Question: What did Dokyung do in this scene?	
Story elements: Character and Event in Target, Identity in Content, Recognition in Thinking	
AI agent	Answer: Dokyung was sitting on the ground. (✗)
Pre-operational stage	Answer: Dokyung is looking at Haeyoung. (✗)
Middle concrete stage	Answer: He is sitting down next to Haeyoung. (✓)
Concrete generalization stage	Answer: He was sat on the hospital bed. (✓)
Formal stage	Answer: Dokyung was sitting in the bench in the hospital within Haeyoung's sight. (✓)

Table 1: A question and answer example sampled in the case study of the Video Turing Test (VTT). The check-mark (✓) and x-mark (✗) represent correct and wrong answers, respectively.

- `Content`: Identity, Feature, Relationship, Means, Context, Sequence, Causality, Motivation
- `Thinking`: Recall, Recognize, Reasoning

The story elements can be assigned on the questions to specify required information, knowledge, or thinking ability by each question. Also, the correct answer rate for each story element is calculated by combining results from the answers. Through this analysis, we can identify the strengths and weaknesses of the video understanding intelligence of each player very specifically. The set of questions for the VTT can be organized so that CogME attributes appear evenly across the questions. This means that all story elements related to video understanding intelligence can be assessed with a set of questions.

The composition of human players

Human intelligence is not uniform, and there are significant differences in the cognitive aspect by age group according to their developmental stages and individual characteristics. Each person has a different pattern of strengths and weaknesses in their performances and related components of cognition. For example, children are generally good at recall but clumsy in reasoning, while some adults usually focus on the main character in a video clip and ignore surroundings. Therefore, judging the human-likeness of an AI agent can be influenced by the diverse characteristics of the interrogator or the human player. Ultimately, it becomes difficult to generalize the results of the test. To evaluate the human-likeness of an AI agent, VTT conducts tests with a group of human players rather than a single human player. Depending on the characteristics of the human player group, the test can derive more concrete and specified results. For example, if test organizers want to evaluate an AI agent based on age groups, human player groups can be organized based on the developmental stages of Piaget's theory (**?**).

Case Study

Here, we provide a case study of the VTT to confirm the feasibility and effectiveness of the proposed test. We explain detailed implementations of the case study in this section.

Video selection. We utilize shot and scene clips sampled from a Korean popular TV drama "Another Miss Oh".

Question selection. We carefully select 30 questions for the case study considering i) the distribution of story elements, ii) question type (i.e., 15 multiple-choice QA vs. 15 open-ended QA), and iii) associated video length for each question (i.e., 17 QA for shot clip and 13 QA for scene clip).

Human players. We organize a human player group that consists of four people from the different cognitive developmental stages of humans (**?**). The age of the four players is 4 years (pre-operational stage), 10 years (middle concrete stage), 14 years (concrete generalization stage) and 20+ years (formal stage), respectively.

AI agent. As an AI agent, we employ two video QA algorithms: (**?**) for multiple-choice QA and (**?**) for open-ended QA.

Juries. As juries for the test, we invite 16 people of ages from the 20s to the 50s and both sexes. Six people of juries are experts for AI and others are non-expert audiences.

Miscellaneous. We divide 30 questions into 5 rounds of 6 questions. All participants are randomized after each round. The juries are provided with a judgment sheet to record one's thoughts of each QA. After each round consisting of 6 QAs, the juries have to guess who is the AI, based on the judgment sheet as a cue.

Results

Question answering and juries' decision

First, we present examples of question answering conducted in the case study. In Table 1, given questions, story elements associated with the questions, and answers generated by players in different developmental stages are described. As shown in the table, our AI agent or few humans fails to answer a given question. At each round after conducting six question and answering pairs, the juries guess who is the AI among the players. As a result, juries perceive the AI around 62.5% and 75% at the first and second rounds, respectively. On the contrary, in the third round and fifth round, juries vote one of the human players as an AI around 62.5% and 87.5%, respectively. Interestingly, in the fourth round, the juries are quite confused and the voting results were a close match. The juries guessing the AI agent as an AI around 55.6% (correct guess), and guessing one of the human players as an AI around 45.4% (incorrect guess). In 4 rounds among a total of 5 rounds, one player receives more than half of the votes with a consistent result[1].

[1]The juries guess the AI agent as an AI around 48.6% on average for 5 rounds.

