Learning Triadic Belief Dynamics in Nonverbal Communication from Videos

Lifeng Fan*, Shuwen Qiu*, Zilong Zheng, Tao Gao, Song-Chun Zhu, Yixin Zhu
UCLA Center for Vision, Cognition, Learning, and Autonomy
{lfan, s.qiu, z.zheng}@ucla.edu, {tao.gao, sczhu}@stat.ucla.edu, yixin.zhu@ucla.edu
https://github.com/LifengFan/Triadic-Belief-Dynamics

Abstract

*Humans possess a unique social cognition capability [43, 20]; nonverbal communication can convey rich social information among agents. In contrast, such crucial social characteristics are mostly missing in the existing scene understanding literature. In this paper, we incorporate different nonverbal communication cues (*e.g.*, gaze, human poses, and gestures) to represent, model, learn, and infer agents' mental states from pure visual inputs. Crucially, such a mental representation takes the agent's belief into account so that it represents what the true world state is and infers the beliefs in each agent's mental state, which may differ from the true world states. By aggregating different beliefs and true world states, our model essentially forms "five minds" during the interactions between two agents. This "five minds" model differs from prior works that infer beliefs in an infinite recursion; instead, agents' beliefs are converged into a "common mind" [31, 47]. Based on this representation, we further devise a hierarchical energy-based model that jointly tracks and predicts all five minds. From this new perspective, a social event is interpreted by a series of nonverbal communication and belief dynamics, which transcends the classic keyframe video summary. In the experiments, we demonstrate that using such a social account provides a better video summary on videos with rich social interactions compared with state-of-the-art keyframe video summary methods.*

1. Introduction

"The human body is the best picture of the human soul."
— Ludwig Wittgenstein [32]

We live in a world with a plethora of animate and goal-directed agents [60], or at least it is how humans perceive and *construct* [49] the world in our *mental state* [24]. The iconic Heider-Simmel display [19] is a quintessential stimulus, wherein human participants are given videos of simple shapes roaming around the space. In this experiment, humans have a strong inclination to interpret the observed featureless motions composed of simple shapes as a story-telling description, such as a hero saving a victim from a bully. This social cognition account of human vision is largely missing in the computational literature of scene understanding or, more broadly, the field of computer vision.

In the field of social cognition, researchers have identified two unique components that distinguish human adults from infants and other primates [43]. The first component is "**representational Theory of Mind (ToM)**," the ability to attribute mental states to oneself and others, to understand that others have perspectives and mental states different from one's own, as well as using these abilities to recognize false belief [39]. In the theoretical construct of mental states, mainstream psychology and related disciplines have traditionally treated *belief* as one simplest form, and therefore one of the building blocks of conscious thought [23]. Belief can be constructed as mental objects with semantic attributes; cognitive states and processes are constituted by the occurrence, transformation, and storage of such information-bearing structure [38]. The second component is the **triadic relations**: *You*, and *Me*, collaboratively looking at, working on, or talking about *This* [47]. Much power of human social cognition depends on the ability to form representations with a triadic structure [43].

To promote social cognition in computer vision, we focus on belief dynamics in nonverbal communication. Here, *belief* is defined as an entity and its attributes (*e.g.*, location), and *belief dynamics* (*i.e.*, the change of belief) are naturally and completely summarized using four categories: *occur* indicates an agent becomes aware of an object at a certain location, *update* means an agent knows the object's attribute was updated, *disappear* denotes that an agent loses track of the object's attribute, and *null* is no change. We emphasize on *triadic relations* emerged during nonverbal communication, including *No Communication*, *Attention Following*, and *Joint Attention* [12, 1]: *No Communication* indicates no social interaction between the two agents, *Attention Following* is a one-way observation, and *Joint Attention* means that two agents have the same intention to share attention on a common stimulus and both know that they are sharing the attention [47]; see an illustration in Fig. 1.

*Lifeng Fan and Shuwen Qiu contributed equally.

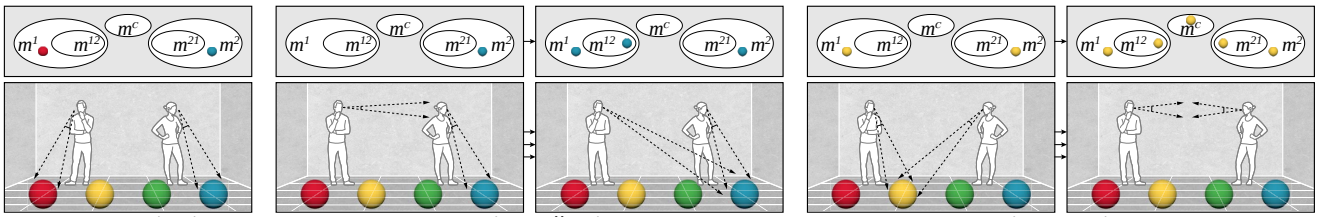

Figure 1: **Triadic belief dynamics in nonverbal communication.** Three types of communication events emerge from social interactions (bottom) and causally construct agents' belief dynamics (top). In this paper, we propose a novel structural mind representation "five minds" and a learning and inference algorithm for belief dynamics based on a hierarchical energy-based model that tracks (i) each agent's mental state (m^1 and m^2), (ii) their estimated belief about other agent's mental state (m^{12} and m^{21}), and (iii) the common mind (m^c). Of note, some events have two phases connected by three arrows.

To account for the two social components computationally, we propose a novel structural mind representation, termed "**five minds**," that includes two first-order self mental states (*i.e.*, the ground-truth mental state), two second-order estimated mental states of each other's mind (may deviate from the ground-truth mental states), and the third-level "common mind." Note that the proposed "five minds" differs from prior models that attempt to infer mental states among agents *recursively* with potentially infinite loops; instead, the "common mind" considers what the two agents share completely transparently without infinite recursion and corresponds to the concept of "common ground" [47].

The proposed "five minds" model is well-grounded to visual inputs, especially in terms of nonverbal communication. For instance, gaze communication uses eye gazes as portals inward to provide agents with glimpses into the inner mental world [12], and pointing gesture serves as "the first uniquely human forms of communication" to ground and reshape mental states [47]. We bring these crucial social components into representing, modeling, learning, and inference of belief dynamics in the computer vision community. Intuitively, the spatiotemporal parsing of social interactions affords the emergence of communication events; these events causally affect belief dynamics. Thus, a hierarchical energy-based model with Bayesian inference is naturally derived to track, maintain, and predict the mental states of all "five minds." To demonstrate the model's efficacy, we collect a new 3D video dataset with eye-tracking devices to facilitate ground-truth labeling. We verify the proposed method on this new 3D video dataset focusing on rich nonverbal social interactions and triadic belief dynamics.

