SocialGesture: Delving into Multi-person Gesture Understanding

Xu Cao[1], Pranav Virupaksha[1], Wenqi Jia[1], Bolin Lai[2], Fiona Ryan[2], Sangmin Lee[3†], James M. Rehg[1†]
[1] University of Illinois Urbana-Champaign [2] Georgia Institute of Technology
[3]Sungkyunkwan University
{xucao2,pranavv3,wenqij5,jrehg}@illinois.edu {bolin.lai,fkryan}@gatech.edu
sangmin.lee@skku.edu

Abstract

Previous research in human gesture recognition has largely overlooked multi-person interactions, which are crucial for understanding the social context of naturally occurring gestures. This limitation in existing datasets presents a significant challenge in aligning human gestures with other modalities like language and speech. To address this issue, we introduce SocialGesture, the first large-scale dataset specifically designed for multi-person gesture analysis. SocialGesture features a diverse range of natural scenarios and supports multiple gesture analysis tasks, including video-based recognition and temporal localization, providing a valuable resource for advancing the study of gesture during complex social interactions. Furthermore, we propose a novel visual question answering (VQA) task to benchmark vision language models' (VLMs) performance on social gesture understanding. Our findings highlight several limitations of current gesture recognition models, offering insights into future directions for improvement in this field. SocialGesture is available at huggingface.co/datasets/IrohXu/SocialGesture.

1. Introduction

Long before the development of spoken language in human history, deictic gestures played an important role in human social communication, serving as one of the earliest communication tools for expressing thoughts, emotions, and intentions. In contemporary communication, both verbal (e.g., speech) and non-verbal (e.g., gesture) cues work in concert. While spoken words express direct meanings, understanding the complete social context often requires more than just the words alone. Gestures play an important role in clarifying intention or points of emphasis, and augment spoken communication with additional social context. Theoretical analysis suggests that both speech and gesture originate from a shared cognitive system, with gestures representing imagery-driven data and speech representing symbolic data [6]. Together, they reflect a unified communicative intent.

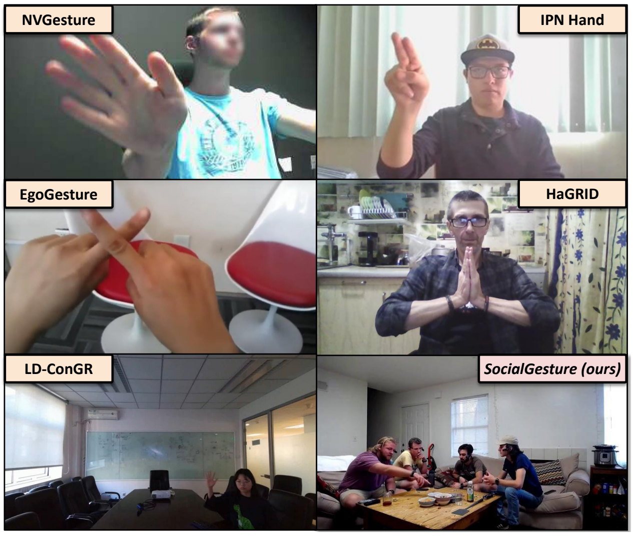


Figure 1. Example frames from six gesture datasets. SocialGesture is the only dataset featuring *multi-person interactions* and focusing on natural gestures with meaningful social communication.

Recent advancements in machine learning have driven research on the understanding of human social communication [27, 58], yet most of the latest social research predominantly focuses on linguistic modalities, often overlooking visual social cues like gestures [28]. While there have been machine learning works on human gestures, most of them have focused on gestures for device manipulation or sign languages [42, 44], and the problem of analyzing natural social gestures between multiple people has not received significant attention. A primary reason for this limitation is the lack of gesture data and annotations that capture multi-person interactions and the use of gestures in realistic social contexts. While recent datasets such as HaGRID [23] and LD-ConGR [31] address high-resolution and long-distance gesture collection (Figure 1), they remain confined to controlled,

[†]Corresponding author

Dataset	**Multi-person**	**Video Clips**	**Instances**	**Label**			
				Category	**Bbox**	**Relation**	**VQA**
Jester [37]	✗	148,092	148,092	✓	✗	✗	✗
NVGesture [39]	✗	1,532	1,532	✓	✓	✗	✗
EgoGesture [60]	✗	2,081	24,161	✓	✓	✗	✗
ChaLearn ConGD [52]	✗	22,535	47,933	✓	✓	✗	✗
IPN Hand [5]	✗	200	4,218	✓	✓	✗	✗
UAV gesture [42]	✗	119	-	✓	✓	✗	✗
iMiGUE [34]	✗	359	18,499	✓	✗	✗	✗
LD-ConGR [31]	✗	542	44,887	✓	✓	✗	✗
SocialGesture (Ours)	✓	9,889	42,533	✓	✓	✓	✓

Table 1. Comparison of SocialGesture with existing video-based gesture datasets. SocialGesture is the first work that addresses multi-person social interaction cases. Our dataset uniquely provides comprehensive annotations including gesture categories, bounding boxes (humans and objects), gesture relations (subject person and target person), and visual question answering pairs.

single-person settings, and do not contain the spontaneous, natural gestures between multiple people that characterize real-life social interactions.

To address this limitation, we introduce SocialGesture, a comprehensive dataset designed specifically for understanding gestures in multi-person social interactions. SocialGesture provides three key advantages over existing datasets:

1. It focuses on videos with multiple people, making it the first multi-person gesture dataset, in contrast to previous datasets which lack human-to-human interaction.
2. SocialGesture includes four types of comprehensive annotations: (a) Social gesture categories, identifying the specific class of a gesture (i.e., pointing, showing, giving, reaching) within social interactions; (b) Temporal-spatial annotations, identifying the appearance of social gestures temporally and spatially through time stamps and bounding boxes. (c) Interaction annotations, capturing interpersonal social dynamics that result from gestures; and (d) Visual question answering (VQA) annotations, which link gestures to descriptive content, facilitating the alignment of gestures with contextual meaning. All of these annotations offer a rich resource for developing models for social gesture understanding.
3. SocialGesture comprises a large volume of diverse and naturally occurring social interactions sourced from YouTube and Ego4D [17], and is the largest multi-person gesture dataset available.

