Supplementary Material for 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

1. Beam Search Algorithm

2. Dataset

Fig. 1 showcase some snapshots from our dataset. Every three rows correspond to one long video, wherein the first row is the third-person view, and the other two rows are the first-person views from two agents. The first video is mainly about *Joint Attention*. The second video includes *No Communication*, *Attention Following* and *Joint Attention*; it also involves second-order false belief. The third video includes *Attention Following*. The fourth video includes *No Communication*.

3. Surveys for Human Studies

Below are the links to the questionnaires for the human subject studies in the keyframe-based video summary task.

- Group 1: https://5minds.typeform.com/to/dh782Z
- Group 2: https://5minds.typeform.com/to/T3hGhN
- Group 3: https://5minds.typeform.com/to/wovakS
- Group 4: https://5mind.typeform.com/to/SpOMu3

4. Additional Quantitative Results

4.1. ROC curve

Fig. 3 show the ROC curves for all five minds in the predicting belief dynamics task. The numbers of belief dynamics denote different categories: 0–*occur*, 1–*disappear*, 2–*update*, and 3–*null*.

5. Additional Qualitative Results

Fig. 2 shows additional qualitative results for the keyframe-based video summary task.

*Lifeng Fan and Shuwen Qiu contributed equally.

Algorithm 1: Infer events via dynamic programming beam search

Input : Extracted feature set $\boldsymbol{\Phi}$, constructed attention graph $\boldsymbol{\mathcal{G}}$, the set of interactive segment proposals V_s, and pre-trained likelihood $p(e_j|\Phi_{\Lambda_j}, \mathcal{G}_{\Lambda_j})$.
Output : Communication events 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**
    /* Propose next m possible events (both the event segment and the event label). */
    $\{e_i\} = Next(V_s, V_e, m)$
    if $\{e_i\}$ *is not empty* **then**
      for *each proposed* e_i **do**
        /* Calculate the posterior probability of V_e via dynamic programming. */
        $p(V_e|\boldsymbol{\Phi}, \boldsymbol{\mathcal{G}}) = DP(V_e, p, e_i, \Phi, \mathcal{G})$
        $V_e = V_e \cup \{e_i\}$
        $\mathcal{B}' = \mathcal{B}' \cup \{V_e, p\}$
      end
    end
    else
      $\mathcal{B}' = \mathcal{B}' \cup \{V_e, p\}$
    end
  end
  if $\mathcal{B}' == \mathcal{B}$ **then**
    return $V_e = Best(\mathcal{B}, 1)$
  end
  else
    /* select n best event parsing with best posterior prob from all candidates. */
    $\mathcal{D} = Best(\mathcal{B}', n)$
    $\mathcal{B} = \mathcal{D}$
  end
end

Figure 1: Sample snapshots of the *Meditation* dataset.