This paper makes four contributions: (i) By incorporating crucial social cognition components, we address a new task of triadic belief dynamics learning and inference from nonverbal communication in natural scenes with rich social interactions. We propose a novel structural mental representation "five minds" by introducing a "common mind," with well-defined and quantized belief and belief dynamics, as well as nonverbal communication events. To the best of our knowledge, ours is the first to tackle such challenging problems in the field of computer vision. (ii) We collect a new 3D video dataset with rich social interactions using eye-tracking devices to facilitate ground-truth labeling; nonverbal communication events and belief dynamics are densely annotated. Such a setup goes beyond toy and symbolic examples presented in the literature, which we believe will serve as a modern benchmark for high-level social learning based on pixel inputs. (iii) We devise a hierarchical energy-based model and a beam-search-based algorithm to simultaneously optimize the learning and inference of nonverbal communication events and belief dynamics. (iv) We provide a benchmark and demonstrate the efficacy of the proposed method in a keyframe-based video summary.

2. Related Work

Nonverbal behavior and human communication Tomasello [47] argues that nonverbal communication is the "unconventionalized and uncoded" form, more foundational than the human natural language. Crucially, instead of merely treating head and body motions as an assembly of skeletons movements (*e.g.*, gaze [27], gesture [35], or interaction [26] in computer vision), we do recognize the underlying intentions behind these motions from the perspective of human social cognition; pointing and iconic gestures have their special meaning to convey the message and establish shared intentionality and common ground [13].

This unique view of nonverbal behavior and communication is largely ignored in modern scene understanding and computer vision. The present work subsumes prior work in gaze, gesture, body motions, and interactions in computer vision by presenting a hierarchical graphical representation, wherein the communication events [12, 11] emerge from the spatiotemporal parsing of low-level signals to maintain the triadic relations and belief dynamics among agents.

Machine ToM ToM has been long regarded as an acid test for human social interaction; impairment of such capability to construe persons in terms of their inner mental lives often results in autism [4]. In literature, modern computational models of ToM often treat the inference of mental states as *infinite* (or approximated by finite) recursions, notably by partially observable Markov decision process (POMDP) [17, 2, 9, 8]. Recent research includes esti-

mating the opponent's sophistication (*i.e.*, recursion) in sequential games [52, 7], representing and updating beliefs through time [56, 37], reasoning about other agent's desires and beliefs based on their actions with a Bayesian account [3], or learning to model agent's mental state and policy in grid world [41]. However, studies have concluded that the default level of recursive reasoning typically could go no deeper than one or two levels [5]; instead, we tend to build and rely on the "common mind" [46, 45] after only one or two levels of recursive reasoning of mental states.

In this paper, we adopt the representation of "common mind," a crucial ingredient to properly interpret the triadic relation *without* infinite recursion. Sharing a similar spirit, the coordinated and joint planning [14, 15, 29, 21] has been extensively studied in symbolic-like environments. Additional efforts have also emerged to recognize false-belief or perspective-taking with more realistic and noisy data in the field of robotics [55, 34], computer vision [10], and natural language processing [36]. However, the problem settings in prior work either lack rich interactions or communications among agents or have relatively confined problem space (at most on object/human tracking). In comparison, the problem setting in this paper considers rich social interactions with nonverbal communication in physical indoor environments captured and synced with multiple Azure Kinect sensors and eye-tracking devices. To tackle the challenges introduced by raw video input, we present a much more expressive hierarchical representation to interpret the interactions and communications among agents.

Keyframe-based video summary Keyframe-based video summary is a practical application of video understanding. In literature, models tend to obtain keyshots for segment-based summary [16], minimize the reconstruction loss [33], or directly compute a frame-level score, measuring the frame's contribution in summarizing the video essence [44]. Various mechanisms and additional cues have been adopted to improve semantics, including temporal dependency [57], subtitles [54] and action features [30]. Although these models are effective in general, they primarily rely on low-level features (*e.g.*, appearance, motion) without much modeling of high-level "agency" of human agents.

Obtaining a better semantic summary for videos with rich human interactions and nonverbal communications necessitates the modeling and understanding of the agents' mental world. To tackle this problem, we incorporate belief dynamics and model nonverbal communications in the video summary task for interaction-rich videos.

3. Representation and Model

In this section, we start by introducing the proposed ToM representation, "five minds," that accounts for the triadic relation and "common mind"; this representation is embedded in a hierarchical graphical model with a six-level structure. Next, to learn a probabilistic distribution over such hierarchically structured data and capture the relations among latent and observable variables, a classic Gibbs energy-based probabilistic formulation with carefully designed and most representative energy terms is derived, capable of parsing the communication events that emerged from the raw pixel inputs and tracking belief dynamics in five minds. At length, we conclude this section with a detailed description of learning and joint inference algorithms.[1]

3.1. Hierarchical Representation

Given the input image sequence $I = \{I_t\}_{t=1,\dots,T}$, the detected human agent i at time t is denoted by $h_t^i = (x_t^i, p_t^i, g_t^i)$, where $x_t^i \in \mathbb{R}^3$ denotes the spatial position, $p_t^i \in \mathbb{R}^{3\times 26}$ the skeleton pose, and $g_t^i \in \mathbb{R}^3$ the gaze direction. Similarly, $o_t^j = (x_t^j, c_t^j, d_t^j)$ denotes the detected object j at time t, where $x_t^j \in \mathbb{R}^3$ denotes the spatial location, $c_t^j \in \mathbb{C}$ the object category, and $d_t^j \in \{1, \dots, N_o\}$ the object ID; $\mathbb{C}$ is the object category set. Let $H = \{h_t^i\}$ and $O = \{o_t^j\}$ denote all the detected human agents and objects in the video. Without loss of generality, we assume a minimal setting for triadic relation with two agents in a video.

Formally, all minds M_t at time t is represented as a set, forming a "five minds" representation:

$$M_t = \{m_t^1, m_t^2, m_t^{12}, m_t^{21}, m_t^c\}, \qquad t = 1, \dots, T, \quad (1)$$

where m_t^1 and m_t^2 denote two agents' mind, m_t^{12} and m_t^{21} denote the agent's belief about the other agent's mind, and m_t^c denotes their common mind. Each mind is defined as $m_t = \{(o_t^i, A(o_t^i)) : i = 1, \dots, N_{o,t}\}$ with a set of objects o^i and their attributes $A(o^i)$ (*e.g.*, 3D location). The state change of M_t, *i.e.*, $\Delta M_t = \{\Delta m_t^1, \Delta m_t^2, \Delta m_t^{12}, \Delta m_t^{21}, \Delta m_t^c\}$, defines the belief dynamics. Here, $\Delta m = \{\Delta(o_t^i, A(o_t^i))\}$ and belief dynamics in each mind $\Delta(o_t^i, A(o_t^i)) \in \{0, 1, 2, 3\}$, correspond to four communication types, *occur*, *disappear*, *update*, and *null*.