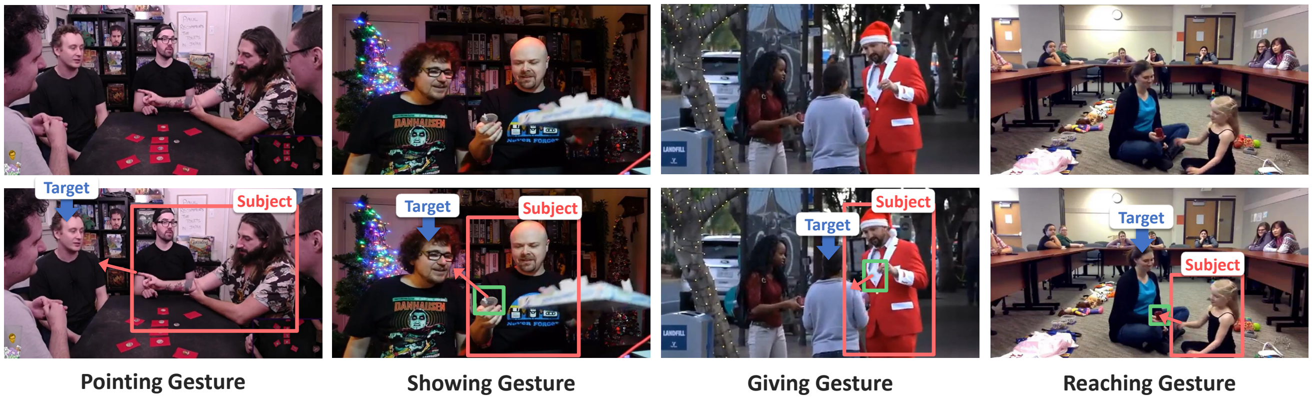
Based on our novel SocialGesture dataset, we introduce diverse gesture-related benchmark tasks from temporal localization and recognition to visual question answering. Our experiments reveal significant challenges in multi-person gesture understanding for current architectures. It is the first step toward building agents that have capabilities to seamlessly understand non-verbal gesture cues in social contexts. The major contributions of this paper are as follows.

- We introduce SocialGesture, the first dataset focusing on multi-person gesture interactions in natural social settings. Unlike existing datasets that are limited to single-person or controlled environments, our dataset captures the complexity of real-world gesture-based social communication.
- We provide comprehensive multi-level annotations including gesture categories, temporal-spatial localization, interaction dynamics, and visual question-answering pairs. This rich annotation scheme enables the development of models for understanding social gestures.
- We perform comprehensive experiments across gesture localization and recognition, evaluating a wide breadth of architectures and observe multiple challenging areas for future works. We further benchmark multiple large vision-language models on our proposed diverse Social Visual Question-Answering tasks.

2. Related Works

2.1. Gesture Datasets

As shown in Figure 1 and Table 1, the majority of gesture recognition datasets are video-based, with recent research emphasizing emotional state analysis, increased capture distance for improved model robustness, and egocentric gesture collection. Notable datasets include LD-ConGR, featuring 542 RGB-D videos representing 10 gestures [31], and HaGRID, which focuses on high-resolution image based gesture recognition across 18 hand gesture classes [23]. EgoGesture, with over 2 million frames from 50 subjects, is tailored for egocentric interactions with wearable devices [60], while iMiGUE [34] and SMG focus on micro-gesture recognition [10], containing 359 and 40 videos, respectively. Earlier datasets such as Jester [37], IPN Hand [5], ChaLearn IsoGD [52], and ChaLearn ConGD [52] have also made significant contributions, covering tasks like hand segmentation, and isolated or continuous gesture classification. In contrast, we are the first to provide a large-scale multi-person gesture



Pointing Gesture

Showing Gesture

Giving Gesture

Reaching Gesture

Figure 2. Examples of the four deictic gesture categories in SocialGesture with subject-target relationships. From left to right: pointing (directing attention), showing (presenting objects), giving (transfer intention), and reaching (acquisition intention) gestures. Red boxes indicate gesture initiators (subjects) and blue notations indicate targets.

dataset, and we are also the first to provide comprehensive coverage of the four main classes of deictic gestures(defined in [38]), which play a crucial role in social communication.

2.2. Social Understanding with Non-verbal Cues

Understanding social interactions has been long studied in the domain of natural language processing [15, 18, 25, 26, 46, 48]. However, the inherent multimodal nature of social interactions ensures that non-verbal elements (such as gestures, gaze and emotions) play an important role in communication [28], which are understudied in previous research [57]. Recently, several investigations have been conducted on these implicit non-verbal cues. Liu *et al.* [31] build a dataset to identify the gestures of single person from a long distance. Unfortunately, they only focus on isolated gestures of a single person, and fail to contextualize gesture recognition in a real-world social scenario. Non-verbal social understanding of multiple people demands more investigation [16, 21, 22, 41]. To this end, we make the first attempt to model multi-party social interactions by defining various social gesture understanding tasks, such as social gesture detection, recognition, and VQA, based on our dataset.

2.3. Foundation Models for Video Understanding

Video foundation models are designed to tackle a wide range of understanding and analysis tasks, including classification, temporal action localization, question answering, captioning, and more. Early approaches to video understanding relied on CNN blocks as visual encoders for video frames, connected to a prediction head or decoder [9, 13]. With the emergence of transformers, new architectures like MViTv2-B [30], VideoSwin [36], UniFormerV2 [29], and VideoMAE-2 [54] tokenize their visual inputs, enabling unified integration with other modalities [56]. Building on this, recently released video-based VLMs like Qwen2-VL [55], LLaVA-NEXT [32], and InternVL-2.5 [12] leverage pretrained visual encoders to extract rich representations. These models align visual tokens with textual tokens, enabling a deeper multimodal understanding that bridges the gap between video and language tasks. We provide the first investigation of the effectiveness of these models in understanding multi-party social gestures via a VQA task.

