Nonverbal Interaction Detection

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

Abstract. This work addresses a new challenge of understanding human nonverbal interaction in social contexts. Nonverbal signals pervade virtually every communicative act. Our gestures, facial expressions, postures, gaze, even physical appearance all convey messages, without anything being said. Despite their critical role in social life, nonverbal signals receive very limited attention as compared to the linguistic counterparts, and existing solutions typically examine nonverbal cues in isolation. Our study marks the first systematic effort to enhance the interpretation of multifaceted nonverbal signals. **First**, we contribute a novel large-scale dataset, called NVI, which is meticulously annotated to include bounding boxes for humans and corresponding social groups, along with 22 atomic-level nonverbal behaviors under five broad interaction types. **Second**, we establish a new task NVI-DET for nonverbal interaction detection, which is formalized as identifying triplets in the form ⟨`individual`, `group`, `interaction`⟩ from images. **Third**, we propose a nonverbal interaction detection hypergraph (NVI-DEHR), a new approach that explicitly models high-order nonverbal interactions using hypergraphs. Central to the model is a dual multi-scale hypergraph that adeptly addresses individual-to-individual and group-to-group correlations across varying scales, facilitating interactional feature learning and eventually improving interaction prediction. Extensive experiments on NVI show that NVI-DEHR improves various baselines significantly in NVI-DET. It also exhibits leading performance on HOI-DET, confirming its versatility in supporting related tasks and strong generalization ability. We hope that our study will offer the community new avenues to explore nonverbal signals in more depth.

Keywords: Nonverbal Interaction · Hypergraph · Social Intelligence

1 Introduction

Nonverbal ⋯ is an elaborate code written nowhere, known by no one, and understood by all [60].

– Edward Sapir (1884 – 1939)

---

* The first two authors contribute equally to this work.

† Corresponding author: Wenguan Wang.

Fig. 1: *Can you read these humans?* Nonverbal interaction (*e.g.*, gaze, gesture) forms the cornerstone of our social life, and serves as the basis of our social intelligence.

Nonverbal behavior allows us to express intention, attention, and emotion in our social life. It includes all communicative acts except words, ranging from facial expression and gaze to bodily contact and physical appearance. By analyzing these cues, we gain valuable insights into people's thoughts, feelings, and purposes, even when they are not explicitly stated (Fig. 1). Though often conducted outside conscious awareness, nonverbal behavior is omnipresent, accounting for nearly two-thirds of all human social interactions [48,55,56]. Picture us entering a busy elevator on a Monday morning; while people inside may not seek verbal exchanges, they likely encourage our entry through nonverbal body adjustments.

This primordial form of communication, as widely studied in cognitive science and psychology [51,62], is deeply ingrained in human nature. From early infancy, human babies pay extra attention to caregivers' behavior and subsequently learn to convey their needs via nonverbal means (*e.g.*, point, gaze following) long before they speak. Endowing machines with the ability to perceive and use nonlinguistic interactions is thus an essential step towards giving computers social intelligence [3], improving human-computer interaction [1,4], and making them more accessible to people with verbal communication challenges [69].

Despite its academic value and practical significance, nonverbal interaction analysis has yet to receive sufficient attention in AI research, especially when compared with the immense progress made in linguistic analysis domains, *i.e.*, NLP and speech recognition. In computer vision, past efforts focus on interpreting nonverbal signals separately, starting from facial expressions [50,91], to gestures [52,89], gaze [16,19,33], and postures [38,82], using specialized benchmarks (*e.g.*, RAF-DB [37], VACATION [16], EgoGesture [89]). However, human social interactions are typically characterized by the initiation of concurrent behaviors. For instance, gaze aversion is always accompanied by angry/sad expressions and arm-crossing which indicates a defensive or wary mentality. Similarly, pointing gesture usually triggers gaze-following in others. The specialized datasets fail to capture this complexity nature, and existing systems developed on them are therefore incapable of interpreting them completely [5,47].

In this work, we present *the first* systematical study of nonverbal interaction in daily social situations. The goal is ambitious – make machines "freely"

communicate with humans in a non-linguistic manner. Our solution hinges on contributions in three fundamental pillars: **data**, **task** and **model**.

Data (§3). We introduce NVI, the first large-scale dataset tailored for *generic* nonverbal interaction understanding. It consists of 13,711 images accompanied with high-quality manual annotations of more than 49K humans involving in 72K social interactions. The image collection covers a wide range of social events such as coffee break in a conference or picnic in a park. Following previous efforts and established terminologies in social psychology [8, 36, 59], we consider 22 atomic-level interactions (16 individual- and 6 group-wise), that span across five broad interaction types, *i.e.*, *gaze*, *touch*, *facial expression*, *gesture*, and *posture*. NVI distinguishes itself from all prior datasets in annotation richness of interactions (see Table 1), and opens the door to study the true effect of multifaceted nonverbal signals in human social life.

Task (§4). Upon building NVI, we define a new task NVI-DET to steer the development of AI models towards generic nonverbal interaction understanding. Current efforts in the field is severely diverged – interactions are described in different ways and addressed in inconsistent granularity, *e.g.*, individual facial expressions, pair-wise human-object interactions, or group-wise gaze communication. In contrast, we propose that all nonverbal signals, regardless of their broad type and the number of participants, can be distilled into a combination of "individual behavior" – an individual act which nevertheless is influenced by others, and "collective behavior" taken together by a group of individuals. With this insight, NVI-DET is formulated to localize each individual and the social group it belongs to, meanwhile identifying the category of the interaction. This can be encapsulated into a triplet ⟨`individual`, `group`, `interaction`⟩. This formulation finds relevance to a well-established, yet different task – human-object interaction detection (HOI-DET) [25]. Unlike HOI-DET that mainly focuses on recognizing actions between human-object pairs, NVI-DET aspires to comprehensively interpret the full spectrum of communactive nonverbal signals, whose patterns are generally more subtle, ambiguous, and involving multiple persons.

Model (§5). Clearly, NVI-DET is a structured task that demands a comprehensive modeling of interaction relationships among individuals. To this end, we propose a novel solution named nonverbal interaction detection hypergraph or NVI-DEHR. It is grounded in a hypergraph structure [18, 22, 92] that is in nature flexible in high-order relation modeling. Concretely, NVI-DEHR first detects human individuals and social groups following DETR [9]. It then constructs two distinct multi-scale hypergraphs: one with human individuals as vertices and the other with social groups as vertices. This dual multi-scale hypergraph structure enables deep exploration of both individual-individual and group-group relationships at varying levels of granularity. Through hypergraph learning, NVI-DEHR obtains enriched feature representations for individuals and social groups. Finally, NVI-DEHR utilizes these updated features to categorize nonverbal interactions for all individual-group pairs. NVI-DEHR represents a seminal approach for generic nonverbal interaction understanding. We believe that it lays a solid foundation for future research in this fast-evolving domain.

Table 1: Comparison of NVI to related datasets. Unlike existing datasets that specialize in one particular interaction type, NVI offers extensive annotations across five common types, promoting a unified understanding of nonverbal social interactions.