ΔM_t along time construct the overall *belief dynamics* $\{\Delta \boldsymbol{\mathcal{M}}\}$, derived from the spatiotemporal parsing of the video. The parsing is represented by a spatiotemporal parse graph [59] $pg = (pt, E)$, a hierarchical graphical model that combines a parse tree pt and the contextual relation E on terminal nodes; Fig. 2 illustrates an example. A parse tree $pt = (V, R)$ includes the vertex set with a six-level hierarchical structure $V = V_r \cup V_b \cup V_e \cup V_s \cup V_f \cup V_t$ and the decomposing rule R, where V_r is the root set with only one node representing the entire video, V_b the set of belief dynamics of "five minds," V_e is the set of communication events, V_s is the set of interactive segments, V_f is the set of frame-based static scenes, and V_t is the set of all the detected instances in an indoor scene. Specifically:

[1] Henceforth, we use the term "mind" in human and animal studies and the term "mental state" in computational models interchangeably.

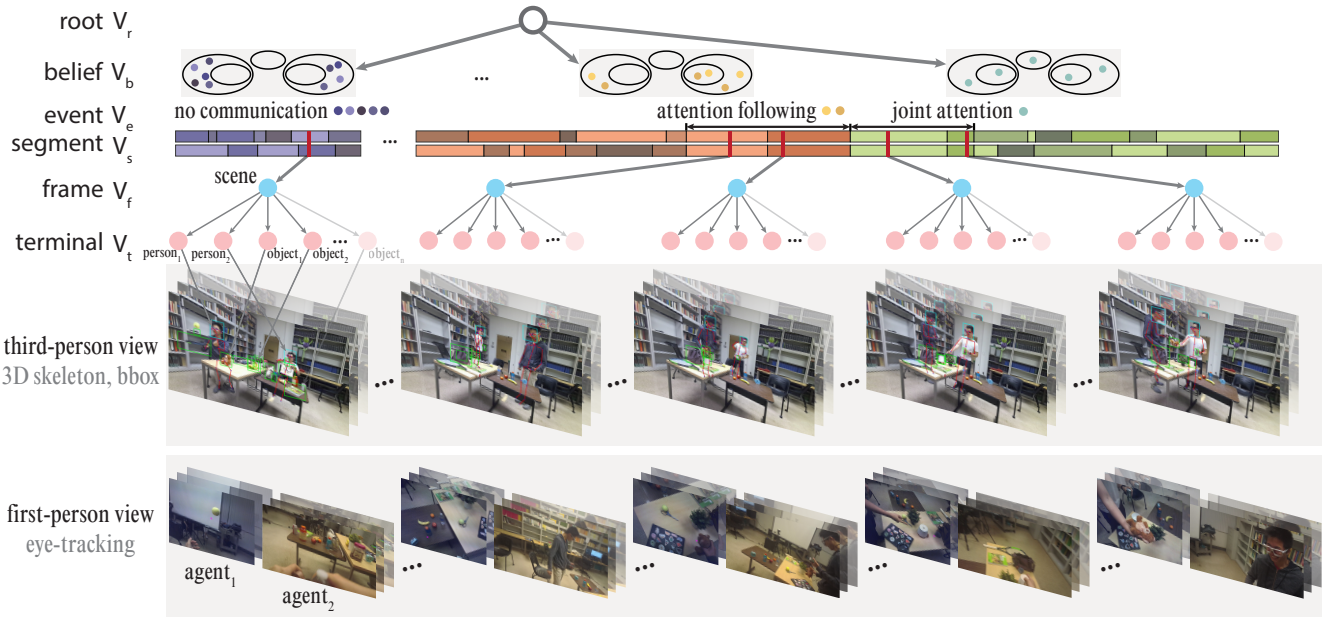

Figure 2: **A parse graph of a social event with a six-level hierarchical structure.** V denotes vertex sets in the hierarchy. The root node V_r corresponds to the entire video. The set of belief dynamics V_b emerges from the lower-level communication events (see also Fig. 1). Communication events in V_e decompose into lower-level interactive segments in V_s; these segments are social primitives learned unsupervisedly. Each frame of the scene in V_f further decomposes into several terminal nodes in V_t, grounded into entities detected from videos. The colored dots in the V_e layer represent belief changes triggered by communication events. Note that belief dynamics are accumulated over time; we only illustrate the most significant changes.

- The belief dynamics $\Delta\mathcal{M}$ are conditioned on communication events V_e, grouped by interactive segments V_s.
- A communication event $e \in V_e$ is one of the three categorical nonverbal communication events: *No Communication*, *Attention Following*, or *Joint Attention*; see Fig. 1.
- An interactive segment $s \in V_s$ is the decomposition of a communication event $e \in V_e$ and represented by the 4D spatiotemporal features $\Phi_s = (\Phi_s^1, \Phi_s^2)$ extracted from detected entities. These features describe social interactions, including both unary Φ_s^1 and pair-wise features Φ_s^2.
- The contextual relation E is represented by an attention graph $\mathcal{G}_s$ formed based on 4D features, wherein the node represents an agent or an object in the scene, and an edge is connected between two nodes if there is directed attention detected among the two entities from visual inputs.

3.2. Probabilistic Formulation

To infer the optimal parse graph pg^* from raw video sequence I, we formulate the video parsing of social events as a maximum a posteriori (MAP) inference problem:

$$\begin{aligned} pg^* &= \arg\max_{pg} P(pg|H,O)P(H,O|I) \\ &= \arg\max_{pg} P(H,O|pg)P(pg)P(H,O|I), \end{aligned} \tag{2}$$

where $P(H,O|I)$ is the detection score of agents and objects in the video, $P(pg)$ is the prior model, and $P(H,O|pg)$ is the likelihood model. Below, we detail the prior model and the likelihood model one by one.

Prior The prior model $P(pg)$ measures the validness of parse graph; all the nodes in the parse graph should be reasonably parsed from the root node. We model the prior probability of pg as a Gibbs distribution: $P(pg) = \frac{1}{Z_1}\exp\{-\mathcal{E}(pg)\} = \frac{1}{Z_1}\exp\{-\mathcal{E}_{aggr} - \mathcal{E}_{evt} - \mathcal{E}_{be}\}$, where $\mathcal{E}_{aggr}$ is the aggregation prior, $\mathcal{E}_{evt}$ the communication event prior, and $\mathcal{E}_{be}$ the belief dynamics prior. Specifically,