3. SocialGesture

3.1. Motivation

Social gestures comprise a fundamental set of human behavioral practices that facilitate interpersonal communication through hand and body movements in social contexts, distinct from sign language. According to [38], gestures can be classified into four primary categories: deictic, beat, iconic, and metaphoric. Deictic gestures further subdivide into pointing, showing, giving, and reaching. Beat gestures are rhythmic movements that emphasize certain words and phrases during speech. Iconic gestures visually represent concrete objects or actions (e.g., making a circle with fingers), while metaphoric gestures symbolize abstract concepts (e.g., making a circular motion to indicate repetition).

Among these categories, deictic gestures deserve particular attention due to their prevalence and significance in social interactions. While iconic and metaphoric gestures appear in specific contexts and beat gestures primarily serve prosodic functions, deictic gestures play a crucial role in establishing joint attention and facilitating object-mediated social interaction. Therefore, our dataset focuses on four fundamental deictic gestures (see Figure 2):

- **Pointing Gesture** involves directing attention to a specific entity in the social scene through finger extension. While commonly executed with an extended index finger and arm, these specific forms are not mandatory requirements. The crucial identifying factor is the communicative intent to guide others' attention to a particular target, whether

Video Content	Source	Proportion	Total Length	# of People	Description
Group social games	YouTube	44.51%	896 min	>=3	A group of people playing social games such as one night werewolf
	Ego4D	21.91%	441 min	>=4	
Variety entertainment	YouTube	22.31%	449 min	>=2	People performing activities such as pranks, challenges, and interviews
Educational play	YouTube	2.53%	51 min	>=2	Children engaged in educational activities with adult guidance
Product reviews	YouTube	2.53%	51 min	>=2	People recommending products such as advent calendars
Party & dinner	YouTube	3.63%	73 min	>=7	A group of people having party & dinner together
Group cooking	YouTube	2.58%	52 min	>=3	People engaged in cooking activities together

Table 2. SocialGesture contains diverse videos including different multi-person social interactions from YouTube and Ego4D [17].

person or object. This distinguishes pointing from incidental hand movements or other gesture types, particularly reaching gestures where the primary intent differs.

- **Showing Gesture** involves presenting an object for others to see it by manipulating or orienting it towards observers. Unlike pointing, the object itself is the focus of attention, and it does not need to be fully supported by the subject, such as when tilting or sliding an object to make its contents visible. The key criterion for a showing gesture is that the object is being presented for visual inspection, regardless of whether the target person actually looks at it or where the subject is directing their gaze.
- **Giving Gesture** involves manipulating an object with the intention of transferring it to another person's possession. It differs from pointing because the object originates with the subject, and from showing because the subject is inviting the other person to take possession of the object. The key is the subject's intention to transfer the object, regardless of whether the recipient actually takes it.
- **Reaching Gesture** manifests as a hand extension toward an object, expressing desire for possession or requesting transfer. It is characterized by full finger extension and often includes pre-grasp configuration and associated body movements like leaning forward. Reaching differs from showing and giving because the object is not in the subject's possession, and it differs from pointing because the subject is expressing a desire to possess the object, not just to draw attention to it. The key aspect of reaching is the intent to obtain the object, regardless of whether the subject can physically retrieve it.

3.2. Data Acquisition

Data collection. We collect raw video data for SocialGesture from various sources, including YouTube channels and Ego4D. Table 2 represents the overall composition of the dataset. Video selection is based on four key criteria: (1) Video quality – we manually check each video to ensure high resolution; (2) Number of people – each video is carefully selected to include between two and ten people, ensuring clear visibility of gestures; (3) Video length – the total length of selected videos ranges from 2 to 30 minutes, with each gesture lasting between 1 to 15 seconds; and (4) Scene diversity – to enhance generalization, we select videos featuring diverse races, genders, and ages.

Pre-processing. For videos longer than 10 minutes, we cut them into video clips of 5 minutes in length. All videos are then converted to 720p (1280×720) with 30 FPS in the initial curation and are then mapped to 360p (640×360) with 5 FPS for further annotation.

Gesture annotation. Each gesture in the SocialGesture dataset is defined as a sequence of video frames where an individual performs an arm or hand movement corresponding to one of four specified deictic gestures: pointing, showing, giving, and reaching. The annotations capture both temporal and spatial aspects of these social interactions through a comprehensive multi-level annotation framework. For temporal annotation, we identify the complete sequence of frames containing a gesture, from initial movement through completion. Within this sequence, we designate a key frame that best captures the defining moment of the gesture - typically when the gesture reaches its most distinctive configuration. This key frame serves as an anchor point for additional annotations. Spatial annotation includes precise bounding box coordinates [x1, y1, x2, y2] for both the gesture initiator and the target. The initiator bounding box encompasses the person performing the gesture, while the target bounding box identifies either the person or object toward which the gesture is directed. These spatial annotations enable analysis of the geometric relationships between interaction participants. Our annotation framework also includes natural language descriptions of each gesture's social context, capturing the broader interaction dynamics beyond pure spatiotemporal coordinates.

3.3. Benchmark Tasks

3.3.1. Social Gesture Temporal Localization

Temporal action localization is a challenging task that requires not only identifying the temporal intervals of all detected gesture instances but also estimating corresponding confidence values. Unlike standard classification tasks, this process involves an end-to-end approach in which classification is integrated within temporal localization, adding considerable complexity. For example, even if a gesture is accurately localized in time, it is treated as a false detection if assigned an incorrect class label.

