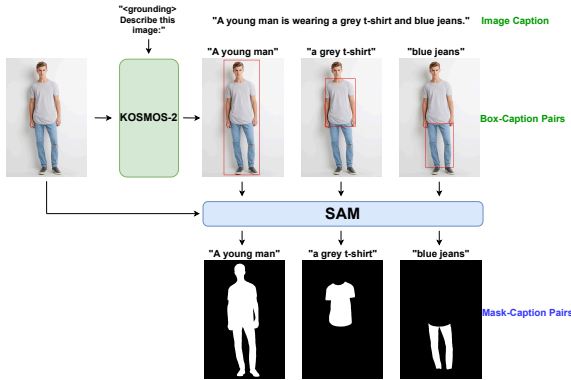


Figure 12. Data generation pipeline to create additional training data for HumanCLIP.

A. Data Generation Pipeline for HumanCLIP

The pipeline used to automatically generate training data for HumanCLIP is shown in Figure 12. It applies KOSMOS-2 and SAM to create diverse mask-caption pairs. First, the pipeline takes the input image and a text prompt, “<grounding>Describe this image:”, and fed to KOSMOS-2. The text prompt is given a <grounding>tag to guide KOSMOS-2 to locate image regions associated with texts in the output caption. This results in an image caption with box-caption pairs. In the second step, we convert the bounding boxes to binary masks by feeding the input image and each bounding box location to SAM. As a result, we are able to obtain diverse mask-caption pairs without any human intervention. **The final 1.3 million pairs of data would be released as the HumanCLIP dataset to facilitate future studies.**

The pipeline shares a similar flow with PartSLIP++ to convert bounding boxes to binary masks with SAM. However, a significant advantage of our method is that **we do not assume text prompts for each mask are provided.** Instead, we adopt KOSMOS-2 to simultaneously generate box-caption pairs with a general text prompt. It ensures a more diverse and comprehensive coverage of contents within each image, enhancing the generalizability of our HumanCLIP model.

B. Few-shot Learning of PartSLIPs

PartSLIP and PartSLIP++ are both claimed as low-shot methods, with their full potential unlocked through few-shot fine-tuning by category. We assess the performance of models that have undergone few-shot fine-tuning compared to the pre-trained checkpoints provided by the au-

Table 6. Few-shot learning efficacy of PartSLIP and PartSLIP++ on 3D human data.

Model	SIZER		CTD	
	Acc.	mIoU	Acc.	mIoU
PartSLIP	84.94	70.79	75.11	55.46
PartSLIP (few-shot)	84.90	70.70	75.65	56.07
PartSLIP++	86.63	72.93	80.73	62.36
PartSLIP++ (few-shot)	86.67	72.99	80.69	62.49

thors. We follow their low-shot setting and prepare point clouds to cover all of the part categories in the dataset with 8 samples as claimed in the paper. These are then rendered from 10 views to obtain images and ground truth bounding boxes. We keep the parameters of the GLIP model frozen and train only the learnable offset for each part. The few-shot results on the SIZER and CTD datasets are shown in Table 6. From the table, it can be observed that there is not a significant improvement in performance for both models on each dataset. Hence, we do not distinguish few-shot settings in our evaluations in the main paper.

C. More Qualitative Comparison

Additional visual comparisons for each dataset with other open-set 3D segmentation methods are shown in Figure 15.

D. Segmentation of Generated 3D Human Models

Due to the rapid development of 3D asset generation techniques, the models to be segmented may not originate from real-world scans. We demonstrate that our method effectively bridges the domain gap for less photorealistic generated data, significantly broadening its applicability across various content types. Figure 13 shows two examples of segmentation results on 3D humans generated from a text-to-3D human generation model. We use the output results from HumanNorm where the 3D humans were generated



Figure 13. Segmentation of 3D humans generated from HumanNorm: an off-the-shelf text-to-3D model.

from the text descriptions: “a DSLR photo of Messi” (left) and “a DSLR photo of Stephen Curry” (right). In both cases, our framework can segment the generated models into distinct parts corresponding to the input prompts. It validates the robustness of our method to segment 3D humans at varying quality and content.

E. Promptable Segmentation

Additional examples of our framework’s promptable segmentation capability is visualized in Figure 16. In the figure, each row represents a different combination of input prompts to highlight our method’s versatility in segmenting any category the user wants.

F. In-the-wild Segmentation

Visual comparisons with PartSLIP and PartSLIP++ on our in-the-wild point dataset is shown in Figure 14.

F.1. In-the-Wild Mesh Dataset Description

We would release the 3D human mesh dataset we use under in-wild segmentation settings, consisting of 15 subjects in different poses to demonstrate different practical scenarios. We build a multi-view capturing system using consumer-level RGB-D cameras. To mimic the in-the-wild quality of 3D models, we utilize only four incomplete views and capture under the daylight environment, resulting in a relatively noisy and non-watertight surface. The 3D models are recon-

structed through unprojecting and fusing. The meshes are created from fused point clouds through Poisson Surface Reconstruction. The generated meshes are then smoothed using a simple Laplacian filter to reduce high-frequency noise. **This dataset will be made open-source**, and contains the following poses and configurations:

- **Sitting Pose:** Subjects were scanned in a natural sitting posture.
- **Arms Stretched Out:** Subjects were instructed to extend their arms fully, providing a clear view of the torso and limbs.
- **Human Object Interaction:** Subjects held various objects such as a bottle, book, or bag.
- **Loose Clothing:** Subjects wore loose clothing, such as hoodies.
- **Multi-Layer Clothing:** Subjects wore multiple visible layers of clothing.

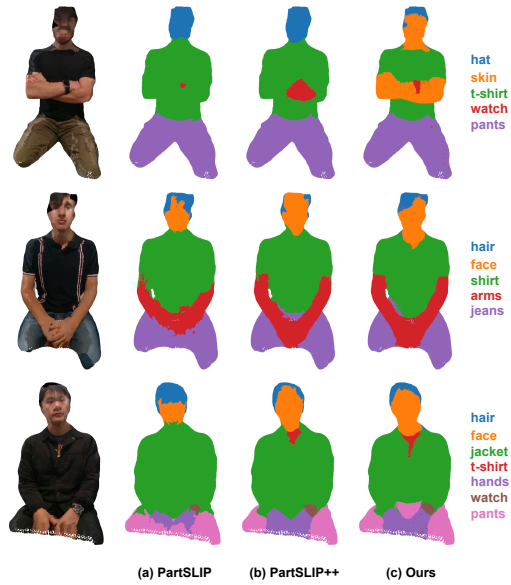


Figure 14. Additional visual comparisons of (a) PartSLIP, (b) PartSLIP++, and (c) Ours on our in-the-wild point cloud dataset.



Figure 15. More Qualitative analysis of segmentation results with PointCLIPv2, SATR, PartSLIP, and PartSLIP++ on the 3D scans



Figure 16. Additional examples of promptable segmentation. Each row shows a different type of combination of input prompts.