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 $\Phi$, constructed attention graph $G$, the set of interactive segment proposals $V_s$, and pre-trained likelihood $p(e_j|\Phi, \lambda, G, \Lambda)$. |
| Output | : Communication events $V_e$ |
| Initialization | $V_0 = \emptyset, B = \{V_e, p = 0\}$ |

1. while True do
2. $B' = \emptyset$
3. for $(V_e, p) \in B$ do
4. /* Propose next $m$ possible events (both the event segment and the event label). */
5. $\{e_i\} = \text{Next}(V_e, V_e, m)$
6. if $\{e_i\}$ is not empty then
7. /* Calculate the posterior probability of $V_e$ via dynamic programming. */
8. $p(V_e|\Phi, G) = DP(V_e, p, e_i, \Phi, G)$
9. $V_e = V_e \cup \{e_i\}$
10. $B' = B' \cup \{V_e, p\}$
11. end
12. else
13. end
14. end
15. if $B' = B$ then
16. return $V_e = \text{Best}(B, 1)$
17. end
18. else
19. /* select n best event parsing with best posterior prob from all candidates. */
20. $D = \text{Best}(B', n)$
21. $B = D$
22. end
23. end
Figure 1: Sample snapshots of the Meditation dataset.
Figure 2: Additional comparisons on video summarization.
Figure 3: ROC Curve