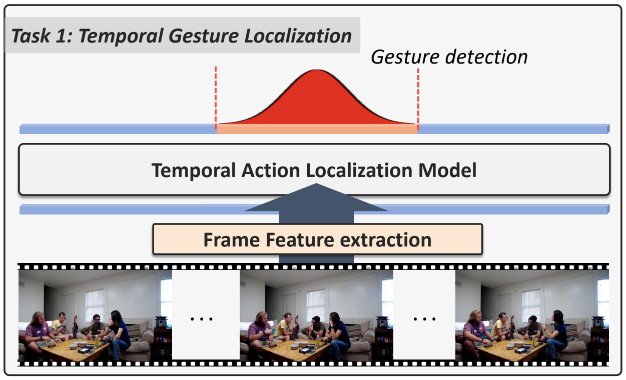

Task 1: Temporal Gesture Localization
Gesture detection
Temporal Action Localization Model
Frame Feature extraction
...
...

Figure 3. Temporal gesture localization task for social gestures

Figure 3 illustrates a general approach to temporal action localization. Each video is initially divided into multiple overlapping video clips, denoted as $X = [X_0, X_1, \ldots, X_K]$, where $X_k \in \mathbb{R}^{T \times H \times W \times C}$. Feature extractors such as I3D or VideoMAE are then employed to extract multiscale feature representations, denoted as $Z_k \in \mathbb{R}^D$, for each video clip. These features are subsequently concatenated into a single representation, $Z \in \mathbb{R}^{T \times D}$. The objective of the temporal action localization model is to generate sequence labels for each video clip.

This task is also benchmarked by several well-known datasets, including THUMOS14 [20] and ActivityNet-1.3 [8]. However, our task presents additional challenges. In SocialGesture videos, multiple individuals are often present simultaneously, and gestures tend to be small, subtle portions of larger actions, making it uniquely challenging to detect and classify gestures accurately.

3.3.2. Social Gesture Recognition

Given the complexity of social gesture localization, we also divide long videos into shorter clips of 2-5 seconds and introduce the social gesture recognition task. This task can be approached in multiple ways. The first approach is isolated gesture recognition, where a long video is segmented into discrete gesture clips. Video action recognition models are then trained to classify the gestures in each clip independently. Alternatively, gestures can be identified continuously throughout the entire video, from beginning to end. Although most action localization methods we discussed in the previous section can typically handle this task, most existing video feature extraction pre-processing methods struggle with consistent multi-person action localization. Therefore, sliding windows remain a common strategy for this purpose: the video is processed in overlapping segments of a predefined window size, with each window analyzed by the model to calculate gesture confidence scores.

We divide recognition into two subtasks (See Figure 4):

- **Task 2-1: Gesture vs Non-gesture Classification** This task aims to distinguish between social gesture and non-gesture video clips. The input is a short video clip $X \in \mathbb{R}^{T \times H \times W \times C}$, where $T < 16$, and the output is a label $\hat{Y}$ indicating whether a gesture is present.

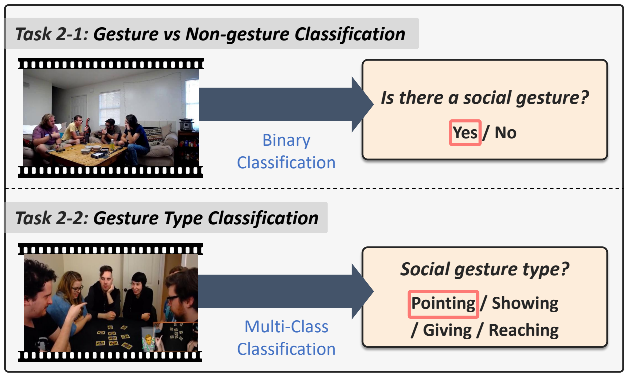

Task 2-1: Gesture vs Non-gesture Classification
Binary Classification
Is there a social gesture?
Yes / No
Task 2-2: Gesture Type Classification
Multi-Class Classification
Social gesture type?
Pointing / Showing / Giving / Reaching

Figure 4. Gesture recognition tasks for social gestures

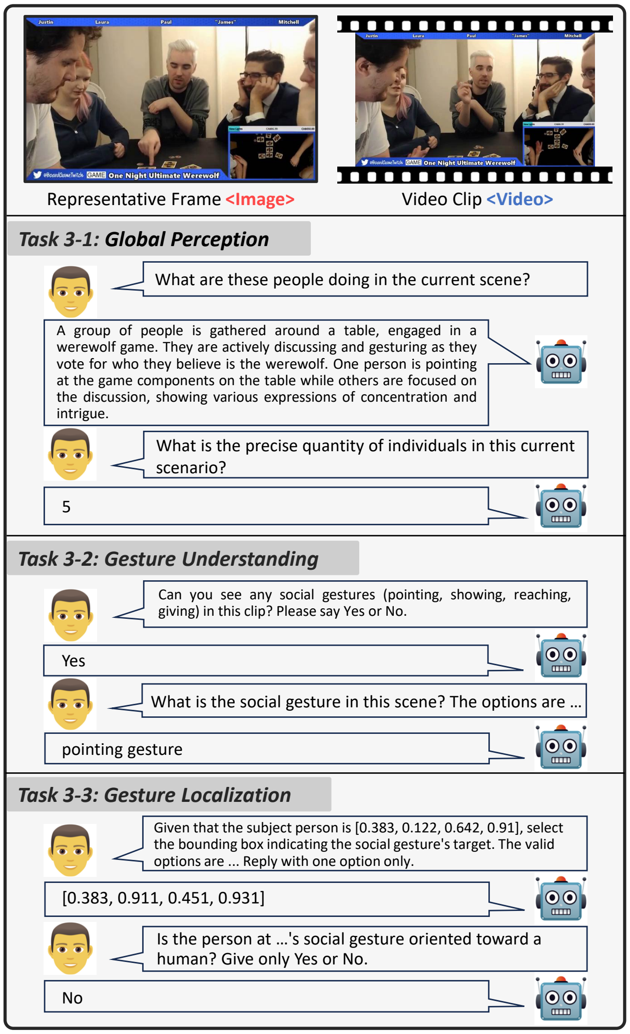
- **Task 2-2: Gesture Type Classification.** This task focuses on classifying video clips containing one of four social gestures: pointing, showing, giving, or reaching. The input and output structures are consistent with those used in the binary classification task. We further explore whether using cropped bounding boxes for each individual within a scene improves the accuracy of social gesture classification. The input is a short video clip cropped by the maximum region of the initiator of a social gesture and the output is a label $\hat{Y}$ indicating four types of social gesture.

