1. Parameter Details

Similar to DHOT and PIHOT, we use ResNet-50 as our feature extractor. During training, we freeze the SAM, depth model, and text encoder, while training the image encoder and image decoder. The learning rate for both the encoder and decoder is set to 5e-5. We use the AdamW optimizer with a weight decay of 1e-4. The batch size is set to 4 per GPU, and the model is trained for 200 epochs. For data augmentation, we randomly crop the input images to 224×224, and apply random flipping, random noise addition, and other transformations. In Eq. 21, α, β, and γ are set to 0.3, 0.1, and 1.0, respectively. The network is optimized using the AdamW optimizer. The batch size is set to 4 per GPU. The experimental environment is Ubuntu 20.04, equipped with 8 NVIDIA A6000 (48GB) GPUs. PyTorch version is 1.11.0, torchvision version is 0.12.0, and Python is 3.8.19.

2. Stronger encoder

To ensure a fairer comparison, we follow the design of DHOT and PIHOT by using ResNet-50 as the encoder. Of course, our method can also be trained with more powerful encoders. We tested the Swin-L encoder on a 4× A100 GPU server, and the results are summarized in Table 1. It is evident that using a more powerful encoder further improves the performance.

Encoder	SC-Acc.	C-Acc.	mIoU	wIoU	AD-Acc.
ResNet-50	46.0	74.9	25.6	30.2	42.3
Swin-Large	**48.0**	**78.1**	**27.8**	**31.5**	**46.3**

Table 1. Performance comparison of different encoders.

3. Inference time

The comparison of inference time between our proposed model and the two latest HOT methods is summarized in Table 2. DHOT has an inference time of 181ms, primarily due to its use of post-processing with weighted outputs during prediction. PIHOT is the slowest, with an inference time of 208ms, as it employs an object restoration model.

In contrast, our proposed method achieves an inference time of 91ms, with the main computational cost coming from the text-prompted segmentation model. Furthermore, if we replace the segmentation framework in our pipeline with a faster model, such as YOLOv8, the inference speed improves significantly to 26ms, albeit with a slight loss in accuracy.

4. Ablation studies

For each component of RJLoss, namely Local Joint Loss and Global Joint Loss, we conducted an ablation study, as shown in Table 3. The baseline setup uses only CE + BE loss, where CE represents cross-entropy loss, and BE refers to binary cross-entropy loss, which is used for computing the image-text similarity loss. It is evident that incorporating either Local Joint Loss or Global Joint Loss improves accuracy. When both are used together, i.e., when applying RJLoss, the results achieve the best performance.

model	DHOT	PIHOT	ours
time(ms)	181	208	91(26)

Table 2. Inference time of three methods on HOT-Annotated dataset.

ResNet-50	SC-Acc.	C-Acc.	mIoU	wIoU	AD-Acc.
CE+BE	44.5	72.3	23.8	28.3	40.1
CE+BE+Local	45.1	73.8	24.2	28.7	40.7
CE+BE+Gloal	44.9	74.0	23.9	28.8	41.0
CE+BE+RJLoss	**46.0**	**74.9**	**25.6**	**30.2**	**42.3**

Table 3. Performance comparison of different loss.

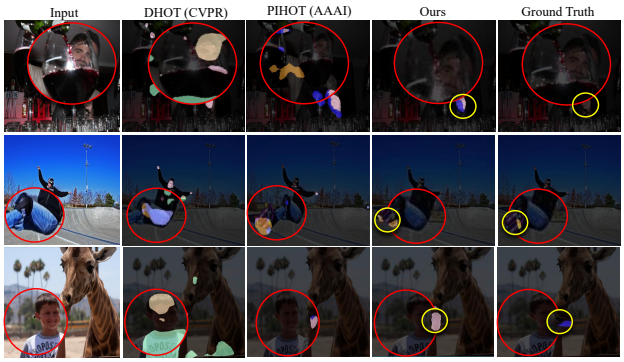

Figure 1. Qualitative results of our contact predictor and DHOT.