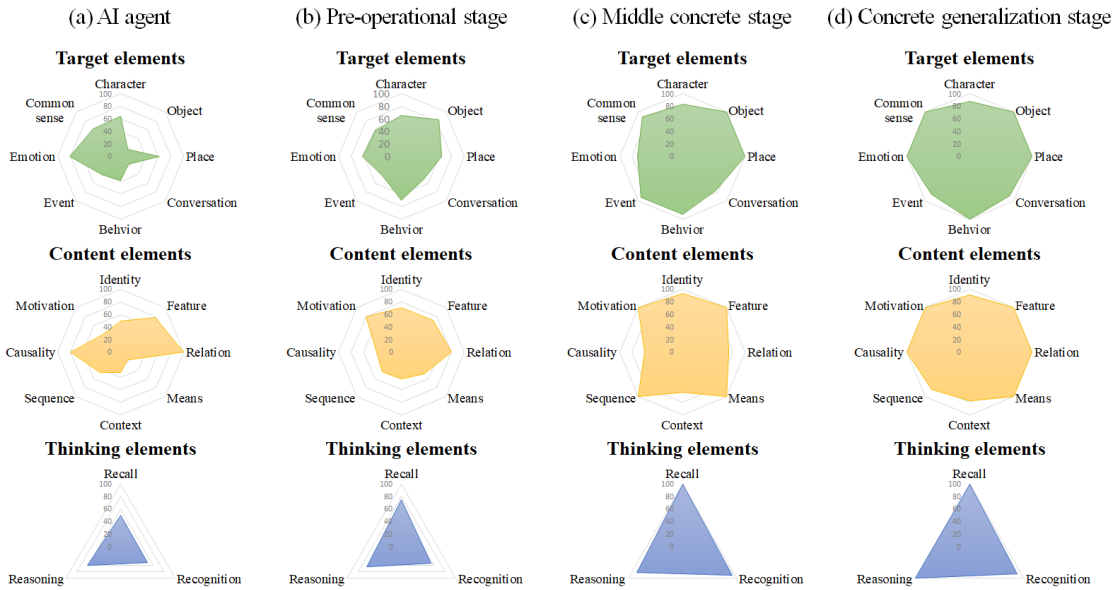

Figure 2: The accuracy-based profiles for each player participated in the case study of the Video Turing Test (VTT). To evaluate each player's video story intelligence, we apply CogME based on cognitive process of human.

Analysis on cognitive modules

We analyze each player's performance based on CogME for specifying their video understanding intelligence. We calculate the correct answer rate (accuracy) for each story element associated with the questions used in the case study. The accuracy-based profiles for each player on story elements in the three cognitive modules (i.e., target, content, thinking) are shown in Figure 2. Here, we highlight two key observations as follows. First, as the developmental stages of human progress, the accuracy of QA across all three modules is gradually improved. This means that some cognitive processes mature in a way that is reflected in the chosen metrics of this test. This result validates the use of these measures to evaluate video understanding intelligence. Second, when comparing the profiles between the AI agent and the human players, we notice that the performance of each component revealed more evenly in human players than the AI agent. Quantitatively, the standard deviation of the accuracy of each story element is 13.78 for a child in the pre-operational stage and 23.13 for the AI agent. In particular, the AI shows deficient performance in Object and Conversation in Target elements, Means and Motivation in Context elements, and Recall in Thinking elements. It is expected for two reasons that the AI agent's profile differs from the previous paper (**?**). This study considers two types of QA, multiple-choice and open-ended, while the previous work use only the former. Furthermore, we employ two distinct video QA algorithms as our AI agents to answer multiple-choice and open-ended QA.

Discussion and Conclusion

In this paper, we introduced the Video Turing Test (VTT), a novel measurement to evaluate a human-likeness of video understanding intelligence. We defined a general format and procedure of VTT and conducted a case study to confirm the feasibility and effectiveness of the proposed test. The case study provided new insight into the association of video understanding intelligence between the AI agent and human players from the different developmental stages. While the case study suggested a new perspective of measurement for video understanding intelligence, we still need to discuss several aspects. First, a 4-years child was included in the test as a human player for comparing the video understanding ability between the AI agent and a human from the pre-operational stage. However, several requirements (e.g., writing as a full sentence, choosing an answer among five answer candidates) were not familiar for the child so that it is hard to interpret the answers as totally reflecting the child's video understanding ability. Furthermore, we did not strictly specify the criteria to pass the VTT. The passing criteria can be defined by considering the detailed design of the VTT such as the composition of human players, juries, and the selected question set. As future work, we expect that the additional case study and its interpretation validates the clearness of VTT. For more objective and analytical evaluation for VTT, we plan to conduct various case studies and suggest appropriate guidelines including the composition of participants and the arrangement of question set.

Acknowledgements

This work was partly supported by the IITP (2015-0-00310-SW.StarLab/20%, 2017-0-01772-VTT/20%, 2018-0-00622-RMI/15%, 2019-0-01371-BabyMind/15%, 2021-0-02068-AIHub/15%) grants and the NRF of Korea (2021R1A2C1010970/15%) grant funded by the Korean government. Consequuntur dignissimos fuga quis expedita nostrum consectetur quam voluptas nemo rerum, corrupti deleniti repellat numquam ratione