- The aggregation prior is defined as $\mathcal{E}_{aggr} = \lambda_1 \frac{N_e}{T}$ to encourage the algorithm to focus more on high-level communication patterns, instead of being trapped into trivial primitives that results in fragmented segmentation.
- The communication event prior leverages transition and co-occurrence frequencies of communication events,

$$\begin{aligned} \mathcal{E}_{evt} = &- \frac{\lambda_2 \sum_{i,j,\mathbb{1}^{trans}(e_i,e_j)=1} \log p^{trans}(e_i,e_j)}{\sum_{i,j}(\mathbb{1}^{trans}(e_i,e_j)=1)} \\ &- \frac{\lambda_3 \sum_{i,j,\mathbb{1}^{occ}(e_i,e_j)=1} \log p^{occ}(e_i,e_j)}{\sum_{i,j}(\mathbb{1}^{occ}(e_i,e_j)=1)}, \end{aligned} \tag{3}$$

  where $p^{trans}(e_i,e_j)$ and $p^{occ}(e_i,e_j)$ are based on frequencies from the dataset, and $\mathbb{1}^{trans}$ and $\mathbb{1}^{occ}$ are indicator functions that reflects the spatiotemporal relations.
- $\mathcal{E}_{be}$ models the prior of belief dynamics, which helps to prune some invalid configurations, such as two consecutive *occurs* or an *occur* after an *update*. The prior model is defined as $\mathcal{E}_{be} = -\lambda_4 \sum_{j=1}^{N_e} \log p^M(\Delta\mathcal{M}_j|e_j)$, and

$$p^M(\Delta\mathcal{M}_j|e_j) = \prod_t p(\Delta M_{t+1}|\Delta M_t, e_j)\, p(\Delta M_t|e_j), \tag{4}$$

  where $\Delta\mathcal{M}_j$ is the set of belief dynamics in event e_j.

Likelihood The likelihood model measures the consistency between the parse graph pg and the ground-truth observed data H and O. Since our model has a hierarchical structure, we split the likelihood into three energy terms, corresponding to the three crucial layers in the parse graph:

$$\begin{aligned} &P(H,O|pg) = P(H,O|V_b,V_e,E) \\ &= \frac{1}{Z_2}\exp\{-\mathcal{E}^{comp}(H,O|V_e,E) \\ &- \mathcal{E}^{evt}(H,O|V_e,E) - \mathcal{E}^{be}(H,O,V_e|\{\Delta\boldsymbol{\mathcal{M}}\})\}. \end{aligned} \tag{5}$$

- The first energy term $\mathcal{E}^{comp}$ constrains the communication event composed by the interactive segments, so that the features within one composition are sufficiently similar, whereas the features between two consecutive compositions are considerably distinct:

$$\begin{aligned} &\mathcal{E}^{comp}(H,O|V_e,E) = \mathcal{E}(\boldsymbol{\Phi}|V_s,E) \\ &= \frac{\lambda_5}{N_e}\sum_{j=1}^{N_e}\left(\frac{1}{T_j}\sum_t \mathcal{D}(\phi_{j,t},\phi_{j,t+1})\right) \\ &- \frac{\lambda_6\sum_{i,j,\mathbb{1}^{trans}(e_i,e_j)=1}\mathcal{D}(\psi(\Phi_i),\psi(\Phi_j))}{\sum_{i,j}(\mathbb{1}^{trans}(e_i,e_j)=1)} \\ &- \frac{\lambda_7\sum_{i,j,\mathbb{1}^{occ}(e_i,e_j)=1}\mathcal{D}(\psi(\Phi_i),\psi(\Phi_j))}{\sum_{i,j}(\mathbb{1}^{occ}(e_i,e_j)=1)} \end{aligned} \tag{6}$$

  where $\Phi_i = \{\phi_{i,t}\}$ is the set of features within the interactive segment s_i, $\psi(\cdot)$ the wavelet transform [40], and $\mathcal{D}(\cdot)$ the Euclidean distance between the two sets of features.
- The second energy term $\mathcal{E}^{evt}$ is the negative communication event classification score with respect to the detected feature set $\boldsymbol{\Phi} = \{\Phi_j\}$ and the constructed attention graph set $\boldsymbol{\mathcal{G}} = \{\mathcal{G}_j\}$. This second term is defined as $\mathcal{E}^{evt}(H,O|V_e,E) = \mathcal{E}(\boldsymbol{\Phi},\boldsymbol{\mathcal{G}}|V_e)$ and encodes all the entities in the scene extracted from visual input, which can be solved by a maximum likelihood estimation (MLE):

$$\begin{aligned} \mathcal{E}(\boldsymbol{\Phi},\boldsymbol{\mathcal{G}}|V_e) &= -\frac{1}{N_e}\sum_{j=1}^{N_e}\lambda_8\log p(\Phi_{\Lambda_j},\mathcal{G}_{\Lambda_j}|e_j) \\ &= -\frac{1}{N_e}\sum_{j=1}^{N_e}\lambda_8\log p(e_j|\Phi_{\Lambda_j},\mathcal{G}_{\Lambda_j}) - C, \end{aligned} \tag{7}$$

  where Λ_j is the set of indexes of the interactive segments decomposed from e_j, and C is a constant.
- The third energy term $\mathcal{E}^{be}$ models the belief dynamics in all five minds, formally defined as

$$\begin{aligned} &\mathcal{E}^{be}(H,O,V_e|\{\Delta\boldsymbol{\mathcal{M}}_j\}) = -\frac{1}{N_e}\sum_{j=1}^{N_e}\lambda_9\log p(\Delta\boldsymbol{\mathcal{M}}_j|H,O,V_e) \\ &= -\frac{1}{N_e}\sum_{j=1}^{N_e}(\frac{1}{T_j}\sum_t\lambda_9\log p(\Delta M_{j,t+1}|g_{j,t+1},e_j,\{\Delta M_{j,t'}\})), \end{aligned} \tag{8}$$

  where $t' \in [t_j^s, t]$, t_j^s is the starting frame of the event e_j, and $g_{j,t+1}$ the attention graph of frame $t+1$ in event e_j.

3.3. Learning

The detailed learning process follows a bottom-up procedure. Specifically, the algorithm (i) parses each frame to extract the entities and relations, (ii) jointly and dynamically parses both interactive segments (proposals generated unsupervisedly by clustering methods) and communication events (with trained likelihood) by beam search, (iii) predicts the belief dynamics (with trained likelihood), and (iv) fine-tunes all the parameters to minimize the overall loss. Algorithm 1 details the overall learning procedure.