5. Visualization

When looking at the second and fourth columns in Figure 1, it becomes apparent that the DHOT method struggles to understand the spatial connection between individuals and objects. Certain non-contact areas are mistakenly categorized as contact areas. The DHOT model does not accurately determine the positions of the boy, the mountain, and the giraffe, resulting in inaccurate detection outcomes (third row). Our predictor can perceive the relationships between the three and accurately detect the actual contact area between the boy and the giraffe. This difference may be due to our proposed human proximal perception mechanism, which can dynamically perceive key depth range around the human. This further demonstrates that the proposed method is better at understanding the relationships between objects

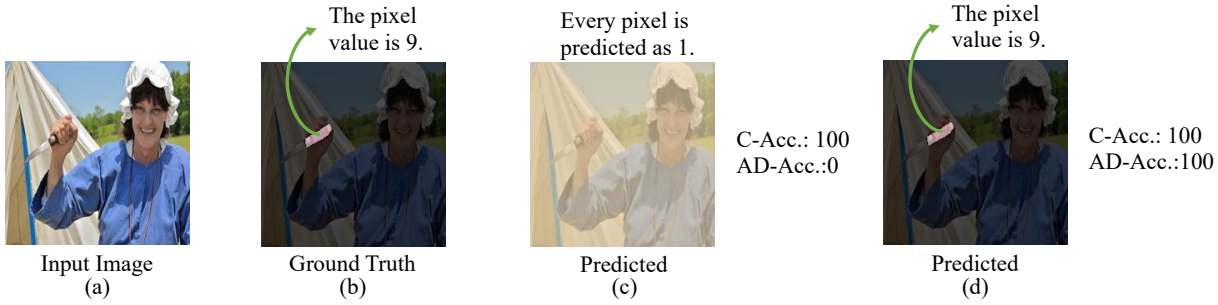

Figure 2. Illustrative diagram of the C-Acc. issue.

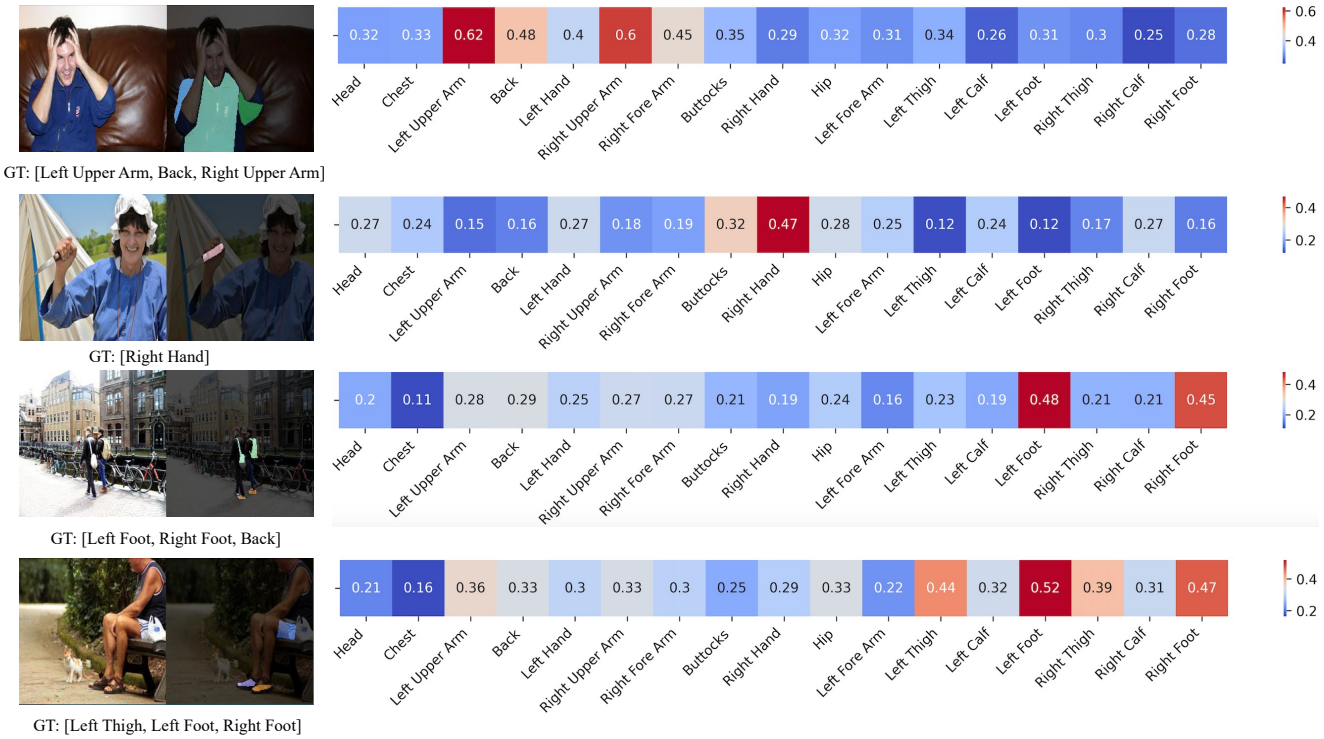

Figure 3. Visualization of text attention.

and more effectively avoids misjudgments caused by visual illusions.

