NL2Contact: Natural Language Guided 3D Hand-Object Contact Modeling with Diffusion Model

Zhongqun Zhang[1], Hengfei Wang[1], Ziwei Yu[2], Yihua Cheng[1] ⋆, Angela Yao[1], and Hyung Jin Chang[1]

[1] University of Birmingham, UK
[2] National University of Singapore, Singapore

We provide additional materials to supplement our main paper. We first provide the additional details of the *ContactDescribe* dataset in Section 1. Furthermore, we provide extra experimental results in Section 2, including more qualitative results on the HO3D dataset [3] (Section 2.1), more quantitative evaluation of the alignment between text description and generated contact (Section 2.2), more visualization of human grasp generation (Section 2.3), more evaluation of in-the-wild text input (Section 2.4), and some failure cases (Section 2.5).

1 More Details of *ContactDescribe* Dataset

More statistics. We further illustrate the data distribution of the grasping behavior over the dataset by quantifying the contact locations. As shown in Figure 1, the index finger is the most frequently utilized when humans grasp an object. Following that, the index and middle fingers are commonly engaged, as people typically manipulate objects using these three fingers, with the other fingers often playing a supportive role. At the level of finger joints, the fingertip is used most frequently as it is the most flexible. The palm's participation in contact is notably less frequent than that of the fingers, typically observed only during a "wrap".

Auto-checker. We leverage ChatGPT to automatically check the generated descriptions. Below we provide examples:

Q: *Is this gasp suitable for using scissors? 'The person uses the scissors for cutting, wrapping with the fingertip of thumb and index, along with the middle joint of the middle finger, while the other fingers form a loose sphere'*
A: *Yes, the described grasp is suitable for using scissors*
Q: *Is this gasp suitable for using scissors? 'The scissors is* ***pinched*** *by the thumb fingertip, in conjunction with the middle joints of the middle and index finger, while the remaining fingers support the object.'*
A: *No, the pinch grasp may lack the fine control needed for accurate cutting with scissors.*

After we obtain the feedback from ChatGPT, we manually revise the descriptions.

⋆ Corresponding author.

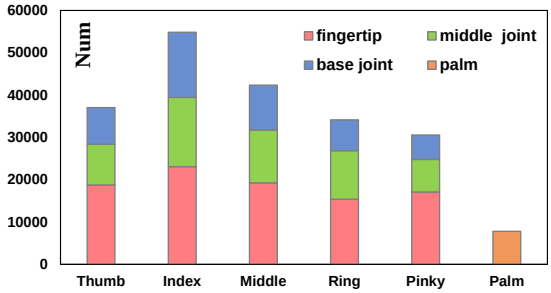

Fig. 1: Data distribution of the *ContactDescribe* dataset. The horizontal axis represents the fingers involved in the contact, the vertical axis represents quantity and different colors represent different hand regions.

2 More Results of NL2Contact

2.1 Qualitative Results on HO3D

We evaluate the generalization capability of our method on the HO3D dataset [3]. We qualitatively compare our method with ContactOpt [2]. The ground-truth of HO3D is annotated by the HOnnotate [3], which is a multi-view RGB-D hand-object tracker. The initial grasp pose is estimated by a baseline pose estimator from Hasson *et al*. [4]. As depicted in Figure 3, ContactOpt consistently produces grasps involving all five fingers, which, while realistic in terms of touch, lacks controllability and diversity in contact. In contrast, by extracting and modeling contact information from the text prompt, our method can accurately generate results that closely align with ground-truth grasps.