Algorithm 1: Learning to parse social events

Input : Video $\{I_{train}\}$, ground truth V_e^* and V_b^*.
Output: Parameter sets Θ_1^* and Θ_2^*, and parse graph pg.
Init. : $H, O, \boldsymbol{\Phi}, \boldsymbol{\mathcal{G}}, \Theta_1^*, \Theta_2^* = \varnothing$; $L_1^*, L_2^* = +\infty$

for I_i *in* $\{I_{train}\}$ **do**
  H_i = humanDetectionWithReID(I_i), $H \leftarrow H \cup H_i$
  O_i = objectDetectionWithReID(I_i), $O \leftarrow O \cup O_i$
  $\boldsymbol{\Phi}_i$ = extractSTFeatures(H_i, O_i), $\boldsymbol{\Phi} \leftarrow \boldsymbol{\Phi} \cup \boldsymbol{\Phi}_i$
  $\boldsymbol{\mathcal{G}}_i$ = buildAttentionGraph(H_i, O_i, $\boldsymbol{\Phi}_i$), $\boldsymbol{\mathcal{G}} \leftarrow \boldsymbol{\mathcal{G}} \cup \boldsymbol{\mathcal{G}}_i$
end
$V_s \leftarrow$ Generate $\{s\}$ by unsupervised clustering.
/* Train likelihood of e_j as in [12] */
Train $p(e_j|\Phi_{\Lambda_j},\mathcal{G}_{\Lambda_j})$ in Eq. (7) with ground-truth V_e^*.
/* Finetune the parameter set Θ_1^*. */
for $\Theta_1^{(i)} = (\lambda_1,\lambda_2,\lambda_3,\lambda_5,\lambda_6,\lambda_7,\lambda_8) \in \Omega_{\Theta_1}$ **do**
  Compute $\mathcal{E}^{comp}$ based on Eq. (6), given $\boldsymbol{\Phi}$ and $\Theta_1^{(i)}$.
  Compute $\mathcal{E}^{evt}$ based on Eq. (7), given $\boldsymbol{\Phi}$, $\boldsymbol{\mathcal{G}}$, $\Theta_1^{(i)}$.
  Infer V_e by dynamic programming beam search; see details in Algorithm 2.
  Calculate error L_1 between V_e and V_e^*.
  if $L_1 < L_1^*$ **then** $L_1^* \leftarrow L_1$. $\Theta_1^* \leftarrow \Theta_1^{(i)}$.
end
/* Train belief dynamics likelihood */
Train $p\left(\Delta M_{j,t+1}|g_{j,t+1},e_j,\{\Delta M_{j,t'}\}\right)$ in Eq. (8) with V_b^*.
/* Finetune the parameter set Θ_2^*. */
for $\Theta_2^{(i)} = (\lambda_4,\lambda_9) \in \Omega_{\Theta_2}$ **do**
  for e_j *in* V_e **do**
    Compute the posterior probability of belief dynamics based on Eqs. (4) and (8).
    Predict the best $\hat{V}_b$ by MAP.
  end
  Calculate error L_2 between the best predicted belief dynamics $\hat{V}_b$ and the ground-truth V_b^*.
  if $L_2 < L_2^*$ **then** $L_2^* \leftarrow L_2$. $\Theta_2^* \leftarrow \Theta_2^{(i)}$.
end

4. Experiment

4.1. Dataset

To verify the efficacy of the proposed algorithm, we collected a new 3D video dataset with rich social interactions. This dataset was shot in both third-person view (with Azure Kinect sensors) and first-person view (with two pairs of glasses, capable of reading videos) to properly mimic human social interactions and the perception of the environment. One pair of glasses is an SMI eye tracker, capable

Algorithm 2: Event inference via DP beam search

Input : $\boldsymbol{\Phi}, \boldsymbol{\mathcal{G}}, V_s, p(e_j|\Phi_{\Lambda_j}, \mathcal{G}_{\Lambda_j})$.
Output : V_e
Initialization: $V_e = \varnothing, \mathcal{B} = \{V_e, p = 0\}, m, n$.

while *True* **do**
  $\mathcal{B}' = \varnothing$
  for $\{V_e, p\} \in \mathcal{B}$ **do**
    $\{e_i\} = Next(V_s, V_e, m)$
    if $\{e_i\} \neq \varnothing$ **then**
    for *each proposed* e_i **do**
      $p(V_e|\boldsymbol{\Phi}, \boldsymbol{\mathcal{G}}) = DP(V_e, p, e_i, \Phi, \mathcal{G})$
      $V_e = V_e \cup \{e_i\}$; $\mathcal{B}' = \mathcal{B}' \cup \{V_e, p\}$
    end
    else $\mathcal{B}' = \mathcal{B}' \cup \{V_e, p\}$
  end
  if $\mathcal{B}' == \mathcal{B}$ **then** return $V_e = Best(\mathcal{B}, 1)$
  else $\mathcal{D} = Best(\mathcal{B}', n)$; $\mathcal{B} = \mathcal{D}$
end

of accurately tracking human eye gazes while recording the video. Another pair is Pivothead glasses, simply recording the first-person view video. Both glasses have a similar look to regular glasses to ensure maximum naturalism during data collection. Crucially, such a setup also eases the challenging ground-truth annotation procedure; the annotators can better understand the belief dynamics with the precise reference of the agents' first-person view attention.

Our dataset is specially designed to cover typical nonverbal communication in rich social interactions. In total, we collected 88 videos (109,331 frames, 72.89 minutes of each sensor) recorded with 12 participants in 7 different scenarios. The participants were asked to perform three types of nonverbal communication naturally, as illustrated in Fig. 1. Among all the nonverbal behaviors in the dataset, we annotated two major types including eye gaze (almost all frames) and pointing (around 5.47% of all frames); note also that our scenarios involve both first-order and second-order false beliefs [36]. We did not provide scripts for performing detailed actions; instead, we only informed the participants about the task and the types of nonverbal communication they can use. This design follows the principles in recent large-scale video dataset collection [25] to ensure maximum realism.

We densely annotated the dataset, including human head bounding boxes, object bounding boxes, pointing, interactive segments, communication events, and the belief dynamics in all "five minds" for all the objects in all the scenes. These ground-truths were first annotated by seven volunteers using Vatic [48] after simple tutorials and later verified by at least one expert. The annotation process relied on synced third-person and first-person views. Of note, having perfect ground-truth annotation is impossible for any high-level semantic task (*e.g.*, belief dynamics) due to its intrinsic ambiguities, which have also been exhibited in other more traditional computer vision tasks (*e.g.*, activity recognition and event segmentation). Here, the goal of annotation is not to provide universally perfect labels at each frame; instead, we hope to use these annotations to provide a reasonable quantitative measurement of the model's performance. In total, 5,975 frames are labeled with pointing gestures. Among communication events, 48.56% is *No Communication*, 32.51% *Attention Following*, and 18.93% *Joint Attention*. 40 videos have first-order false beliefs, and 13 videos have second-order false beliefs. 26 videos are reserved as the testing set, and the rest 62 videos are used for model training and validation. Detailed statistics of belief dynamics in the dataset are tabulated in Table 1.

The pre-process procedure including following steps:

- Azure Kinect SDK tracks 26 3D joints of each agent.
- Detectron2 [51] detects objects.
- Deepsort [50] tracks objects.
- Object RGB and category features are matched and aligned between the third- and first-person views.
- Gaze360 [53] estimates 3D human gazes.
- We use the detected objects, depth maps, and camera parameters to recover 3D object point clouds.

Combining with 3D information and multi-views, features that can be potentially extracted from our dataset would be advantageous compared to 2D data, especially in cases that require handling complex occlusions, which is crucial for multi-agent human-object and human-human interactions.

