Nonverbal Interaction Detection
Supplementary Material

Jianan Wei[1*], Tianfei Zhou[2*], Yi Yang[1], and Wenguan Wang[1†]

[1]Zhejiang University [2]Beijing Institute of Technology
https://github.com/weijianan1/NVI

This document provides more details to supplement our main manuscript. We first give additional analyses about NVI in §A and present more implementation details on HOI-DET in §B. Subsequently, additional quantitative results of our NVI-DEHR are summarized in §C. Finally, we delve into an in-depth discussion about social impact, potential limitations and future directions in §D.

A Additional Dataset Analysis

More Statistics. We investigate the distribution of individuals engaged in group-wise interaction as illustrated in Fig. S1. It can be observed that the size of the gaze group exhibits considerable diversity, ranging from 2 to 12, while the touch group predominantly comprises 2 or 3 individuals. Furthermore, we present a detailed quantitative analysis of human behaviors depicted in each image (as shown in Fig. S2), including the quantitative statistics of human instances, gaze, touch, facial expression, gesture, posture.

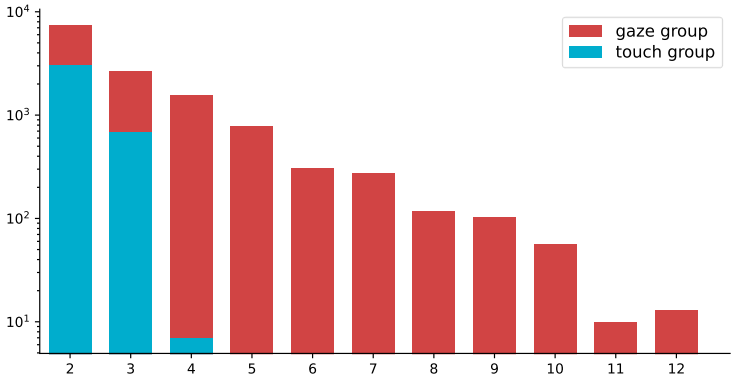

Fig. S1: Group size in group-wise interactions (§A).

More Examples. To better showcase the diversity of our NVI, we provide additional examples from various social environments involving diverse nonverbal interactions and different numbers of individuals, depicted in Fig. S3.

* The first two authors contribute equally to this work.
† Corresponding author: Wenguan Wang.

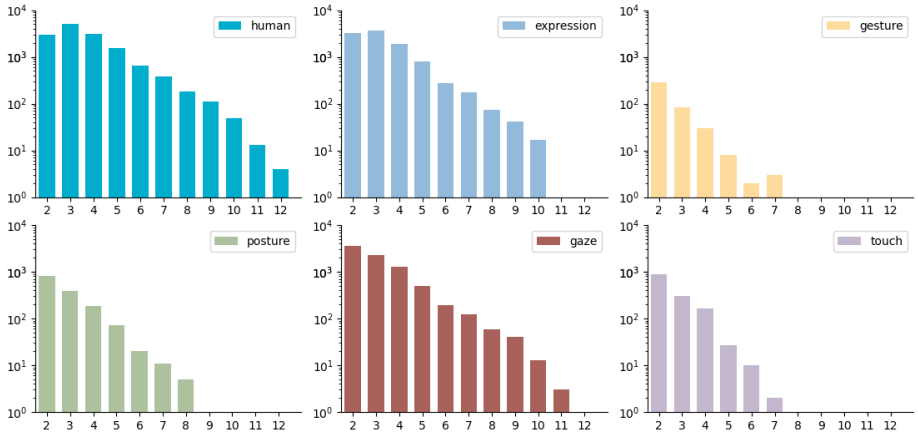

Fig. S2: The quantitative statistics of human instances, gaze, touch, facial expression, gesture, posture in each image (§A).

B More Implementation Details on HOI-DET

Training Objective. Following [4,7–9], the HOI detection loss used in this work comprises four parts: a box regression loss $\mathcal{L}_b$, a generalized IoU loss $\mathcal{L}_{\text{GIoU}}$, a cross-entropy loss $\mathcal{L}_c^o$ for object classification and a cross-entropy loss $\mathcal{L}_c^a$ for action recognition. The overall loss is the weighted sum of these parts:

$$\mathcal{L} = \lambda_b \mathcal{L}_{\text{b}} + \lambda_{\text{GIoU}} \mathcal{L}_{\text{GIoU}} + \lambda_c^o \mathcal{L}_c^o + \lambda_c^a \mathcal{L}_c^a, \tag{1}$$

where $\lambda_b = 2.5$, $\lambda_{\text{GIoU}} = 1$, $\lambda_c^o = 1$, $\lambda_c^a = 2$.