2.2 More Quantitative Evaluation of Contact Similarity Ratio

Our work proposes the first hand-object modeling from text descriptions, and there are no standardized measures to assess the generation. Therefore, we define a new metric, Contact Similarity Ratio (CSR), for reference. As shown in Figure **??**, we define a 16-bit touch code to represent the 16 hand parts sequentially, where 1 indicates contact and 0 indicates no contact. We extract touch codes from descriptions using keyword detection and compute touch codes from grasp GT using hand part label of MANO. We compare the two touch codes to evaluate the quality of the generated text. The CSR denotes the ratio of correct bits to the total bits. Our manually annotated prompts have 100% CSR, and the text generated from LLMs achieves 99.94% CSR. This shows the high quality of the generated text descriptions. We also use the CSR to evaluate the alignment. We respectively convert text input into touch codes and compute touch codes from the generated grasp. Table 1 shows that our method achieves 92.4% CSR, which significantly outperforms other methods.

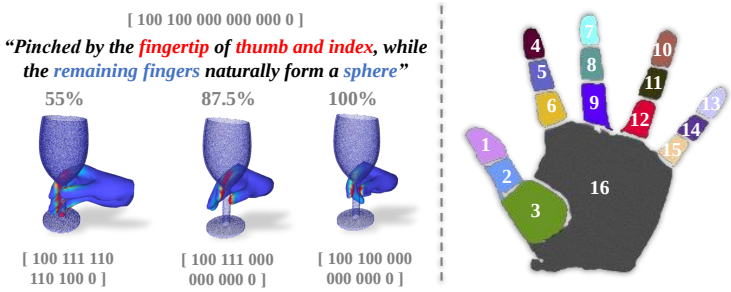

Fig. 2: Illustration of Touch Code and Contact Similarity Ratio.

2.3 More Visualization of Human Grasp Generation

We present additional visualizations using the out-of-domain unseen objects from the HO3D dataset. The generated grasps from GraspTTA [5] and ContactGen [6] appear realistic in terms of touch, they are not conducive to practical use of the object. For instance, as depicted in Figure 4, our method is capable of generating a suitable grasp for opening a bottle, whereas other methods often merely make contact with it.

2.4 More Evaluation of In-the-wild Text Description

We recruited ten non-expert participants. Each participant was shown a grasp and asked to write a text describing it. These texts serve as the in-the-wild text. One case is shown below, where the text contains everyday language such as "the second finger". We first input the description directly into our model. The generation failed to understand unknown words, such as "the second finger". We then used ChatGPT to refine the text description. We provided a text example and the description, asking ChatGPT to revise the text based on the example. With just one correction, we were able to input the refined text into our model and achieve the correct generation.We also computed the average CSR metric (please refer to R2-Q1 for the definition) for the ten participants. The CSR is 85.2% for in-the-wild texts and 93.7% for refined texts, where we only input the texts into ChatGPT once for refinement.

2.5 Failure Case

We demonstrate failure cases in our approach. Wrong descriptions can result in our method generating nonsensical hand-object contact. This is illustrated in Figure 5, where the generated contact positions of the fingers for grasping a YCB object [1] align with the textual description but are nonsensical for grasping. This situation arises when the wrong language fails to provide sufficient constraints, prompting the implementation of geometric constraints to ensure that the hand contacts the nearest surface of the object.

Table 1: Results of evaluations of the generated contact.

Method	Contact Similarity Ratio↑
GrabNet	53.1%
GraspTTA	55.9%
ContactGen	79.0%
NL2Contact (Ours)	92.4%