Although collecting and processing a new 3D video dataset is challenging, it is the only viable direction to go if we hope to study social cognition on natural videos in indoor environments. First, as such a study requires dense annotations for evaluations, it demands specific hardware and computational power (*e.g.*, eye-tracking glasses, Azure Kinect sensors) to collect the raw data in a way that could ease the annotation process. Hence, crowd-sourcing the dataset is not an option. Second, the ideal dataset would cover rich social interactions with nonverbal communication. The closest existing dataset is presented in Fan *et al.* [12] using clips collected from TV shows and movies; however, its nonverbal communication and belief dynamics are sparse. As a significant upgrade to existing datasets, we hope this new dataset paves the way towards ToM modeling in natural and complex indoor environments.

4.2. Task 1: Predicting Belief Dynamics

To directly evaluate the proposed algorithm, we predict belief dynamics in all five minds on our dataset and evaluate based on the Macro-average of Precision and F1-score.

To make a fair comparison, we consider the following five **baselines**: (i) **Chance** is a weak baseline, *i.e.*, randomly assigning a belief dynamic label; (ii) **CNN** uses the pre-trained ResNet-50 [18] to extract image features and adopts an MLP to classify the belief dynamics; (iii) **CNN w/ HOG-LSTM** feeds both the ResNet-50 features of the entire image and the HOG [6] of the local image patch gazed by the agent to an LSTM [22], followed by an MLP to predict belief dynamics; (iv) **CNN w/ HOG & memory** adds the history of predicted belief dynamics on top of ResNet-50

Table 1: **Statistics of belief dynamics in our dataset.** The numbers of belief dynamics denote different categories: 0–*occur*, 1–*disappear*, 2–*update*, 3–*null*. The belief dynamics are imbalanced by its inherent sparse nature, with *null* most frequent and *occur*/*disappear* rare; it is one of the many challenges that make the inference of belief dynamics difficult.

Five Minds	m^1				m^2				m^{12}				m^{21}				m^c			
Belief Dynamics	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
No Communication	49	0	501	78017	36	2	442	78087	1	1	0	78514	3	1	0	78528	0	0	0	78545
Attention Following	36	6	401	36953	33	2	457	36899	23	6	128	37238	23	5	264	37104	0	0	0	37370
Joint Attention	15	5	324	26136	17	0	340	26119	32	6	290	26153	28	1	267	26180	32	6	166	26276

Table 2: **Quantitative results on predicting belief dynamics of five minds.** The best scores are marked in **bold**.

	Macro-average of Precision (↑)						Macro-average of F1-score (↑)					
Model	m^1	m^2	m^{12}	m^{21}	m^c	Avg.	m^1	m^2	m^{12}	m^{21}	m^c	Avg.
Chance	0.250	0.250	0.250	0.250	0.250	0.250	0.103	0.104	0.102	0.101	0.100	0.102
CNN	0.250	0.250	0.250	0.250	0.250	0.250	0.171	0.167	0.169	0.174	0.250	0.186
CNN w/ HOG-LSTM	0.250	0.250	0.250	0.250	0.250	0.250	0.167	0.132	0.205	0.182	0.250	0.187
CNN w/ HOG & memory	0.277	0.279	0.266	0.267	0.259	0.270	0.285	0.285	0.246	0.250	0.155	0.244
Features w/ memory	0.272	0.278	0.253	0.260	0.256	0.264	0.274	0.288	0.230	0.227	0.191	0.242
Init. seg.	0.371	**0.418**	0.293	0.301	0.265	0.330	0.346	0.366	0.302	0.314	0.274	0.320
Event prior	0.384	0.409	0.294	0.307	0.264	0.332	0.365	0.364	0.305	0.324	0.273	0.326
Uniform event	0.385	0.413	0.293	0.310	0.267	0.334	0.363	0.366	0.304	0.328	0.278	0.328
Ours-full	**0.397**	0.415	**0.316**	**0.315**	**0.278**	**0.344**	**0.431**	**0.443**	**0.351**	**0.349**	**0.299**	**0.375**

and HOG features; (v) **Features w/ memory** uses the same sets of features as our methods (see details below) with the history of predicted belief dynamics; all features are concatenated and fed into an MLP to predict belief dynamics. We only used the annotations of belief dynamics in all five minds when training the deep learning models.

To assess the contributions and efficacy of essential components in the proposed method, we derive the following variants as **ablation study**: (i) **Init. seg.** directly uses the initial interactive segments generated by unsupervised clustering as event segments, without additional temporal clustering by beam search; (ii) **Event prior** only uses event prior for event assignment without the event likelihood; (iii) **Uniform event** replaces event posterior with uniform distribution, and randomly assign event labels for all segments.

A suite of 4D spatiotemporal features Φ_s (see Section 3.1) are adopted to ground our methods to raw video inputs. Unary feature Φ_s^1 concerns a single agent and concatenates features of human poses, hand-object distances, and estimated gaze and attention. Pair-wise feature Φ_s^2 focuses on the relations between two agents and concatenates features of the relative human poses between two agents, relative gaze angles, and relative hand joint distances. All models are implemented in PyTorch using ADAM optimizer [28] and trained on an Nvidia TITAN RTX GPU.

Quantitatively, our full model achieves the best performance on predicting belief dynamics of five minds, measured by the macro-average of both Precision and F1-score; see comparisons in Table 2. Overall, the CNN baseline and its variant with HOG-LSTM perform the worst on this challenging task, only slightly better than randomly guessing. The performances of *CNN w/ HOG & memory* and *Features w/ memory* are improved after incorporating the history of the estimated belief dynamics. The results indicate that the performance bottleneck to infer belief dynamics does not lie in the low-level features or representations; instead, it heavily depends on high-level semantics to distinguish similar segments and events to predict mental states and belief dynamics in nonverbal communication correctly.

The ablation study further reveals the effects of various model components. Without temporal segment re-clustering by dynamic-programming-based beam search, *Init. seg.* performs worse. Compared to using the posterior event distribution in our full model, the performance would drop if using either the biased event distribution prior or the uniform event distribution. Our full model yields the best performance on this challenging task with carefully derived hierarchical representation and joint learning algorithms.

4.3. Task 2: Keyframe-based Video Summary

We apply the proposed method on keyframe-based video summary on videos with rich social interactions. For comparison, we choose three state-of-the-art methods as baselines: DPP-LSTM [57], FCSN [42], and DSN [58]. To adopt the proposed method for this task, we sum over the predicted probabilities of belief dynamics *occur* and *disappear*, which indicate a significant belief change in agents' minds; we use it as the score indicating the frame's contribution in summarizing the video content. To quantitatively compare our method with three baselines, we conduced a study with 33 human participants, who were neither experts on the task, nor were they knowledgeable about the video. After seeing the entire video, participants were presented with top keyframes from different methods (in a counterbalanced order to avoid bias) and asked to select the best keyframe-based summary that describes the observed video. The proposed method outperforms the state-of-the-art methods significantly; see the detailed comparisons in Fig. 4.