For each task, we also strive to sample the training set to achieve a balanced label distribution. This adjustment is necessary because pointing gestures are significantly more frequent than showing, giving, and reaching in real-world communication contexts.

3.3.3. Social Gesture Visual Question Answering

Prior to our work, there has been very limited research on designing video captioning and Visual Question Answering (VQA) specifically for human gesture datasets [49]. This gap may be due to the lack of multi-person interactions in existing video datasets, reducing the necessity for enhanced alignment between the gesture and language modalities. To fill this gap, we introduce the first VQA task focused on human gestures (Figure 5). Our VQA can be divided into three subtasks: (1) Global Perception, (2) Gesture Understanding, and (3) Gesture Localization.

The Global Perception task serves as a foundational test to assess whether VLMs can comprehend basic events in short video clips, including scene descriptions and counting the number of humans present. We generate data for this task through a semi-automated QA generation pipeline based on GPT-4o [2], similar to the instruction-tuning data pair generation in LLaVA [33]. Initially, we manually create short descriptions for each video series, including the social background. We prompt these descriptions with the key frame of the video clip into the GPT-4o pipeline to generate QA pairs for human counting and scene description. Each QA



Task 3-1: Global Perception

What are these people doing in the current scene?

A group of people is gathered around a table, engaged in a werewolf game. They are actively discussing and gesturing as they vote for who they believe is the werewolf. One person is pointing at the game components on the table while others are focused on the discussion, showing various expressions of concentration and intrigue.

What is the precise quantity of individuals in this current scenario?

5

Task 3-2: Gesture Understanding

Can you see any social gestures (pointing, showing, reaching, giving) in this clip? Please say Yes or No.

Yes

What is the social gesture in this scene? The options are ...

pointing gesture

Task 3-3: Gesture Localization

Given that the subject person is [0.383, 0.122, 0.642, 0.91], select the bounding box indicating the social gesture's target. The valid options are ... Reply with one option only.

[0.383, 0.911, 0.451, 0.931]

Is the person at ...'s social gesture oriented toward a human? Give only Yes or No.

No

Figure 5. The question-answer pairs of SocialGesture. We omit the options of each question in the figure. The bounding box defined by [top-left x, top-left y, bottom-right x, bottom-right y])]. The definition will be provided together with the system prompts.

pair is then reviewed by a human annotator. Figure 5 shows the example of global perception (Task 3-1). Human counting is used as one VQA subtask in the VLM benchmarking experiments.

For Gesture Understanding, we utilize gesture classification annotations to create questions related to social gesture classification (Figure 5, Task 3-2). This results in two types of questions: (1) detecting whether a social gesture occurs, and (2) classifying between the four types of social gestures.

Gesture Localization leverages bounding boxes from our annotations to generate questions that test the model's spatial and temporal comprehension. These questions include providing the bounding box localization of the initiator, then determining whether the gesture's target is a human, and locating the target. We present an example of gesture localization (Task 3-3) in Figure 5. These questions are designed to challenge temporal-spatial reasoning in a multi-person setting, requiring analysis to produce the correct responses.

Feature	0.3	0.4	0.5	0.6	0.7	Avg mAP
I3D [9]	24.85	16.31	9.31	2.22	0.96	10.73
R(2+1)D [51]	14.38	10.25	7.23	2.81	1.77	7.29
VideoMAEV2 [54]	**27.23**	**25.05**	**13.33**	**5.28**	**2.76**	**14.73**

Table 3. Temporal action localization for four social gestures by different feature extractors with ActionFormer [59].

Stride	0.3	0.4	0.5	0.6	0.7	Avg mAP
16	11.09	8.60	5.52	2.67	1.82	5.94
8	24.85	16.31	9.31	2.22	0.96	10.73
4	**31.64**	**29.13**	**19.30**	**13.77**	**2.11**	**19.19**

Table 4. Explore the influence of stride for I3D and ActionFormer.

4. Experiments

4.1. Experimental Setup

We split the raw videos (long video) into training set (293 videos) and test set (79 videos), and then cut each long video into video clips. The 9,889 video clips used in video action recognition and social gesture VQA tasks come from the same train-test split in the raw videos. We also cut another 4,304 video clips from the raw videos as the non-gesture class. All experiments are conducted with 2 NVIDIA L40S GPUs and 2 H100 GPUs to make sure the batch size is 16 and learning rate is 5e-4 for all models. For all methods in Table 3, 5, 6, and 7, we apply the same data augmentation strategies used in [29, 50]. To solve the class imbalanced issue, we also adjust the size of training set for each task.

4.2. Social Gesture Temporal Localization

Baselines. We used two-stream I3D [9], R(2+1)D [51], and VideoMAEV2 [54] with the same sliding window of 16 frames and different stride sizes for feature extraction. Then, these video features are used as input for ActionFormer [59] during model training. We used mAP@[0.3:0.1:0.7] as the main evaluation metric and also reported the average mAP.