Dataset	Year	Scale	Condition	Nonverbal Interaction Taxonomy		Task Goal
				# Atomic Class	# Broad Class	
MSR-Action3D [38]	2010	23797 images	controlled	20	1 (posture)	Action cls.
UTKinect-Action [82]	2012	6220 images	controlled	10	1 (posture)	Action cls.
GroupDiscovery [37]	2014	1,176 images	wild	7	1 (posture)	Posture cls./det.
RAF-DB [37]	2017	29,672 images	wild	18	1 (facial expression)	Facial expression cls.
Affectnet [50]	2017	450,000 images	wild	8	1 (facial expression)	Facial expression cls.
EgoGesture [89]	2018	2,081 videos	controlled	83	1 (gesture)	Gesture cls./det.
LD-ConGR [41]	2022	542 videos	controlled	10	1 (gesture)	Gesture cls./det./seg.
VideoCoAtt [15]	2018	380 videos	wild	1	1 (gaze)	Shared-gaze det.
VACATION [16]	2019	300 videos	wild	4	1 (gaze)	Gaze det.
PANDA [79]	2020	600 images	wild	10	4 (gaze, touch, facial expression, posture)	Human det./trk.
NVI (Ours)	2023	13,711 images	wild	22	5 (gaze, touch, facial expression, gesture, posture)	Generic nonverbal interaction det.

Experiments (§6). Our solution is verified through extensive experiments. First, we benchmark the NVI dataset by modifying a set of HOI-DET models for NVI-DET, and offer in-depth discussions on the newly introduced dataset and task. Second, we find that our model yields the best performance in NVI-DET, improving the adapted baselines by solid margins. Third, the model exhibits strong generalization capability, as evidenced by its remarkable performance on two standard HOI-DET benchmarks, *i.e.*, V-COCO [25] and HICO-DET [10].

2 Related Work

Nonverbal Interaction Understanding. Nonverbal communication is a highly efficient and pervasive means for interpersonal exchange. It is integral to our social intelligence [7], which has been argued to be indispensable and perhaps the most important for success in life [3]. When it comes to computers, however, they are socially ignorant. This gap has led to the emergence of social signal processing (SSP) [71] that aims at providing computers with the ability to sense and understand human social signals. Analyzing facial expressions has ever been a core focus in this area [46,80,91], which has achieved tremendous progress. Many other computational methods to automatically detect social cues from images and videos have also been proposed, including recognizing gestures [52,89], detecting eye movements [12,13,33], inferring emotions from body posture [61], or detecting social saliency [54]. Despite these advancements, they explore social signals in isolation based on highly specialized datasets, some of which are even created in controlled conditions (see Table 1). This limits them to interpret subtle meanings of social interactions which are typically transmitted via a combination of multiple signals rather than just one at a time. Though [2,32] attempt to construct a complex system that can accommodate various SSP rec-

ognizers to forecast multiple nonverbal social signals, they necessitate substantial engineering endeavors, posing constrains on the development of this field.

Human Interactions with Objects. Recently, there has been a strong push on uncovering human actions with objects, a task known as HOI-DET [10, 21, 24, 34, 57]. While these actions are also nonverbal, they differ from the types of signals we study in this article. Most actions within HOI-DET occur with *explicit intentions*, such as the action `ride bicycle`. Our work, on the other hand, focuses on the signals that display mostly *implicit intentions* but still produce social awareness. In practice, these unconscious signals are often strong indicators of the initiation of human-object interactions [63]. Moreover, HOI-DET deals with pair-wise actions, whereas NVI-DET allows for varied forms of higher-order interactions, making it a more nuanced and challenging task.

Hypergraph Learning. In recent years, graph learning has become prevalent for understanding human-centric visual relationships, notably in human parsing [76–78, 85, 93–95] and HOI detection [20, 57, 70, 73, 75]. Concurrently, hypergraphs have also garnered notable attention for its effectiveness in modeling and learning complex data correlation [23, 92]. A hypergraph generalizes a standard binary graph; it consists of vertices and hyperedges, and each hyperedge can connect an arbitrary number of vertices, rather than pair-wise connections in standard graphs. Learning on hypergraphs then turns into the process of passing information along hyperedges, facilitating message exchanges among complex relational data. Recently, deep hypergraph learning algorithms [6, 18, 22, 31, 84] have been proposed and applied to solve computer vision tasks, *e.g.*, image classification [81], pose estimation [83], and mesh reconstruction [29]. Drawing inspiration from these advances, we design a hypergraph structure to model high-order nonverbal interactions among individuals and social groups. By integrating it with a Transformer, our model shows immediate performance improvements, highlighting its efficacy in processing complex nonverbal social signals.

3   NVI Dataset

NVI is built on PIC 2.0 [42] that focuses on human-centric relation segmentation. PIC 2.0 is labeled solely with segmentation masks for entities (*i.e.*, humans, objects), and their action/spatial relations (*e.g.*, kicking ball, behind table). NVI enriches it by densely labeling social groups (see Fig. 2). To further enhance the quality of NVI, we exclude images that merely contain a single individual in PIC 2.0. Next, we first define the nonverbal interaction taxonomy used in NVI (§3), then detail the annotation process (§3), and report dataset statistics (§3).

Nonverbal Interaction Taxonomy. We categorize nonverbal interactions following a hierarchical taxonomy (see Fig. 3(a)). We first define five broad interaction types that are recognized as the most important behavioural cues in psychology research [26, 36, 59]. Among them, two (*i.e.*, *gaze*, *touch*) are group-wise, typically involving multiple individuals, while the other three (*i.e.*, *facial expression*, *gesture*, *posture*) are mainly observed in individuals. Further, each of these broad types is subdivided into a variety of atomic-level behaviors. The hi-

Fig. 2: Examples of the NVI dataset, showing that our dataset covers rich nonverbal signals in diverse social scenes. Bounding box annotations of individuals are marked by `green rectangles`. To enhance demonstration and clarity, `red arrows` and `numerical identifiers` are incorporated additionally.
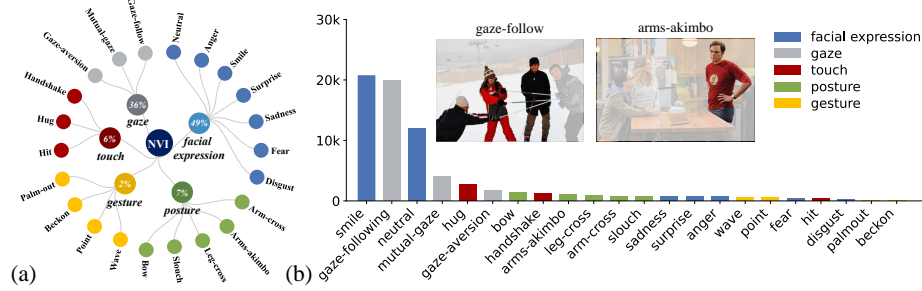

Fig. 3: Dataset statistics. (a) Nonverbal interaction taxonomy (§3). (b) Distribution of atomic-level nonverbal behaviors (§3).

erarchical taxonomy not only allows our dataset to capture nonverbal signals at multiple levels of abstraction, but also streamlines the annotation process (§3).