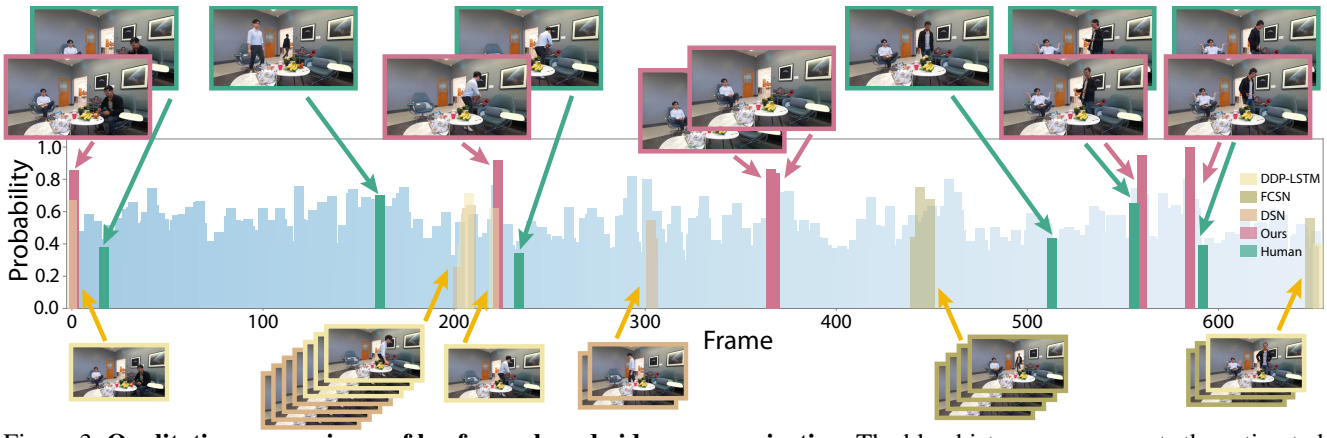

Figure 3: **Qualitative comparisons of keyframe-based video summarization.** The blue histogram represents the estimated probability of belief dynamics (including *occur* and *disappear*) by our model. The top keyframes chosen by human participants are shown next to our model's prediction. Overall, baseline models tend to predict frames merely based on visual patterns, making the top keyframes similar to each other and grouped in clusters. In comparison, with the proper modeling of belief dynamics, our method tends to capture the moment with significant belief changes in "five minds," resulting in a more uniform set of keyframes along the time. Note that human participants also demonstrate similar behaviors when choosing the top keyframes for video summarization. Please refer to the *supplementary material* for additional qualitative results.

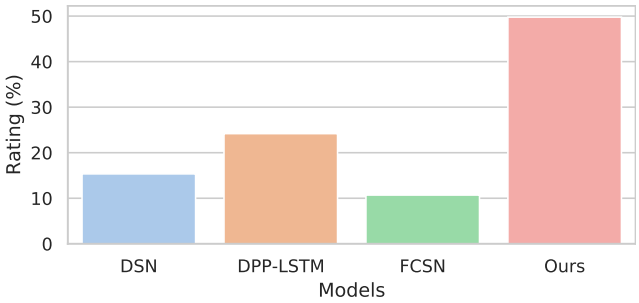

Figure 4: **Human ratings of keyframe-based video summary.** Our model outperforms state-of-the-art methods significantly on videos with rich interactions.

We further discuss a qualitative comparison shown in Fig. 3; please refer to the *supplementary material* for additional qualitative comparisons. In this example, agent A (in a black jacket) puts a teddy bear on the desk and leaves the room. Agent B (in a white shirt) later hides the teddy bear into a bag. When agent A comes back, he cannot find the teddy bear at the original location, showing his confusion to agent B by spreading his hands. Agent B pretends to have no clue about what happened and also spreads his hands. Agent A looks around for the lost teddy bear helplessly. As shown on the top of Fig. 3 (in green), keyframes chosen by human participants give a complete and refined summary of the video content. Our model produces a similar story digest—the most similar one compared with human judgments among all the methods. In essence, our method captures almost all the crucial moments of the story by modeling belief dynamics during social interactions. The keyframes generated by other baseline methods fail to capture the key moments spanning across the entire story; they tend to group the predicted keyframes in selected moments.

Taken together, the result presented here is no surprise. When watching and summarizing the video with rich social interactions, humans primarily understand the story from a higher level, typically considering the issues going on in the mental world instead of purely looking at the visual motions. As such, by introducing higher-level multi-agent belief dynamics into the keyframe modeling and selecting procedure, the generated keyframes can be mostly optimized to understand social interactions better.

5. Conclusion

This paper studies two critical components in understanding multi-agent social interaction in 3D real scenes, *i.e.*, understanding nonverbal communication and belief dynamics in "five minds," with a particular focus on a structured mental representation of "common mind." A six-level hierarchical graphical model is devised to account for the parsing of belief dynamics in "five minds," nonverbal communication events, the 4D spatiotemporal interactive segments, and the detected entities and relations from raw visual inputs. We propose an energy-based probabilistic model and a beam-search-based algorithm to learn and infer communication events and belief dynamics jointly. Experimental results show that our model captures the sparse belief dynamics in all five minds and facilitates generating more comprehensive keyframe-based video summarization. We believe such a unique social aspect of scene understanding could have broader applications in various future tasks.

Acknowledgements The authors thank Baoxiong Jia at UCLA and Siyuan Qi at Google Inc. for helpful discussions. Tao Gao was supported in part by DARPA PA 19-03-01 and ONR MURI N00014-16-1-2007. Other authors were supported in part by ONR MURI N00014-16-1-2007, ONR N00014-19-1-2153, and DARPA XAI N66001-17-2-4029.

References

[1] Henny Admoni and Brian Scassellati. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63, 2017. 1

[2] Chris Baker, Rebecca Saxe, and Joshua Tenenbaum. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2011. 2

[3] Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009. 3

[4] Simon Baron-Cohen, Ruth Campbell, Annette Karmiloff-Smith, Julia Grant, and Jane Walker. Are children with autism blind to the mentalistic significance of the eyes? *British Journal of Developmental Psychology*, 13(4):379–398, 1995. 2

[5] Colin F Camerer, Teck-Hua Ho, and Juin-Kuan Chong. A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3):861–898, 2004. 3

[6] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005. 6

[7] Harmen de Weerd, Denny Diepgrond, and Rineke Verbrugge. Estimating the use of higher-order theory of mind using computational agents. *The BE Journal of Theoretical Economics*, 18(2), 2017. 3

[8] Harmen De Weerd, Rineke Verbrugge, and Bart Verheij. How much does it help to know what she knows you know? an agent-based simulation study. *Artificial Intelligence*, 199:67–92, 2013. 2

[9] Prashant Doshi, Xia Qu, Adam Goodie, and Diana L Young. Modeling recursive reasoning by humans using empirically informed interactive pomdps. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2010. 2