Results and Findings. We observe that almost all feature extractors underperform on this task in Table 3. Notably, despite the strong performance of features from VideoMAE V2 [54] in other action localization datasets, such as THUMOS 2014 [20] and ActivityNet [8], they only achieve an average mAP of 14.73 for social gesture recognition. To investigate the effect of the stride size in the sliding window on the average mAP, we conducted an ablation study in Table 4 using different stride sizes of the I3D feature in SocialGesture. We found that reducing the stride from 16 to 4 led to an increase in average mAP; however, the overall results remained insufficient. This shortfall can be attributed to the

Model	Pretrain	Param	Acc (%)
TSN-R50 [53]	Kinetics-400 [24]	24M	78.77
TANet-R50 [35]	Kinetics-400 [24]	26M	71.22
SlowFast-R50 [13]	Kinetics-400 [24]	35M	80.82
SlowFast-R101 [13]	Kinetics-400 [24]	63M	79.59
TimeSformer-L [7]	Kinetics-400 [24]	121M	78.71
MViTv2-B [30]	Kinetics-400 [24]	51M	83.29
VideoSwin-B [36]	Kinetics-400 [24]	88M	81.70
VideoSwin-L [36]	Kinetics-400 [24]	200M	83.44
UniFormerV2-B/16 [29]	CLIP [43]	115M	**84.43**
UniFormerV2-L/14 [29]	CLIP [43]	354M	80.93

Table 5. Experiments on binary classification for social gesture and non-gesture.

fact that all feature extractors were pretrained on datasets that lack multi-person interactions, causing the features to be poorly aligned with the specific requirements of our task.

4.3. Social Gesture Recognition

Baselines. Our baselines for social gesture recognition includes different CNN-based video models such as TSN [53], TANet [35], SlowFast [13], and Vision Transformer based models such as TimeSformer [7], MViTv2 [30], VideoSwin [36], UniFormerV2 [29]. All these models are pretrained with Kinetics-400 [24] dataset or using CLIP encoder [43]. We use Accuracy for gesture and non-gesture binary classification in Table 5 and Top 1 Accuracy for four social gesture classification in Table 6, and 7.

Results and Findings. The results in Table 5 indicate that most existing action recognition models can differentiate social gestures and non-gesture from short video clips. Despite fine-tuning these models using over 10,000 video clips for the binary classification task, their performance remains suboptimal. The SOTA baseline, UniFormerV2-B/16, achieves an accuracy of only 84.43%. Notably, Transformer-based models generally outperform CNN-based models.

Table 6 presents results for the four-class social gesture classification task, revealing that all models struggle to effectively classify these gestures—despite the simplicity of the task for human annotators. The highest performance comes from VideoSwim-L, achieving only 56.18%. Given that individual gestures can be challenging to recognize in multi-person environments, we conducted further experiments, as shown in Table 7. Specifically, we extracted the region corresponding to each individual using ground truth bounding boxes and fine-tuned the same models with these per-person inputs. UniFormerV2-B/16 achieves the best performance of 64.72% top 1 accuracy. This may be because social gestures inherently involve fine-grained and subtle movements. Consequently, recognizing such gestures is difficult from both the extracted per-person region and the full-frame perspective. Addressing the challenge of social gesture understanding requires advanced models capable of disentangling the relationship between the subject and the target person.

Model	Pretrain	Param	Top1 Acc (%)
TSN-R50 [53]	Kinetics-400 [24]	24M	54.83
TANet-R50 [35]	Kinetics-400 [24]	26M	45.39
SlowFast-R50 [13]	Kinetics-400 [24]	35M	45.17
SlowFast-R101 [13]	Kinetics-400 [24]	63M	46.07
TimeSformer-L [7]	Kinetics-400 [24]	121M	53.03
MViTv2-B [30]	Kinetics-400 [24]	51M	37.98
VideoSwin-B [36]	Kinetics-400 [24]	88M	53.93
VideoSwin-L [36]	Kinetics-400 [24]	200M	**56.18**
UniFormerV2-B/16 [29]	CLIP [43]	115M	55.51
UniFormerV2-L/14 [29]	CLIP [43]	354M	50.34

Table 6. Experiments on classification for four social gestures.

Model	Pretrain	Param	Top1 Acc (%)
TSN-R50 [53]	Kinetics-400 [24]	24M	58.43
TANet-R50 [35]	Kinetics-400 [24]	26M	46.29
SlowFast-R50 [13]	Kinetics-400 [24]	35M	45.17
SlowFast-R101 [13]	Kinetics-400 [24]	63M	42.92
TimeSformer-L [7]	Kinetics-400 [24]	121M	63.60
MViTv2-B [30]	Kinetics-400 [24]	51M	35.28
VideoSwin-B [36]	Kinetics-400 [24]	88M	60.00
VideoSwin-L [36]	Kinetics-400 [24]	200M	63.60
UniFormerV2-B/16 [29]	CLIP [43]	115M	**64.72**
UniFormerV2-L/14 [29]	CLIP [43]	354M	60.45

Table 7. Experiments on classification for four social gestures after extract each subject person's bounding box.

4.4. SocialGesture VQA

From the results of previous tasks, we found that distinguishing social gestures is still very challenging for current SOTA methods in gesture localization and recognition. This difficulty highlights the unique nature of social gesture, where models need to have a unified view of multi-person social interactions and can distinguish details that are hard to detect. This is something traditional video analysis models still struggle with [1, 14]. Inspired by recent progress in multimodal video grounding and video-based VQA [11, 40], we aim to leverage the reasoning capabilities of the latest VLMs and evaluate their performance on this complex task.

Baselines. To assess performance on the SocialGesture VQA benchmark, we chose several SOTA publicly available VLMs, including Qwen2-VL [55], Qwen2.5-VL [4], InternVL-2.5 [12], and LLAVA-NEXT-Video (LLaVA 1.6) [32]. Additionally, we included closed-source multimodal LLMs such as GPT-4o-mini [2] and Claude-3.7-sonnet [3] for a broader comparison. Note that GPT-4o [2] was excluded, as it was used to generate some of the VQA instruction-following training data, which could lead to biased results.