– *Gaze*, vital for inferring other's intentions [14], often serves as the first form of communication for interactants. Following [16], we consider three specific gaze behaviors in NVI, *i.e.*, `gaze-following`, `mutual-gaze`, and `gaze-aversion`.
– *Touch*, also known as haptics, is capable of conveying emotions, establishing connections, and regulating interpersonal dynamics [66]. In NVI, we categorize touch into three atomic behaviors: `handshake`, `hug`, and `hit`.
– *Facial expressions* are indicators of people's emotional state, typically conveyed via movements of the lips, eyes, brows, cheeks, and furrows [67]. NVI includes: `neutral`, `anger`, `smile`, `surprise`, `sadness`, `fear`, and `disgust`.
– *Gestures* arise from hand movements, are part of our communicative repertoire from infancy. NVI contains four gestures: `wave`, `point`, `beckon`, and `palm-out`.
– *Postures* are configurations of human body when standing or sitting. In NVI, we study five classes: `arm-cross`, `leg-cross`, `slouch`, `arms-akimbo`, and `bow`.

Dataset Annotation. For each image in NVI, annotators are instructed to proceed with the following steps: **1)** identify each social group and its broad

interaction type t; **2)** determine the specific atomic-level interaction category associated to t; **3)** add bounding boxes for all individuals; **4)** label the location of the group as the minimal bounding box encompassing all individuals. Six volunteers are involved in the annotation process. To maintain annotation coverage and accuracy, annotations of each image are double-checked by a specialist.

Dataset Statistics. NVI contains 13,711 images in total, and we follow the protocol in PIC 2.0 to split them into 9,634, 1,418 and 2,659 for `train`, `val` and `test`, respectively. On average, NVI contains 3.6 human instances, involving in 1.88 gaze, 0.32 touch, 2.60 facial expression, 0.08 gesture, and 0.36 posture per image. The distribution of atomic-level nonverbal behaviors is shown in Fig. 3(b). As seen, `smile` is the most frequent behavior in NVI, while `beckon` rarely occurs.

4   NVI-DET Task

Task Definition. As summarized in Table 1, current research in nonverbal interaction understanding tends to examine different types of interactions independently, each within its own dataset. This hinders AI models from gaining a thorough understanding of complex human behaviors and poses challenges in properly assessing their true ability in real-world social events. NVI-DET is a new task to address this limitation, which encourages models to interpret a full range of nonverbal signals. It has a three-fold objective aimed at: 1) human individual detection, 2) social group detection, and 3) interaction discovery of each individual with their respective group. Formally, accomplishing NVI-DET demands the capability to identify all triplets with the form $\langle$`individual`, `group`, `interaction`$\rangle$. While seemingly similar to the objective of HOI-DET, NVI-DET is much more challenging due to the requirement of distinguishing various heterogeneous nonverbal signals, which are frequently subtle and appear concurrently.

Metric. Inspired by [45,65], we utilize mean Recall@K (mR@K) as our primary evaluation metric, which computes Recall@K (R@K) for each category and then averages all scores:

$$mR@K = \frac{1}{|\mathcal{S}^\tau|} \sum_{\tau_{iou} \in \mathcal{S}^\tau} \frac{1}{C} \sum_{c \in [1...C]} \left(\sum_{p \in P_c} \mathbb{1}_{\{p \ is \ \mathrm{TP}\}} \right) / |G_c|, \tag{1}$$

where $|\cdot|$ is the cardinality of a set, $\mathcal{S}^\tau = \{0.25, 0.5, 0.75\}$ and τ_{iou} is the IoU threshold for assigning predicted individuals and social groups to ground truth. And, C is the number of social interaction categories, P_c represents the set of predicted triplets corresponding to the social interaction type c, while G_c denotes the set of ground truth triplets associated with the same social interaction type c. This metric continuity reduces barriers to entry. In particular, there are two key reasons why mR@K is employed for our needs: 1) it is more robust to incomplete annotations in NVI as compared to mean Average Precision (mAP) [45], and 2) it rationally takes into account long-tailed distributions found in NVI (Fig. 3(b)), since it treats all categories equally. This is a significant advantage over the alternative R@K that exhibits reporting bias [49]. For thorough evaluation, we adopt mR@25, mR@50, mR@100, along with their average (AR).
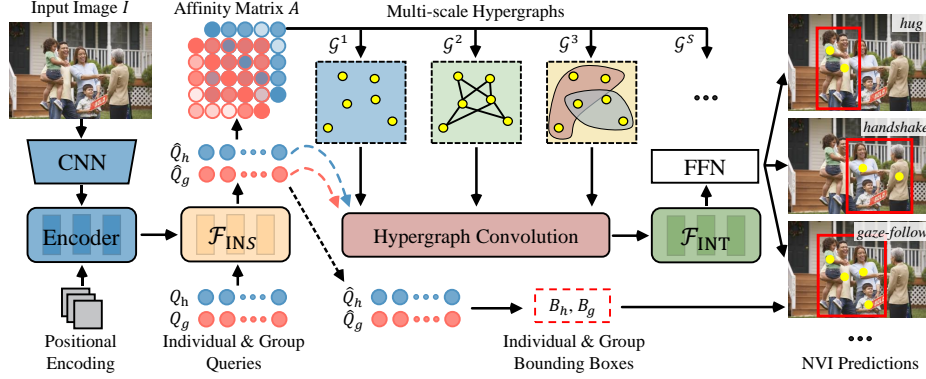
Fig. 4: Overall architecture of the proposed NVI-DEHR model. Given an image, the visual encoder is first applied to extract features, followed by an instance decoder that locates human-object pairs. Next, a dual multi-scale hypergraph is designed to model complex interactions between individuals and social groups via hypergraph convolutions. Lastly, an independent transformer decoder is employed to predict the nonverbal interaction categories for each individual-group pair (§5).

5   NVI-DEHR Model

5.1   Preliminary of Hypergraph

A hypergraph can be viewed as a higher-order form of graph whereby edges can link more than two nodes. Denote $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ as a hypergraph, where $\mathcal{V}$ is a set of vertices and $\mathcal{E}$ is a set of hyperedges. Each vertex $v \in \mathcal{V}$ is associated with an initial embedding $\boldsymbol{v}$. A hyperedge $e \in \mathcal{E}$ is a subset of $\mathcal{V}$, indicating the vertices it connects. For convenience, the corresponding hypergraph connectivity structure is usually represented by a binary incidence matrix $\boldsymbol{H} \in \{0,1\}^{|\mathcal{V}| \times |\mathcal{E}|}$, where $\boldsymbol{H}(v, e) = 1$ if node $v \in e$, otherwise $\boldsymbol{H}(v, e) = 0$. Moreover, we follow [18,23] to define the degree of hyperedge e and vertex v as $\delta(e) = \sum_{v \in \mathcal{V}} \boldsymbol{H}(v, e)$ and $d(v) = \sum_{e \in \mathcal{E}} \boldsymbol{H}(v, e)$, respectively. $\boldsymbol{D}_e \in \mathbb{R}^{|\mathcal{E}| \times |\mathcal{E}|}$ and $\boldsymbol{D}_v \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$ are the diagonal matrices of the hyperedge and vertex degrees.

5.2   Nonverbal Interaction Detection Hypergraph

Fig. 4 presents the architecture of our NVI-DEHR. It is built upon the popular encoder-decoder detection structure [9], consisting of a shared visual encoder for feature extraction, and two decoders for set-based prediction of NVI triplets ⟨individual, group, interaction⟩. A unique aspect of the model is that a dual multi-scale hypergraph is introduced to bridge the two decoders. It explicitly models complex interactional contexts among individuals and social groups, facilitating the learning of high-order cues essential for interaction recognition.