6. Regarding τ

Ideally, τ should be adjustable for each image. However, in our previous attempts, we found that learning a single parameter for each image was difficult. Instead, we learn a general parameter that can adapt to the majority of images. This approach simplifies the training process while optimizing the model performance at the dataset level. Moreover, experimental results have confirmed that this method yields promising results.

7. Text Prompt

In the main text, the text prompt template used in the text encoder is "A [body part] of the human body is in contact with an object." By replacing [body part] with the corresponding body part, we obtain the specific text prompts. However, this template is not fixed. We also tested other templates, such as "A [body part] touches an object." We found that the results remained unaffected. Since both sentences effectively describe the contact between body parts and objects. We believe that any text prompt template conveying this information should work similarly. Of course, the template should not be too long, as excessive length may reduce the sensitivity to key information.

8. Text attention

In the main text, our proposed framework utilizes an image encoder and a text encoder to compute text-image similarity, which provides attention to the features in the image decoder. Additionally, the image encoder and text encoder are trained using the image-text matching loss (BE, binary cross-entropy loss) introduced in the main text. To intuitively assess the effectiveness of this module, we visualize the text attention, as shown in Figure 3.

The Figure 3 presents heatmaps illustrating the attention distribution across different body parts for five different test cases. Each row corresponds to a distinct scenario where specific body parts receive higher attention, represented by warmer colors (red). The x-axis labels indicate various body regions (e.g., "Head", "Left Upper Arm", "Right Foot"), while the numbers inside each cell denote the confidence scores assigned to each region. The ground truth (GT) labels indicate the target regions expected to receive attention. The visualization highlights how our model effectively focuses on relevant body parts in accordance with the given task, demonstrating its ability to capture meaningful spatial relationships.

9. Issues with the C-Acc. Metric.

C-Acc. denotes the accuracy of classifying pixels on the human body, which is a binary classification. When computing C-Acc., the predicted background pixels are first set to 0, while the predicted foreground pixels (i.e., human body parts) are set to 1. Then, the overlap is calculated based on the nonzero pixel regions in the ground truth. However, if the entire image is predicted as the foreground, C-Acc. will always be 100, regardless of whether the predicted contact categories are correct or whether background pixels are incorrectly classified as foreground. This is clearly incorrect.

As shown in Figure 2, subfigures (a) and (b) represent the input image and the corresponding ground truth, respectively, while (c) and (d) illustrate two different prediction results. In the ground truth, the right hand of the human body is in contact with an object, and the pixels in the contact region are labeled as 9. When all pixels are predicted as a nonzero value, as in subfigure (c), we observe that the C-Acc. value is 100, which is clearly incorrect. To address this issue, we propose a new evaluation metric, AD-Acc., to replace C-Acc.. We find that in the case of subfigure (c), the AD-Acc. value approaches 0, which is the correct outcome. Only when the predicted map exactly matches the ground truth, as in subfigure (d), do both metrics reach 100. Therefore, the visualization of subfigures (c) and (d) intuitively demonstrates the correctness of our proposed AD-Acc. metric.

9.1. Limitations and Future Directions

At present, the proposed method has some limitations, specifically, it is only applicable for analyzing 2D images and has not been expanded to learn 3D human-object interaction trends from 2D images. Furthermore, the method's performance may decline when both people and objects are obstructed. In our upcoming research, we will address these challenges and investigate possible enhancements.

Several potential development directions have been considered to overcome these limitations. One method involves creating a single multi-task learning model that combines human segmentation, depth estimation, and HOT prediction, improving the method's strength and precision by optimizing all tasks together. Another potential avenue is to develop a new HOT detection system designed for live 3D video feeds, with the ability to anticipate HOT connections in more complex environments. Furthermore, implementing better methods for dealing with obstructions, such as integrating more detailed contextual information or utilizing techniques for combining multiple perspectives, could improve accuracy in challenging surroundings. In general, these instructions are designed to address existing constraints and enhance the use of HOT prediction in real-life situations.