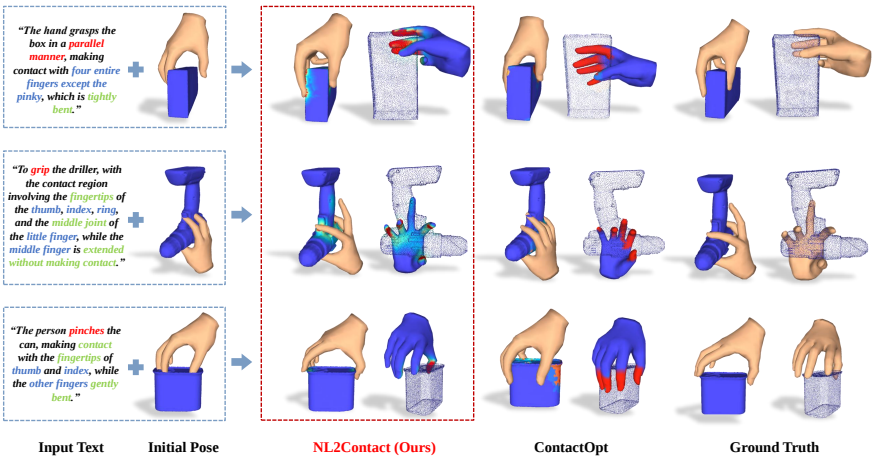

Fig. 3: Qualitative comparison of grasp pose optimization on HO3D dataset [3]. It can be seen that ContactOpt [2] always generates a grasp with all five fingers engaged, whereas our method accurately achieves the closest result to the ground truth.

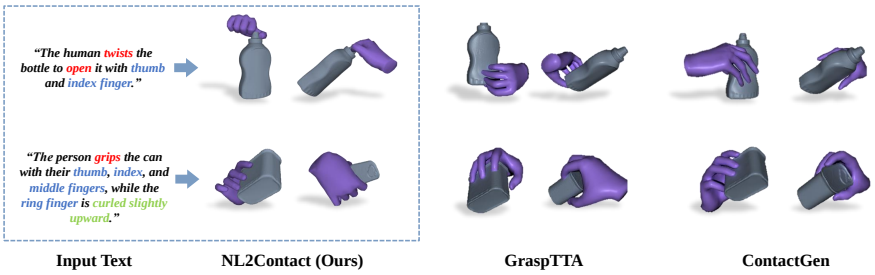

Fig. 4: Visualization of human grasp generation on HO3D dataset [3]. By leveraging text input, our method can generate a suitable grasp for using the object, whereas other methods often merely make contact with it.

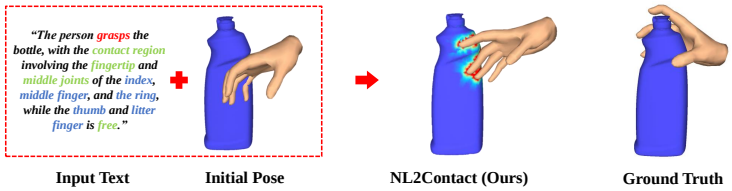

Fig. 5: Failure case of our method. Given a wrong description (e.g. the thumb contact), a generated contact is consistent with the text but is a nonsensical grasp.

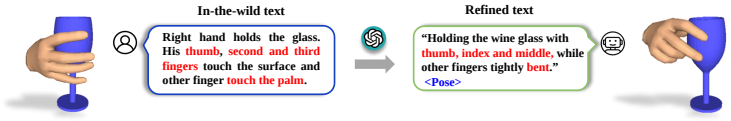

Fig. 6: Case study of in-the-wild text input.

References

1. Calli, B., Singh, A., Walsman, A., Srinivasa, S., Abbeel, P., Dollar, A.M.: The ycb object and model set: Towards common benchmarks for manipulation research. In: ICAR (2015)
2. Grady, P., Tang, C., Twigg, C.D., Vo, M., Brahmbhatt, S., Kemp, C.C.: ContactOpt: Optimizing contact to improve grasps. In: CVPR (2021)
3. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: Honnotate: A method for 3D annotation of hand and object poses. In: CVPR (2020)
4. Hasson, Y., Varol, G., Laptev, I., Schmid, C.: Towards unconstrained joint hand-object reconstruction from RGB videos. In: 3DV (2021)
5. Jiang, H., Liu, S., Wang, J., Wang, X.: Hand-object contact consistency reasoning for human grasps generation. In: ICCV (2021)
6. Liu, S., Zhou, Y., Yang, J., Gupta, S., Wang, S.: Contactgen: Generative contact modeling for grasp generation. In: CVPR (2023)