Visual Encoder. As normal, a visual encoder is adopted to map an input image $I \in \mathbb{R}^{H_0 \times W_0 \times 3}$ into a 3D feature map $I \in \mathbb{R}^{H \times W \times D}$. The encoder shares a similar structure as DETR [9], consisting of a conventional CNN backbone (*e.g.*, ResNet-50 [28]) for initial feature extraction, followed by a standard Transformer encoder [9] to further refine the features by integrating global contextual cues.

Instance Decoder. Given I, our model employs a query-based Transformer decoder to detect human individuals and social groups. The decoder $\mathcal{F}_{\text{INS}}$ takes two distinct sets of learnable queries as inputs, *i.e.*, $Q_h \in \mathbb{R}^{N \times C}$ and $Q_g \in \mathbb{R}^{N \times C}$, and transform them into output embeddings. Subsequently, they are independently decoded into bounding box coordinates for either individuals or groups through a feed forward network (FFN), $\mathcal{F}_{\text{FFN}}$. Here Q_h and Q_g serve separately for the decoding of human individual or social group boxes. N and C denote the number and dimension of these queries. The entire process can be written as:

$$\text{query updating:} \quad \hat{Q}_h, \hat{Q}_g = \mathcal{F}_{\text{INS}}(I, Q_h + P, Q_g + P),$$
$$\text{box prediction:} \quad B_h, B_g = \mathcal{F}_{\text{FFN}}(\hat{Q}_h, \hat{Q}_g). \tag{2}$$

Here we add a learnable position guided embedding $P \in \mathbb{R}^{N \times C}$ to the queries so as to assign the human query and group query at the same position as a pair [40]. $\hat{Q}_h = [\hat{h}_1, \hat{h}_2, \ldots, \hat{h}_N] \in \mathbb{R}^{N \times C}$ and $\hat{Q}_g = [\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_N] \in \mathbb{R}^{N \times C}$ are the updated queries, while B_h and B_g denote the box predictions for humans and groups.

Multi-Scale Hypergraph. Nonverbal signals are in nature complicated since they are subtle and often involve multiple participants. While direct composition of multiple individual features works well for tasks like HOI-DET [11, 35, 40], solving NVI-DET necessitates a more comprehensive relational understanding of individuals. To this end, NVI-DEHR performs multi-scale hypergraph learning to explore nonverbal interactions from a multi-granularity perspective.

Multi-scale hypergraph construction. We utilize two distinct multi-scale hypergraphs, $\mathcal{G}_h$ to model human-human relationships implying that the individuals participate in the same social group, and $\mathcal{G}_g$ to model group-group relationships implying that homologous groups are paired with different individuals. For clarity, we only explain the construction of $\mathcal{G}_h$, while $\mathcal{G}_g$ follows a same process. Concretely, $\mathcal{G}_h$ is comprised of a set of hypergraphs $\{\mathcal{G}_h^1, \mathcal{G}_h^2, \ldots, \mathcal{G}_h^S\}$, where $\mathcal{G}_h^s = (\mathcal{V}_h, \mathcal{E}_h^s)$ denotes a hypergraph at scale s. The vertex set $\mathcal{V}_h$ is consistent across all scales, and the vertex $v_i \in \mathcal{V}_h$ represents the i-th human query, *i.e.*, we have $|\mathcal{V}_h| = N$. The hyperedge set $\mathcal{E}_h^s = \{e_1^s, e_2^s, \ldots, e_{M_s}^s\}$ models group-wise relations with M_s hyperedges, each of which includes s vertices in $\mathcal{V}$. As in §5.1, the topology of each $\mathcal{G}_h^s$ is represented by an incidence matrix $H_h^s \in \mathbb{R}^{N \times M_s}$.

We define the hyperedges based on the distance-based construction strategy [23]. Initially, the embedding of vertex v_i is set to $v_i = \hat{h}_i \in \mathbb{R}^C$. Then, we compute an affinity matrix $A \in \mathbb{R}^{N \times N}$ for vertex pairs, wherein each element A_{ij} measures the similarity between v_i and v_j:

$$A_{ij} = v_i^\top v_j / \|v_i^\top\| \|v_j\|. \tag{3}$$

Based on the affinity matrix, we form hyperedges at various scales. For the *1st* scale (*i.e.*, $s = 1$), vertices are independently treated in the graph without

any edges and the incidence matrix $\boldsymbol{H}_h^0$ is thus an identity matrix. For other scales, each hyperedge e_i^s is formed as a cluster of vertices, which is identified by searching for a $s \times s$ sub-matrix within $\boldsymbol{A}$ that exhibits the highest density:

$$e_i^s = \arg\max_{\mathcal{O} \subseteq \mathcal{V}_h} ||\boldsymbol{A}_{\mathcal{O},\mathcal{O}}||_{1,1}, \quad s.t. \ |\mathcal{O}| = s \ \text{and} \ v_i \in \mathcal{O}. \tag{4}$$

Here $\|\cdot\|_{1,1}$ represents the $L_{1,1}$ matrix norm, *i.e.*, the summation of the absolute values of all elements in the matrix. The objective of Eq. 4 is to identify and connect the most closely related vertices. The first constraint confines the size of each cluster, while the second ensures the inclusion of v_i in the group. The optimization problem is tackled via a vertex-centric greedy algorithm, which, at each iteration i, selects the vertex v_i first and then includes additional $s-1$ vertices based on their affinity to v_i. In this way, for scales $s>1$, we have $M_s = N$.

Multi-scale hypergraph learning. With multi-scale hypergraphs $\mathcal{G}_h$ and $\mathcal{G}_g$, we perform message exchange among vertices through a sequence of L hyperedge convolutional layers [18]. For each scale s, the convolution operation at layer $l \in [L]$ can be formulated as:

$$\boldsymbol{V}_h^{s,(l)} = (\boldsymbol{D}_{h,v}^s)^{-\frac{1}{2}} \boldsymbol{H}_h^s (\boldsymbol{D}_{h,e}^s)^{-1} \boldsymbol{H}_h^{s\top} (\boldsymbol{D}_{h,v}^s)^{-\frac{1}{2}} \boldsymbol{V}_h^{s,(l-1)} \theta_h^{s,(l)}, \tag{5}$$

$$\boldsymbol{V}_g^{s,(l)} = (\boldsymbol{D}_{g,v}^s)^{-\frac{1}{2}} \boldsymbol{H}_g^s (\boldsymbol{D}_{g,e}^s)^{-1} \boldsymbol{H}_g^{s\top} (\boldsymbol{D}_{g,v}^s)^{-\frac{1}{2}} \boldsymbol{V}_g^{s,(l-1)} \theta_g^{s,(l)}, \tag{6}$$

where Eq. 5 and Eq. 6 are applied for human hypergraph and group hypergraph, respectively. The $\theta_h^{s,(l)}$ and $\theta_g^{s,(l)}$ are learnable parameters of the l-th layer at scale s. $\boldsymbol{D}_{\cdot,v}^s / \boldsymbol{D}_{\cdot,e}^s$ denote diagonal matrices of vertex and hyperedge degrees for either $\mathcal{G}_h^s$ or $\mathcal{G}_g^s$, which are computed from corresponding incidence matrices (§5.1). $\boldsymbol{V}_h^{s,(0)} / \boldsymbol{V}_g^{s,(0)}$ are matrices with initial vertex embedding in $\mathcal{G}_h^s / \mathcal{G}_g^s$. After L convolutions, the final embedding for each vertex is obtained by aggregating information across various scales via a multilayer perceptron (MLP):

$$\begin{aligned} \boldsymbol{F}_h &= \mathrm{MLP}([\boldsymbol{V}_h^{1,(L)}, \boldsymbol{V}_h^{2,(L)}, \dots, \boldsymbol{V}_h^{S,(L)}]) \quad \in \mathbb{R}^{N \times C}, \\ \boldsymbol{F}_g &= \mathrm{MLP}([\boldsymbol{V}_g^{1,(L)}, \boldsymbol{V}_g^{2,(L)}, \dots, \boldsymbol{V}_g^{S,(L)}]) \quad \in \mathbb{R}^{N \times C}, \end{aligned} \tag{7}$$

where $\boldsymbol{F}_h$ and $\boldsymbol{F}_g$ are the final embedding matrices of all human and group vertices, respectively. $'[\cdot, \cdot]'$ denotes tensor concatenation.