Evaluation Metrics. Since all VQA pairs can be transformed into a multi-option output, we use rule-based accuracy metrics to evaluate the VQA tasks. Each model was

Model	Param	Global Perception (%)	Gesture Understanding (%)		Gesture Localization (%)	
		HumanCount@Acc	GestureDet@Acc	GestureClass@Acc	TargetLoc@Acc	TargetClass@Acc
Random Select	-	-	50.00	25.00	18.90	50.00
InternVL-2.5 [12]	2B	22.37	63.34	74.45	13.33	39.79
InternVL-2.5 [12]	8B	35.58	73.94	**81.80**	17.38	57.28
LLaVA-NeXT-Video [32, 61]	7B	30.63	71.92	34.08	5.39	59.90
Qwen2-VL [55]	7B	60.18	74.15	67.18	9.58	59.19
Qwen2.5-VL [4]	72B	**69.97**	**75.69**	54.82	**29.37**	**61.38**
Claude 3.7 Sonnet [3]	-	**63.27**	69.64	61.73	**25.48**	64.20
GPT-4o-mini [2]	-	53.40	**73.67**	**68.52**	21.29	**65.12**

Table 8. Experiments on zero-shot performance of SOTA VLMs, including closed sourced multimodal LLMs for Social Gesture VQA

Model	Input	Global Perception (%)	Gesture Understanding (%)		Gesture Localization (%)	
		HumanCount@Acc	GestureDet@Acc	GestureClass@Acc	TargetLoc@Acc	TargetClass@Acc
Qwen2-VL-7B	video	60.18	74.15	67.18	9.58	59.19
Qwen2-VL-7B + LoRA [19] SFT	video	82.55	33.07	84.52	49.21	41.19
Qwen2-VL-7B + Full SFT	video	83.64	33.07	81.13	51.25	38.51

Table 9. Experiments on SFT performance of SOTA VLMs, including closed sourced multimodal LLMs for Social Gesture VQA

given sufficient context for the task, followed by a question and multiple answer choices. Accuracy was determined by comparing the model's response with the correct answer. We evaluated models across all subtasks including global perception, gesture understanding, and gesture localization in Section 3.3.3. The accuracy (Acc) in Table 8 is defined as: $\mathrm{Acc} = \frac{1}{N}\sum_{k=1}^{N} \mathrm{MATCH}_k$, where N is the total number of VQA samples and MATCH_k is a binary value. If the option in the model output matches the ground truth option, it is 1, otherwise it is 0.

Results and Findings. In Table 8, the experimental results show that while most VLMs can handle global perception and gesture understanding tasks without finetuning, they all struggle significantly with gesture localization tasks. This trend is understandable, as localization tasks inherently involve more ambiguity compared to perception and understanding tasks. While there is not a very clear correlation between the number of parameters and performance, Qwen2.5-VL, the largest of our open sourced models, performed well overall. To explore the potential for improvement, we finetuned Qwen2-VL-7B in Table 9, which led to enhanced performance in most metrics, except those related to gesture binary classification and target identification. Performance drops in some metrics can be attributed to easy overfitting on binary cases. An additional noteworthy observation is that despite being pretrained on vast multimodal datasets, none of the SOTA VLMs achieved over 70% accuracy on the simplest task: counting humans in social interaction scenes.

These findings indicate that the SocialGesture tasks are significantly more challenging for VLMs than typical VQA benchmarks. This difficulty stems from the natural complexity of real-world social interaction videos used in our benchmark. Unlike controlled environments, our dataset contains realistic scenarios where people frequently overlap with each other, are partially visible, or extend beyond the camera frame, creating natural occlusion cases. In addition, there exists high interdependence between tasks in SocialGesture VQA, which makes it challenging. For instance, when models process such videos, they should first assess the overall scene despite these visual challenges, counting people and understanding actions. This process involves high-level reasoning and the ability to sequentially infer relationships—skills that are currently lacking in SOTA VLMs. Developing models capable of such contextual social visual reasoning remains a key challenge for future research.

5. Conclusion

We introduce SocialGesture, the first dataset and benchmark specifically designed for understanding and analyzing multi-person social gestures. We provide a comprehensive definition of tasks related to social gesture recognition, including a novel VQA dataset that encompasses global perception, gesture understanding, and spatial localization. To advance the field of human gesture and social interaction understanding, we benchmarked a variety of current state-of-the-art methods, including action recognition models and vision-language models (VLMs). Our findings reveal that multi-person social gesture understanding requires incremental visual reasoning—area where existing computer vision models fall short. This work offers insights into the complexities of multi-person social interactions and highlights the need for continued research into developing models capable of sophisticated visual reasoning. We hope that SocialGesture will motivate further innovations in gesture understanding.

Acknowledgment

We are thankful to the Keymakr for their great work in data annotation. This work was supported in part by the NRF grant funded by the Korea government(MSIT) (No. RS-2025-00563942).

References

[1] Qaisar Abbas, Mostafa EA Ibrahim, and M Arfan Jaffar. Video scene analysis: an overview and challenges on deep learning algorithms. *Multimedia Tools and Applications*, 77 (16):20415–20453, 2018. 7

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5, 7, 8

[3] Anthropic. Claude 3 family. `https://www.anthropic.com/news/claude-3-family`, 2024. Accessed: 2024-05-27. 7, 8

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 7, 8

[5] Gibran Benitez-Garcia, Jesus Olivares-Mercado, Gabriel Sanchez-Perez, and Keiji Yanai. Ipn hand: A video dataset and benchmark for real-time continuous hand gesture recognition. In *2020 25th international conference on pattern recognition (ICPR)*, pages 4340–4347. IEEE, 2021. 2

[6] Paolo Bernardis and Maurizio Gentilucci. Speech and gesture share the same communication system. *Neuropsychologia*, 44(2):178–190, 2006. 1

[7] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021. 7

[8] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 5, 6

[9] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 3, 6

[10] Haoyu Chen, Henglin Shi, Xin Liu, Xiaobai Li, and Guoying Zhao. Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis. *International Journal of Computer Vision*, 131(6):1346–1366, 2023. 2

[11] Qirui Chen, Shangzhe Di, and Weidi Xie. Grounded multi-hop videoqa in long-form egocentric videos. *arXiv preprint arXiv:2408.14469*, 2024. 7