[10] Benjamin Eysenbach, Carl Vondrick, and Antonio Torralba. Who is mistaken? *arXiv preprint arXiv:1612.01175*, 2016. 3

[11] Lifeng Fan, Yixin Chen, Ping Wei, Wenguan Wang, and Song-Chun Zhu. Inferring shared attention in social scene videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[12] Lifeng Fan, Wenguan Wang, Siyuan Huang, Xinyu Tang, and Song-Chun Zhu. Understanding human gaze communication by spatio-temporal graph reasoning. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 5, 6

[13] Margaret Gilbert. *On social facts*. Princeton University Press, 1992. 2

[14] Barbara Grosz and Sarit Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 1996. 3

[15] Barbara J Grosz and Luke Hunsberger. The dynamics of intention in collaborative activity. *Cognitive Systems Research*, 7(2-3):259–272, 2006. 3

[16] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2014. 3

[17] Yanlin Han and Piotr Gmytrasiewicz. Learning others' intentional models in multi-agent settings using interactive pomdps. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[19] Fritz Heider and Marianne Simmel. An experimental study of apparent behavior. *The American journal of psychology*, 57(2):243–259, 1944. 1

[20] Robert A Hinde. *Biological bases of human social behaviour*. McGraw-Hill, 1974. 1

[21] Mark K Ho, James MacGlashan, Amy Greenwald, Michael L Littman, Elizabeth Hilliard, Carl Trimbach, Stephen Brawner, Josh Tenenbaum, Max Kleiman-Weiner, and Joseph L Austerweil. Feature-based joint planning and norm learning in collaborative games. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2016. 3

[22] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 6

[23] Terence Horgan. Review essay: From cognitive science to folk psychology: Computation, mental representation, and belief. *Philosophy and Phenomenological Research*, 52(2):449–484, 1992. 1

[24] David Hume. *An Abstract of a Treatise of Human Nature, 1740*. CUP Archive, 1938. 1

[25] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-chun Zhu. Lemma: A multi-view dataset for learning multi-agent multi-task activities. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020. 6

[26] Hanbyul Joo, Tomas Simon, Mina Cikara, and Yaser Sheikh. Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[27] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2019. 2

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2014. 7

[29] Max Kleiman-Weiner, Mark K Ho, Joseph L Austerweil, Michael L Littman, and Joshua B Tenenbaum. Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2016. 3

[30] Robert Laganière, Raphael Bacco, Arnaud Hocevar, Patrick Lambert, Grégory Païs, and Bogdan E Ionescu. Video summarization from spatio-temporal features. In *TRECVid Video Summarization Workshop*, 2008. 3

[31] Hector J Levesque, Philip R Cohen, and José HT Nunes. On acting together. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 1990. 1

[32] Ludwig Ludwig Wittgenstein. *Philosophical investigations. Philosophische Untersuchungen.* Macmillan, 1953. 1

[33] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[34] Grégoire Milliez, Matthieu Warnier, Aurélie Clodic, and Rachid Alami. A framework for endowing an interactive robot with reasoning capabilities about perspective-taking and belief management. In *International Symposium on Robot and Human Interactive Communication*, 2014. 3

[35] Pradyumna Narayana, Ross Beveridge, and Bruce A Draper. Gesture recognition: Focus on the hands. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[36] Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Thomas L Griffiths. Evaluating theory of mind in question answering. *arXiv preprint arXiv:1808.09352*, 2018. 3, 6

[37] Alessandro Panella and Piotr Gmytrasiewicz. Interactive pomdps with finite-state models of other agents. *Autonomous Agents and Multi-Agent Systems*, 31(4):861–904, 2017. 3

[38] David Pitt. Mental representation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2020 edition, 2020. 1

[39] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978. 1

[40] Kevin Quennesson, Elias Ioup, and Charles L Isbell. Wavelet statistics for human motion classification. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2006. 5

[41] Neil C Rabinowitz, Frank Perbet, H Francis Song, Chiyuan Zhang, SM Eslami, and Matthew Botvinick. Machine theory of mind. *arXiv preprint arXiv:1802.07740*, 2018. 3

[42] Mrigank Rochan, Linwei Ye, and Yang Wang. Video summarization using fully convolutional sequence networks. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2018. 7

[43] Rebecca Saxe. Uniquely human social cognition. *Current opinion in neurobiology*, 16(2):235–239, 2006. 1

[44] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[45] Stephanie Stacy, Qingyi Zhao, Minglu Zhao, Max Kleiman-Weiner, and Tao Gao. Intuitive signaling through an "imagined we". In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2020. 3

[46] Ning Tang, Stephanie Stacy, Minglu Zhao, Gabriel Marquez, and Tao Gao. Bootstrapping an imagined we for cooperation. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2020. 3

[47] Michael Tomasello. *Origins of human communication*. MIT press, 2010. 1, 2

[48] Carl Vondrick, Donald Patterson, and Deva Ramanan. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision (IJCV)*, 101(1):184–204, 2013. 6

[49] Peter A White. Ideas about causation in philosophy and psychology. *Psychological bulletin*, 108(1):3, 1990. 1

[50] Nicolai Wojke and Alex Bewley. Deep cosine metric learning for person re-identification. In *Proceedings of Winter Conference on Applications of Computer Vision (WACV)*, 2018. 6

[51] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 6

[52] Michael Wunder, Michael Kaisers, John Robert Yaros, and Michael Littman. Using iterated reasoning to predict opponent strategies. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2011. 3

[53] Yanyu Xu, Yanbing Dong, Junru Wu, Zhengzhong Sun, Zhiru Shi, Jingyi Yu, and Shenghua Gao. Gaze prediction in dynamic 360 immersive videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[54] Haoran Yi, Deepu Rajan, and Liang-Tien Chia. Semantic video indexing and summarization using subtitles. In *Pacific-Rim Conference on Multimedia*, 2004. 3

[55] Tao Yuan, Hangxin Liu, Lifeng Fan, Zilong Zheng, Tao Gao, Yixin Zhu, and Song-Chun Zhu. Joint inference of states, robot knowledge, and human (false-) beliefs. In *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*, 2020. 3

[56] Luke Zettlemoyer, Brian Milch, and Leslie P Kaelbling. Multi-agent filtering with infinitely nested beliefs. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2009. 3

[57] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016. 3, 7

[58] Kaiyang Zhou, Yu Qiao, and Tao Xiang. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 7

[59] Song-Chun Zhu and David Mumford. A stochastic grammar of images. *Foundations and Trends® in Computer Graphics and Vision*, 2(4):259–362, 2007. 3

[60] Yixin Zhu, Tao Gao, Lifeng Fan, Siyuan Huang, Mark Edmonds, Hangxin Liu, Feng Gao, Chi Zhang, Siyuan Qi, Ying Nian Wu, et al. Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. *Engineering*, 6(3):310–345, 2020. 1