Interaction Decoder. Last, we leverage an independent query-based Transformer decoder to predict the nonverbal interaction categories for each individual-group pair. Instead of random query initialization, we propose to create nonverbal interaction query $\boldsymbol{Q}_n \in \mathbb{R}^{N \times C}$ dynamically based on high-order features of individuals $\boldsymbol{F}_h$ and groups $\boldsymbol{F}_g$:

$$\boldsymbol{Q}_n = (\boldsymbol{F}_h + \boldsymbol{F}_g)/2. \tag{8}$$

Note that these two types of features can be directly added, since they are position-aligned in Eq. 2. Subsequently, the interaction decoder takes the query $\boldsymbol{Q}_n$ alongside image feature $\boldsymbol{I}$ as input, and predict interaction categories as:

$$\boldsymbol{P} = \mathcal{F}_{\mathrm{INT}}(\boldsymbol{I}, \boldsymbol{Q}_n), \tag{9}$$

where $\boldsymbol{P}$ is the nonverbal interaction predictions for N individual-group pairs.

Table 2: **NVI-DET results** on NVI `val` and `test` (§6.1).

Method	val				test			
	mR@25	mR@50	mR@100	AR	mR@25	mR@50	mR@100	AR
m-QPIC [64]	**56.89**	69.52	78.36	68.26	59.44	71.46	80.07	70.32
m-CDN [87]	55.57	71.06	78.81	68.48	59.01	72.94	82.61	71.52
m-GEN-VLKT [40]	50.59	70.87	80.08	67.18	56.68	74.32	84.18	71.72
NVI-DEHR (Ours)	54.85	**73.42**	**85.33**	**71.20**	**59.46**	**76.01**	**88.52**	**74.67**

Table 3: **Results** for individual- and group-wise interactions on NVI `val` (§6.2).

Method	individual				group			
	mR@25	mR@50	mR@100	AR	mR@25	mR@50	mR@100	AR
m-QPIC [64]	**52.23**	66.09	75.98	64.77	**69.18**	78.62	84.85	77.55
m-CDN [87]	50.67	68.23	76.74	65.21	68.66	78.60	84.34	77.20
m-GEN-VLKT [40]	44.98	68.51	78.30	63.93	67.84	79.47	87.12	78.14
NVI-DEHR (Ours)	49.37	**70.04**	**83.82**	**67.74**	69.47	**82.45**	**89.35**	**80.42**

5.3   Detailed Network Architecture

Network Architecture. We utilize ResNet-50 [28] as the CNN backbone in all experiments. Following DETR [9], the Transformer encoder consists of six standard Transformer layers, while both the instance $\mathcal{F}_{\text{INS}}$ (*cf.* Eq. 2) and interaction $\mathcal{F}_{\text{INT}}$ (*cf.* Eq. 9) Transformer decoders incorporate three layers. By default, we set the number of queries N to 64, the number of channels C to 256, the number of hypergraph scales S to 5, and adopt $L=2$ hyperedge convolutional layers.

Training Objective. We follow [9, 40, 64] to perform end-to-end training by assigning a bipartite matching prediction with each groundtruth using the Hungarian algorithm. The loss function is: $\mathcal{L} = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_{\text{GIoU}} + \lambda_3 \mathcal{L}_c$, consisting of three parts: a $\mathcal{L}_1$ loss and a generalized IoU loss $\mathcal{L}_{\text{GIoU}}$ to assess localization accuracy, a focal loss $\mathcal{L}_c$ to evaluate interaction classification. The coefficients are empirically set to: $\lambda_1 = 2.5$, $\lambda_2 = 1$, $\lambda_3 = 2$, in accordance with QPIC [64].

Reproducibility. Our model is implemented using PyTorch and trained on 4 NVIDIA GeForce RTX 3090 GPUs. Testing is carried out on the same machine.

6   Experiment

6.1   Experiment on NVI-DET

Competitors. To better benchmark NVI and verify the proposed model, we adapt three state-of-the-art HOI-DET approaches, *i.e.*, QPIC [64], CDN [87], GEN-VLKT [40], for nonverbal interaction detection, denoted as m-QPIC, m-CDN, and m-GEN-VLKT. We modify their interaction prediction head to align with our NVI-DET task. For m-GEN-VLKT that relies on CLIP, we modify its text prompt to the format of 'A photo of a person [nonverbal interaction]'.

Implementation Details. For fairness, we train all models for 90 epochs. The learning rate is set to be 1e-4 for the initial 60 epochs and decreased to 1e-5

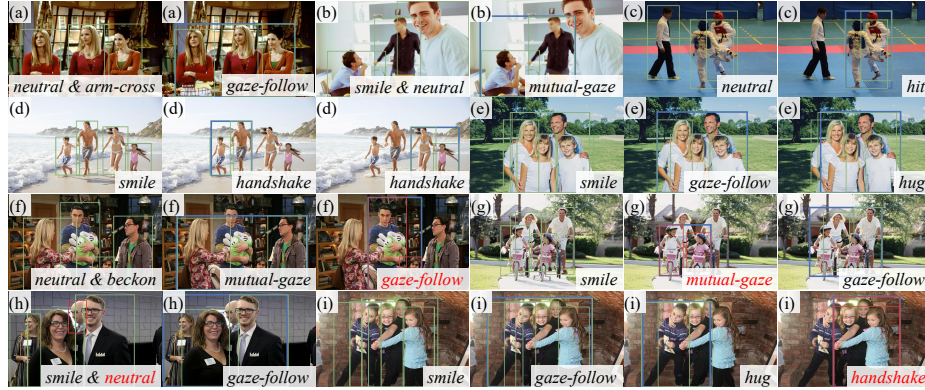**Fig. 5: Visualization of NVI-DET** results on NVI `val`. Predictions are generated with $\{K = 30, \tau_{iou} = 0.5\}$. Individuals correctly identified are in green bounding boxes, whilst those not detected are in red. Blue bounding boxes are utilized to denote social groups. True positive predictions are in black font, while false positives are in red (§6.1).

for the remaining epochs. AdamW is used as the optimizer. Common data augmentation techniques are applied, including random horizontal flipping, color jittering, and random scaling. Moving on to the procedural details during training, it is important to note that the bounding box of social group for "individual interaction" is identical to the individual bounding box. During inference, the ground-truths of triplets that encompass individual interactions do not entail bounding box of social group, akin to body motion categories in V-COCO [25].