[12] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 3, 7, 8

[13] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 3, 7

[14] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 7

[15] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*, 2019. 3

[16] Shreya Ghosh, Zhixi Cai, Abhinav Dhall, Dimitrios Kollias, Roland Goecke, and Tom Gedeon. Mrac track 1: 2nd workshop on multimodal, generative and responsible affective computing. *arXiv preprint arXiv:2409.07256*, 2024. 3

[17] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2, 4

[18] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, page 2122. NIH Public Access, 2018. 3

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 8

[20] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017. 5, 6

[21] Simindokht Jahangard, Zhixi Cai, Shiki Wen, and Hamid Rezatofighi. Jrdb-social: A multifaceted robotic dataset for understanding of context and dynamics of human interactions within social groups. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22087–22097, 2024. 3

[22] Wenqi Jia, Miao Liu, Hao Jiang, Ishwarya Ananthabhotla, James M Rehg, Vamsi Krishna Ithapu, and Ruohan Gao. The audio-visual conversational graph: From an egocentric-exocentric perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26396–26405, 2024. 3

[23] Alexander Kapitanov, Karina Kvanchiani, Alexander Nagaev, Roman Kraynov, and Andrei Makhliarchuk. Hagrid–hand gesture recognition image dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4572–4581, 2024. 1, 2

[24] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 7

[25] Bolin Lai, Hongxin Zhang, Miao Liu, Aryan Pariani, Fiona Ryan, Wenqi Jia, Shirley Anugrah Hayati, James Rehg, and Diyi Yang. Werewolf among us: Multimodal resources for modeling persuasion behaviors in social deduction games. *Association for Computational Linguistics: ACL 2023*, 2023. 3

[26] Bongseok Lee and Yong Suk Choi. Graph based network with contextualized representations of turns in dialogue. *arXiv preprint arXiv:2109.04008*, 2021. 3

[27] Sangmin Lee, Bolin Lai, Fiona Ryan, Bikram Boote, and James M Rehg. Modeling multimodal social interactions: new challenges and baselines with densely aligned representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14585–14595, 2024. 1

[28] Sangmin Lee, Minzhi Li, Bolin Lai, Wenqi Jia, Fiona Ryan, Xu Cao, Ozgur Kara, Bikram Boote, Weiyan Shi, Diyi Yang, et al. Towards social ai: A survey on understanding social interactions. *arXiv preprint arXiv:2409.15316*, 2024. 1, 3

[29] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Unlocking the potential of image vits for video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1632–1643, 2023. 3, 6, 7

[30] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4804–4814, 2022. 3, 7

[31] Dan Liu, Libo Zhang, and Yanjun Wu. Ld-congr: A large rgb-d video dataset for long-distance continuous gesture recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3304–3312, 2022. 1, 2, 3

[32] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 3, 7, 8

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 5

[34] Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10631–10642, 2021. 2

[35] Zhaoyang Liu, Limin Wang, Wayne Wu, Chen Qian, and Tong Lu. Tam: Temporal adaptive module for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13708–13718, 2021. 7

[36] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 3, 7

[37] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The jester dataset: A large-scale video dataset of human gestures. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 2

[38] David McNeill. *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992. 3

[39] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4207–4215, 2016. 2

[40] Fangzhou Mu, Sicheng Mo, and Yin Li. Snag: Scalable and accurate video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18930–18940, 2024. 7

[41] Yujia Peng, Jiaheng Han, Zhenliang Zhang, Lifeng Fan, Tengyu Liu, Siyuan Qi, Xue Feng, Yuxi Ma, Yizhou Wang, and Song-Chun Zhu. The tong test: Evaluating artificial general intelligence through dynamic embodied physical and social interactions. *Engineering*, 34:12–22, 2024. 3

[42] Asanka G Perera, Yee Wei Law, and Javaan Chahl. Uav-gesture: A dataset for uav control and gesture recognition. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 1, 2

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7

[44] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Sign language recognition: A deep survey. *Expert Systems with Applications*, 164:113794, 2021. 1

[45] Fiona Ryan, Ajay Bati, Sangmin Lee, Daniel Bolya, Judy Hoffman, and James M Rehg. Gaze-lle: Gaze target estimation via large-scale learned encoders. 2025. 1

[46] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *language*, 50(4):696–735, 1974. 3

[47] Yuehao Song, Xinggang Wang, Jingfeng Yao, Wenyu Liu, Jinglin Zhang, and Xiangmin Xu. Vitgaze: gaze following with interaction features in vision transformers. *Visual Intelligence*, 2(1):1–15, 2024. 1

[48] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373, 2000. 3

[49] Kosei Tanada, Shigemichi Matsuzaki, Kazuhito Tanaka, Shintaro Nakaoka, Yuki Kondo, and Yuto Mori. Pointing gesture understanding via visual prompting and visual question answering for interactive robot navigation. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*. 5

[50] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 6

[51] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 6

[52] Jun Wan, Yibing Zhao, Shuai Zhou, Isabelle Guyon, Sergio Escalera, and Stan Z Li. Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 56–64, 2016. 2

[53] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 7

[54] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023. 3, 6

[55] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3, 7, 8

[56] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024. 3

[57] Diyi Yang, Dirk Hovy, David Jurgens, and Barbara Plank. The call for socially aware language technologies, 2024. 3

[58] Ceng Zhang, Junxin Chen, Jiatong Li, Yanhong Peng, and Zebing Mao. Large language models for human-robot interaction: A review. *Biomimetic Intelligence and Robotics*, page 100131, 2023. 1

[59] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pages 492–510. Springer, 2022. 6

[60] Yifan Zhang, Congqi Cao, Jian Cheng, and Hanqing Lu. Egogesture: A new dataset and benchmark for egocentric hand gesture recognition. *IEEE Transactions on Multimedia*, 20(5):1038–1050, 2018. 2

[61] Y Zhang, B Li, H Liu, Y Lee, L Gui, D Fu, J Feng, Z Liu, and C Li. Llava-next: A strong zero-shot video understanding model. 2024. 8