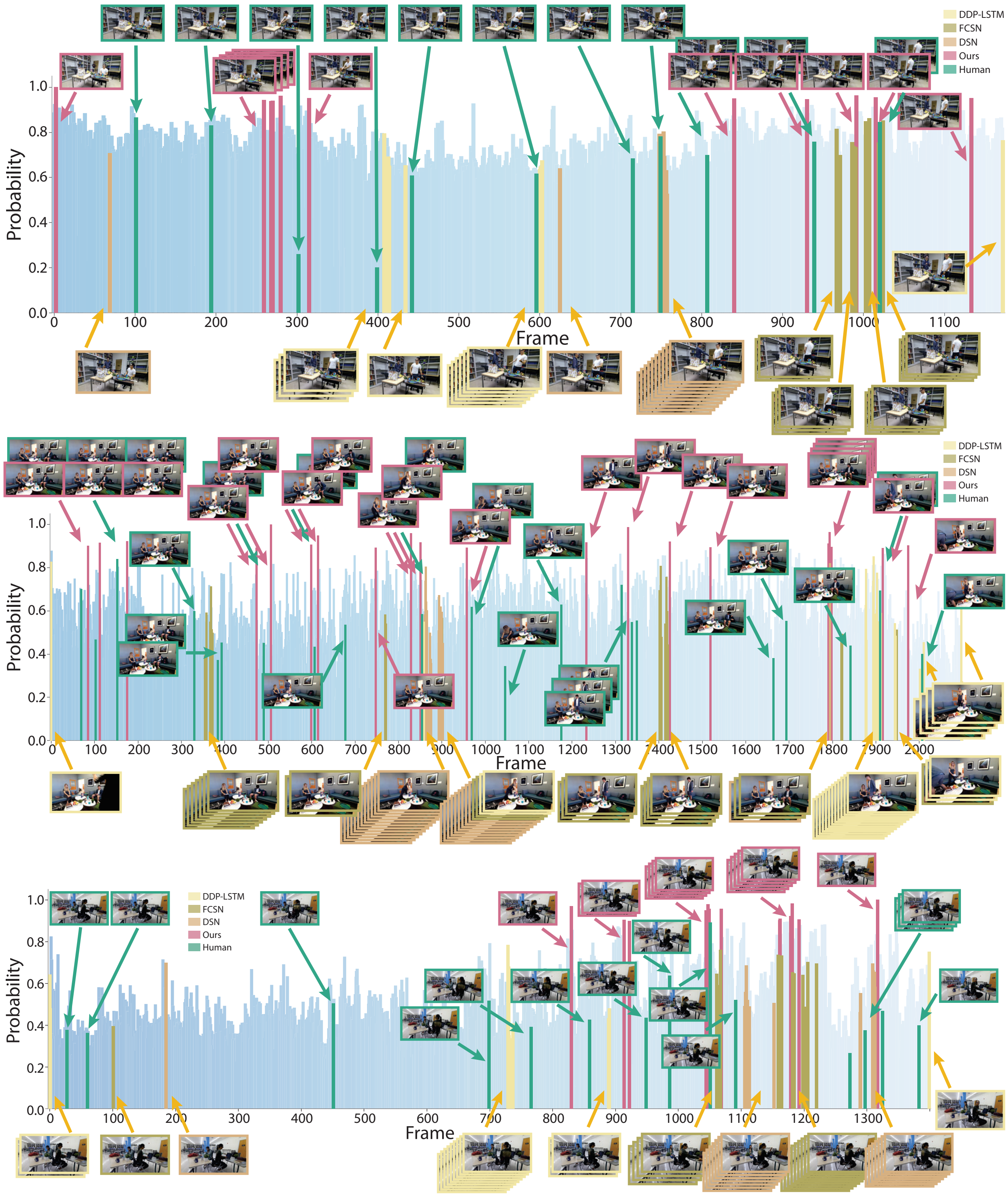

Figure 2: Additional comparisons on video summarization.

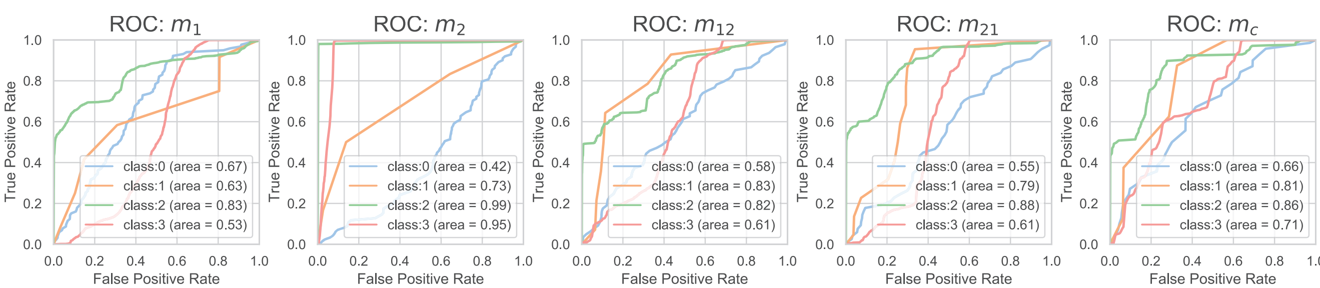

Figure 3: ROC Curve