Quantitative Results. Table 2 presents the benchmarking results on NVI `val` and `test`. As seen, m-QPIC, which is a simple adaptation of DETR from object detection to interaction detection, produces the worst performance among the comparative approaches due to the lack of explicit relational reasoning. m-CDN performs much better than m-QPIC by introducing an independent interaction decoder to account for interactional relations. Notably, the CLIP-based method m-GEN-VLKT performs worse in NVI-DET, revealing that transferring knowledge from visual language models to NVI-DET seems to be more difficult than to HOI-DET. NVI-DEHR surpasses all the baselines, reaching **71.20** and **74.67** AR on NVI `val` and `test`, respectively. Furthermore, we explore how the models perform for individual- and group-wise interactions. Table 3 shows our model consistently outperforms others in both sets across all metrics except for mR@25.

Qualitative Results. In Fig. 5, we show nonverbal interaction detection results of our model on NVI `test`. Our model can accurately localize the interactions, and successfully recognize different nonverbal interactions. For instance, in case (e), it competently forms a gaze-following group from three out of four individuals based on their gaze direction, and also astutely recognizes the man's hugging behavior in the image, although the bounding box for the hugging group is somewhat undersized. We also present some failure cases (shown in red), which may

Table 4: **HOI-DET results** on HICO-DET [10] `test` and V-COCO [25] `test` (§6.2).

Method	Backbone	HICO-DET			V-COCO	
		Full	Rare	Non-Rare	AP_{role}^{S1}	AP_{role}^{S2}
InteractNet [24][CVPR18]	R50-FPN	9.94	7.16	10.77	40.0	48.0
UnionDet [34][ECCV20]	R50	17.58	11.72	19.33	47.5	56.2
PPDM [39][CVPR20]	HG-104	21.73	13.78	24.10	-	-
HOTR [35][CVPR21]	R50	23.46	16.21	25.60	55.2	64.4
QPIC [64][CVPR21]	R50	29.07	21.85	31.23	58.8	61.0
ODM [72][ECCV22]	R50-FPN	31.65	24.95	33.65	-	-
UPT [88][CVPR22]	R50	31.66	25.94	33.36	59.0	64.5
CDN [87][NeurIPS21]	R50	31.78	27.55	33.05	62.3	64.4
Iwin [68][ECCV22]	R50-FPN	32.03	27.62	34.14	60.5	-
DOQ [58][CVPR22]	R50	33.28	29.19	34.50	63.5	-
GEN-VLK [40][CVPR22]	R50	33.75	29.25	35.10	62.4	64.4
HOICLIP [53][CVPR23]	R50	34.54	30.71	35.70	63.5	64.8
NVI-DEHR(Ours)	R50	**35.30**	**31.43**	**36.64**	**64.1**	**65.3**

Table 5: **Analysis of multi-scale hypergraph** on NVI `val` (§6.3).

S	mR@25	mR@50	mR@100	AR
1	53.39	69.81	81.90	68.37
2	53.44	70.31	82.62	68.79
3	53.52	70.45	83.92	69.30
4	54.21	73.34	84.36	70.64
5	**54.85**	**73.42**	**85.33**	**71.20**
6	54.59	73.11	85.24	70.98

Table 6: **Impact of hyperedge convolutions** on NVI `val` (§6.3).

L	mR@25	mR@50	mR@100	AR
0	53.50	69.44	81.71	68.22
1	53.76	71.74	83.61	69.70
2	**54.85**	**73.42**	**85.33**	**71.20**
3	54.36	72.47	84.85	70.56

be due to the ambiguity and subtlety of nonverbal interactions. In case (g), the mutual gaze of the two little girls could easily be mistaken for gaze aversion, as their gazes do not align perfectly. Likewise, the handshake behavior in case (i) could be confused with hugging, due to the unusual pose involved. It is notable that individuals heavily obscured in the image also present a challenge to the NVI-DET, as shown in case (h). These results highlight the challenging aspect of the NVI-DET task, meanwhile offering opportunities for future exploration.

6.2   Experiment on HOI-DET

Datasets. To fully assess the capability of our model, we evaluate it in the HOI-DET using two popular datasets, *i.e.*, V-COCO [25] and HICO-DET [10]. **Evaluation Metrics.** In accordance with conventions [24,25,57,93], the mean average precision (mAP) calculated on HOI triplets is used as the metric. **Implementation Details.** During training, we use the HOI-DET loss in [40,53] for network optimization. The model is trained with the same setting used in Section 6.1. Inspired by [30,40,58,74,86], our HOI classifier is initialized with the HOI embeddings generated by the text encoder of CLIP, and adopts a variant of the cross-attention module [53] to utilize visual representation from CLIP.

During inference, we first filter out low-confidence predictions and then apply non-maximum suppression to reduce overlapped predictions, following [40,58,86].
Quantitative Results. Table 4 presents the detection performance of our model against 12 state-of-the-art HOI-DET methods in the two datasets . Our model achieves the best performance in all settings. Particularly, in the HICO-DET dataset, our model outperforms the best-performing approach [53], with promising gains of 0.76, 0.72, and 0.94 for Full, Rare, and Non-Rare categories, respectively. In V-COCO, our approach achieves scores of **64.1** in terms of AP_{role}^{S1} and **65.3** in terms of AP_{role}^{S2}. These results verify the flexibility and generalizability of our model in handling various relational reasoning tasks.

6.3   Diagnostic Experiment

Number of Hypergraph Scales S. We first examine how the number S of scales in multi-scale hypergraph construction impacts model performance, as reported in Table 5. We observe that the model obtains an AR score of 68.37 at $S = 1$, with all humans and social groups are treated as independent vertices in the hypergraph. As S rises, model performance progressively improves, reaching a best performance of 71.20 AR at $S = 5$. It aligns with the research conducted by [17], which underscores that the majority of the speech in discussions involving 10 or more participants is produced by only the top 4-5 contributors.
Number of Hyperedge Convolutions L. In Table 6, we examine the effect of the number L of hypergraph convolution layers. Here $L = 0$ represents a variant of our model without hypergraph learning. It yields a score of 68.22 AR. By introducing more layers, model performance improves as L increases, and tends to stabilize at $L = 2$, with a significant boost to 71.20. We argue that increasing L leads to noise spreading in the graph, which could impair the final prediction.

7   Conclusion

This work makes a substantial step towards automatic interpretation of human nonverbal behaviors in everyday social environments. We challenge the conventional paradigm, which isolates social signals for standalone study, by instead examining a variety of common nonverbal signals (*i.e.*, *gaze*, *facial expression*, *gesture*, *posture*, *touch*) collectively. To open this avenue, we create a richly annotated dataset NVI, formalize the nonverbal interaction detection task NVI-DET, and devise a baseline model NVI-DEHR based on hypergraph learning. NVI-DEHR model achieves impressive performance in two interaction detection tasks, *i.e.*, NVI-DET and HOI-DET. Despite this, our experiments reveal that NVI-DET is considerably complex, regarding pluralistic relation among individuals, and we are now far from tackling this problem. We hope that our study will serve as valuable resources to foster more extensive exploration in this field.