Evaluation Metrics. We adopt the mean Average Precision (mAP) for evaluation. A HOI detection is considered a true positive when the human is correctly linked to the corresponding object using the appropriate verb, and this human-object pair is accurately localized (The accuracy of localization is evaluated by measuring the overlap between the bounding boxes). For V-COCO [2], we report the mAPs for two scenarios: scenario 1 (S1) including the 4 body motions and scenario 2 (S2) excluding the HOI classes without object. Regarding HICO-DET [1], we assess performance across three settings: the complete set of 600 HOI categories (Full), a subset of 138 rare categories with fewer than 10 training images (Rare), and the remaining 462 categories (Non-rare).

C Additional Quantitative Results on NVI-DET

As seen in Table. S1, we conduct further analysis breaking down performance by interaction category. It can be observed that our NVI-DEHR demonstrates superior performance in all categories except the *posture* category, with the marginal additional costs of our model. It's worth noting that all models encounter a sharp performance decline for the *gesture* category, which could potentially be attributed to the severe long-tailed distribution within this category; for instance, `palm-out` and `beckon` are the two least frequent behaviors in NVI.

Table S1: Performance of interaction category on NVI `val` (§C).

Method	Params	FLOPs	Expression	Posture	Gesture	Touch	Gaze
m-QPIC	41.42M	56.10	77.25	66.91	40.26	80.68	74.41
m-CDN	41.41M	51.66	77.98	68.11	39.25	80.88	73.52
m-GEN-VLKT	41.71M	55.18	78.91	**78.91**	36.69	81.37	74.91
NVI-DEHR(Ours)	42.71M	59.89	**79.37**	72.94	**42.13**	**81.60**	**79.24**

D Discussion

Social Impact. NVI-DET takes a significant step towards creating socially-aware AI models with capabilities of generic nonverbal interaction understanding, and can benefit a variety of applications, like robotics, healthcare, and digital human. The proposed NVI-DEHR and NVI have no evident negative impact to society. Nevertheless, there is a risk that someone could use it for malicious purposes, *e.g.*, widespread surveillance, invasion of privacy, and potential abuse of personal information. Therefore, we strongly advocate for the well-intended application of the proposed method, while simultaneously underscoring the importance of employing the dataset in a responsible and ethical manner.
Limitation. From a feasibility perspective, we carefully select the five most representative types and 22 subcategories of them to construct NVI. But, the constrained samples may fall short of capturing the full spectrum of nonverbal interactions that take place in real-world scenarios, which could hinder the applications of NVI-DET in more complex and diverse situations. Although our image-only NVI, as a pionerring endeavor, is capable of delivering ample clues for the identification of nonverbal behaviors in most instances, there are occasional occurrences of ambiguity, like subtle facial expression and slight gaze-shift movements, akin to ambiguous actions like "throw/catch frisbee" in V-COCO.
Future Work. Moving forward, we plan to extend our NVI with temporal data for an in-depth analysis of nonverbal behaviors and enrich the variety of nonverbal interactions, like proximity *i.e.*, the physical distances involved during the interactions [3]. Inspired by previous works [5, 6, 10] in HOI-DET, which integrate simultaneous cues such as human pose or spatial relation from static images to mitigate label ambiguity, we intend to further exploit the co-occurrence of social signals in NVI-DET to recognize nonverbal interactions effectively.

References

1. Chao, Y.W., Liu, Y., Liu, X., Zeng, H., Deng, J.: Learning to detect human-object interactions. In: WACV (2018) 2
2. Gupta, S., Malik, J.: Visual semantic role labeling. arXiv preprint arXiv:1505.04474 (2015) 2
3. Hans, A., Hans, E.: Kinesics, haptics and proxemics: Aspects of non-verbal communication. IOSR Journal of Humanities and Social Science (IOSR-JHSS) **20**(2), 47–52 (2015) 3

4. Liao, Y., Zhang, A., Lu, M., Wang, Y., Li, X., Liu, S.: Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In: CVPR (2022) 1
5. Liu, X., Li, Y.L., Lu, C.: Highlighting object category immunity for the generalization of human-object interaction detection. In: AAAI (2022) 3
6. Liu, Y., Chen, Q., Zisserman, A.: Amplifying key cues for human-object-interaction detection. In: ECCV (2020) 3
7. Ning, S., Qiu, L., Liu, Y., He, X.: Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models. In: CVPR (2023) 1
8. Tamura, M., Ohashi, H., Yoshinaga, T.: Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In: CVPR (2021) 1
9. Zhang, A., Liao, Y., Liu, S., Lu, M., Wang, Y., Gao, C., Li, X.: Mining the benefits of two-stage and one-stage hoi detection. In: NeurIPS (2021) 1
10. Zhang, Y., Pan, Y., Yao, T., Huang, R., Mei, T., Chen, C.W.: Exploring structure-aware transformer over interaction proposals for human-object interaction detection. In: CVPR (2022) 3

Fig. S3: Illustrative examples of NVI (§A).