References

1. Admoni, H., Scassellati, B.: Data-driven model of nonverbal behavior for socially assistive human-robot interactions. In: ICMI (2014) 2
2. Alameda-Pineda, X., Staiano, J., Subramanian, R., Batrinca, L., Ricci, E., Lepri, B., Lanz, O., Sebe, N.: Salsa: A novel dataset for multimodal group behavior analysis. IEEE TPAMI **38**(8), 1707–1720 (2015) 4
3. Albrecht, K.: Social intelligence: The new science of success (2006) 2, 4
4. Andrist, S., Tan, X.Z., Gleicher, M., Mutlu, B.: Conversational gaze aversion for humanlike robots. In: HRI. pp. 25–32 (2014) 2
5. Aviezer, H., Trope, Y., Todorov, A.: Body cues, not facial expressions, discriminate between intense positive and negative emotions. Science **338**(6111), 1225–1229 (2012) 2
6. Bai, S., Zhang, F., Torr, P.H.: Hypergraph convolution and hypergraph attention. Pattern Recognition **110**, 107637 (2021) 5
7. Bliege Bird, R., Ready, E., Power, E.A.: The social significance of subtle signals. Nature Human Behaviour **2**(7), 452–457 (2018) 4
8. Burgoon, J.K., Manusov, V., Guerrero, L.K.: Nonverbal communication. Routledge (2021) 3
9. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV (2020) 3, 8, 9, 11
10. Chao, Y.W., Liu, Y., Liu, X., Zeng, H., Deng, J.: Learning to detect human-object interactions. In: WACV (2018) 4, 5, 13
11. Chen, M., Liao, Y., Liu, S., Chen, Z., Wang, F., Qian, C.: Reformulating hoi detection as adaptive set prediction. In: CVPR (2021) 9
12. Cheng, Y., Huang, S., Wang, F., Qian, C., Lu, F.: A coarse-to-fine adaptive network for appearance-based gaze estimation. In: AAAI (2020) 4
13. Cheng, Y., Zhang, X., Lu, F., Sato, Y.: Gaze estimation by exploring two-eye asymmetry. IEEE TIP **29**, 5259–5272 (2020) 4
14. Emery, N.J.: The eyes have it: the neuroethology, function and evolution of social gaze. Neuroscience & Biobehavioral Reviews **24**(6), 581–604 (2000) 6
15. Fan, L., Chen, Y., Wei, P., Wang, W., Zhu, S.C.: Inferring shared attention in social scene videos. In: CVPR (2018) 4
16. Fan, L., Wang, W., Huang, S., Tang, X., Zhu, S.C.: Understanding human gaze communication by spatio-temporal graph reasoning. In: CVPR (2019) 2, 4, 6
17. Fay, N., Garrod, S., Carletta, J.: Group discussion as interactive dialogue or as serial monologue: The influence of group size. Psychological science **11**(6), 481–486 (2000) 14
18. Feng, Y., You, H., Zhang, Z., Ji, R., Gao, Y.: Hypergraph neural networks. In: AAAI (2019) 3, 5, 8, 10
19. Funes Mora, K.A., Monay, F., Odobez, J.M.: Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In: Proceedings of the Symposium on Eye Tracking Research and Applications (2014) 2
20. Gao, C., Xu, J., Zou, Y., Huang, J.B.: Drg: Dual relation graph for human-object interaction detection. In: ECCV (2020) 5
21. Gao, C., Zou, Y., Huang, J.B.: ican: Instance-centric attention network for human-object interaction detection. In: BMVC (2018) 5
22. Gao, Y., Feng, Y., Ji, S., Ji, R.: Hgnn+: General hypergraph neural networks. IEEE TPAMI **45**(3), 3181–3199 (2022) 3, 5

23. Gao, Y., Zhang, Z., Lin, H., Zhao, X., Du, S., Zou, C.: Hypergraph learning: Methods and practices. IEEE TPAMI **44**(5), 2548–2566 (2020) 5, 8, 9
24. Gkioxari, G., Girshick, R., Dollár, P., He, K.: Detecting and recognizing human-object interactions. In: CVPR (2018) 5, 13
25. Gupta, S., Malik, J.: Visual semantic role labeling. arXiv preprint arXiv:1505.04474 (2015) 3, 4, 12, 13
26. Hall, J.A., Horgan, T.G., Murphy, N.A.: Nonverbal communication. Annual Review of Psychology **70**, 271–294 (2019) 5
27. 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)
28. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 9, 11
29. Huang, B., Ju, J., Li, Z., Wang, Y.: Reconstructing groups of people with hypergraph relational reasoning. In: ICCV (2023) 5
30. Iftekhar, A., Chen, H., Kundu, K., Li, X., Tighe, J., Modolo, D.: What to look at and where: Semantic and spatial refined transformer for detecting human-object interactions. In: CVPR (2022) 13
31. Jiang, J., Wei, Y., Feng, Y., Cao, J., Gao, Y.: Dynamic hypergraph neural networks. In: IJCAI (2019) 5
32. Joo, H., Simon, T., Cikara, M., Sheikh, Y.: Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In: CVPR (2019) 4
33. Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., Torralba, A.: Gaze360: Physically unconstrained gaze estimation in the wild. In: CVPR (2019) 2, 4
34. Kim, B., Choi, T., Kang, J., Kim, H.J.: Uniondet: Union-level detector towards real-time human-object interaction detection. In: ECCV (2020) 5, 13
35. Kim, B., Lee, J., Kang, J., Kim, E.S., Kim, H.J.: Hotr: End-to-end human-object interaction detection with transformers. In: CVPR (2021) 9, 13
36. Knapp, M.L., Hall, J.A., Horgan, T.G.: Nonverbal communication in human interaction. Cengage Learning (2013) 3, 5
37. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: CVPR (2017) 2, 4
38. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3d points. In: CVPRW (2010) 2, 4
39. Liao, Y., Liu, S., Wang, F., Chen, Y., Qian, C., Feng, J.: Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In: CVPR (2020) 13
40. 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) 9, 11, 13, 14
41. Liu, D., Zhang, L., Wu, Y.: Ld-congr: A large rgb-d video dataset for long-distance continuous gesture recognition. In: CVPR (2022) 4
42. Liu, S., Wang, Z., Gao, Y., Ren, L., Liao, Y., Ren, G., Li, B., Yan, S.: Human-centric relation segmentation: Dataset and solution. IEEE TPAMI **44**(9), 4987–5001 (2021) 5
43. Liu, X., Li, Y.L., Lu, C.: Highlighting object category immunity for the generalization of human-object interaction detection. In: AAAI (2022)
44. Liu, Y., Chen, Q., Zisserman, A.: Amplifying key cues for human-object-interaction detection. In: ECCV (2020)
45. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: ECCV (2016) 7

46. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: CVPRW (2010) 4
47. Meeren, H.K., van Heijnsbergen, C.C., de Gelder, B.: Rapid perceptual integration of facial expression and emotional body language. Proceedings of the National Academy of Sciences **102**(45), 16518–16523 (2005) 2
48. Mehrabian, A., Ferris, S.R.: Inference of attitudes from nonverbal communication in two channels. Journal of Consulting Psychology **31**(3), 248 (1967) 2
49. Misra, I., Lawrence Zitnick, C., Mitchell, M., Girshick, R.: Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In: CVPR (2016) 7
50. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE TAC **10**(1), 18–31 (2017) 2, 4
51. Moore, C., Corkum, V.: Social understanding at the end of the first year of life. Developmental Review **14**(4), 349–372 (1994) 2
52. Narayana, P., Beveridge, R., Draper, B.A.: Gesture recognition: Focus on the hands. In: CVPR (2018) 2, 4
53. Ning, S., Qiu, L., Liu, Y., He, X.: Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models. In: CVPR (2023) 13, 14
54. Park, H., Jain, E., Sheikh, Y.: 3d social saliency from head-mounted cameras. In: NeurIPS (2012) 4
55. Pentland, A.: Honest signals: how they shape our world. MIT press (2010) 2
56. Philpott, J.S.: The relative contribution to meaning of verbal and nonverbal channels of communication: A meta-analysis. Unpublished master's thesis, University of Nebraska, Lincoln (1983) 2
57. Qi, S., Wang, W., Jia, B., Shen, J., Zhu, S.C.: Learning human-object interactions by graph parsing neural networks. In: ECCV (2018) 5, 13
58. Qu, X., Ding, C., Li, X., Zhong, X., Tao, D.: Distillation using oracle queries for transformer-based human-object interaction detection. In: CVPR (2022) 13, 14
59. Richmond, V.P.: Nonverbal behavior in interpersonal relations (2008) 3, 5
60. Sapir, E.: The unconscious patterning of behavior in society. In: Selected Writings of Edward Sapir in Language, Culture, and Personality (1949) 1
61. Schindler, K., Van Gool, L., De Gelder, B.: Recognizing emotions expressed by body pose: A biologically inspired neural model. Neural Networks **21**(9), 1238–1246 (2008) 4
62. Smith, L.B., Jayaraman, S., Clerkin, E., Yu, C.: The developing infant creates a curriculum for statistical learning. Trends in Cognitive Sciences **22**(4), 325–336 (2018) 2
63. Strabala, K., Lee, M.K., Dragan, A., Forlizzi, J., Srinivasa, S.S.: Learning the communication of intent prior to physical collaboration. In: IEEE RO-MAN (2012) 5
64. Tamura, M., Ohashi, H., Yoshinaga, T.: Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In: CVPR (2021) 11, 13
65. Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.: Unbiased scene graph generation from biased training. In: CVPR (2020) 7
66. Thayer, S.: History and strategies of research on social touch. Journal of Nonverbal Behavior **10**, 12–28 (1986) 6
67. Tian, Y.I., Kanade, T., Cohn, J.F.: Recognizing action units for facial expression analysis. IEEE TPAMI **23**(2), 97–115 (2001) 6

68. Tu, D., Min, X., Duan, H., Guo, G., Zhai, G., Shen, W.: Iwin: Human-object interaction detection via transformer with irregular windows. In: ECCV (2022) 13
69. Uluer, P., Kose, H., Gumuslu, E., Barkana, D.E.: Experience with an affective robot assistant for children with hearing disabilities. IJSR **15**(4), 643–660 (2023) 2
70. Ulutan, O., Iftekhar, A., Manjunath, B.S.: Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In: CVPR (2020) 5
71. Vinciarelli, A., Salamin, H., Pantic, M.: Social signal processing: Understanding social interactions through nonverbal behavior analysis. In: CVPRW. pp. 42–49 (2009) 4
72. Wang, G., Guo, Y., Wong, Y., Kankanhalli, M.: Chairs can be stood on: Overcoming object bias in human-object interaction detection. In: ECCV (2022) 13
73. Wang, H., Zheng, W.s., Yingbiao, L.: Contextual heterogeneous graph network for human-object interaction detection. In: ECCV (2020) 5
74. Wang, S., Duan, Y., Ding, H., Tan, Y.P., Yap, K.H., Yuan, J.: Learning transferable human-object interaction detector with natural language supervision. In: CVPR (2022) 13
75. Wang, W., Yang, Y., Pan, Y.: Visual knowledge in the big model era: Retrospect and prospect. Frontiers of Information Technology & Electronic Engineering (2024) 5
76. Wang, W., Zhang, Z., Qi, S., Shen, J., Pang, Y., Shao, L.: Learning compositional neural information fusion for human parsing. In: ICCV (2019) 5
77. Wang, W., Zhou, T., Qi, S., Shen, J., Zhu, S.C.: Hierarchical human semantic parsing with comprehensive part-relation modeling. IEEE TPAMI **44**(7), 3508–3522 (2021) 5
78. Wang, W., Zhu, H., Dai, J., Pang, Y., Shen, J., Shao, L.: Hierarchical human parsing with typed part-relation reasoning. In: CVPR (2020) 5
79. Wang, X., Zhang, X., Zhu, Y., Guo, Y., Yuan, X., Xiang, L., Wang, Z., Ding, G., Brady, D., Dai, Q., et al.: Panda: A gigapixel-level human-centric video dataset. In: CVPR (2020) 4
80. Wang, Y., Sun, Y., Huang, Y., Liu, Z., Gao, S., Zhang, W., Ge, W., Zhang, W.: Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In: CVPR (2022) 4
81. Wu, X., Chen, Q., Li, W., Xiao, Y., Hu, B.: Adahgnn: Adaptive hypergraph neural networks for multi-label image classification. In: ACM MM (2020) 5
82. Xia, L., Chen, C.C., Aggarwal, J.K.: View invariant human action recognition using histograms of 3d joints. In: CVPRW (2012) 2, 4
83. Xu, X., Zou, Q., Lin, X.: Adaptive hypergraph neural network for multi-person pose estimation. In: AAAI (2022) 5
84. Yadati, N., Nimishakavi, M., Yadav, P., Nitin, V., Louis, A., Talukdar, P.: Hypergcn: A new method for training graph convolutional networks on hypergraphs. In: NeurIPS (2019) 5
85. Yang, L., Li, L., Xin, X., Sun, Y., Song, Q., Wang, W.: Large-scale person detection and localization using overhead fisheye cameras. In: ICCV (2023) 5
86. Yuan, H., Jiang, J., Albanie, S., Feng, T., Huang, Z., Ni, D., Tang, M.: Rlip: Relational language-image pre-training for human-object interaction detection. In: NeurIPS (2022) 13, 14
87. 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) 11, 13
88. Zhang, F.Z., Campbell, D., Gould, S.: Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In: CVPR (2022) 13

89. Zhang, Y., Cao, C., Cheng, J., Lu, H.: Egogesture: A new dataset and benchmark for egocentric hand gesture recognition. IEEE TMM **20**(5), 1038–1050 (2018) 2, 4
90. 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)
91. Zhao, G., Huang, X., Taini, M., Li, S.Z., PietikäInen, M.: Facial expression recognition from near-infrared videos. Image and Vision Computing **29**(9), 607–619 (2011) 2, 4
92. Zhou, D., Huang, J., Schölkopf, B.: Learning with hypergraphs: Clustering, classification, and embedding. In: NeurIPS (2006) 3, 5
93. Zhou, T., Qi, S., Wang, W., Shen, J., Zhu, S.C.: Cascaded parsing of human-object interaction recognition. IEEE TPAMI **44**(6), 2827–2840 (2021) 5, 13
94. Zhou, T., Wang, W., Liu, S., Yang, Y., Van Gool, L.: Differentiable multi-granularity human representation learning for instance-aware human semantic parsing. In: CVPR (2021) 5
95. Zhou, T., Yang, Y., Wang, W.: Differentiable multi-granularity human parsing. IEEE TPAMI **45**(7), 8296–8310 (2023) 5