Measuring Children's Mindreading Ability with Machine Comprehension

Yuliang Yan[1,2], **Xiaohua Wang**[1,2], **Xiang Zhou**[1,2], **Xiaoqing Zheng**[1,2,*], **Xuanjing Huang**[1,2]
[1]School of Computer Science, Fudan University, Shanghai, China
[2]Shanghai Key Laboratory of Intelligent Information Processing
{ylyan21,xiaohuawang22}@m.fudan.edu.cn
{zhengxq,xjhuang}@fudan.edu.cn

Abstract

Recently, much research in psychology has benefited from the advances in machine learning techniques. Some recent studies showed that it is possible to build automated scoring models for children's mindreading. These models were trained on a set of manually-labeled question-response pairs, which were collected by asking children to answer one or two questions after a short story is told or a video clip is played. However, existing models did not take the features of the stories and video clips into account when scoring, which obviously will reduce the accuracy of the scoring models. Furthermore, considering that different psychological tests may contain the same questions, this approach cannot be extended to other related psychological test datasets. In this study, we proposed a multi-modal learning framework to leverage the features extracted from the stories and videos related to the questions being asked during the children's mindreading evaluation. Experimental results show that the scores produced by the proposed models agree well with those graded by human experts, highlighting the potential of the proposed network architecture for practical automated children's mindreading scoring systems[1].

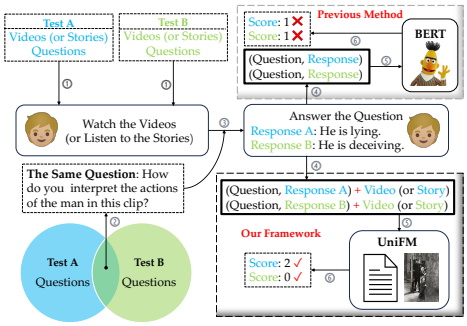

Figure 1: Suppose there are two different psychological tests that contain some identical questions. Previous methods only considered question-answer pairs as input to the model. Therefore, when faced with question-answer pairs as shown in the figure, the model failed to learn how to score these two pairs since the same question and the children's answers were close. However, our framework can handle such situations. By incorporating the background information from psychological tests into the input, the model can utilize the additional information to learn how to distinguish semantically similar question-answer pairs, thus avoiding incorrect evaluation results. In general, our framework can be extended to multiple related datasets, requiring the training of only two models: one for cases with text-only input and another for cases with video information included. These two models are unified within the **UniFM** framework.

1 Introduction

In the field of psychology, the cognitive process of inferring others' mental states through the observation of their actions and verbal expressions is commonly referred to as "mindreading". This intriguing phenomenon has garnered considerable attention from various disciplines, including psychologists, neuroscientists, economists, and computer scientists, over the course of many years (Hughes and Devine, 2015). Research suggests that children who exhibit exceptional mindreading abilities tend to possess a healthier psychological well-being and superior social skills. Specifically, those who perform well on mindreading assessments are more likely to enjoy popularity among their peers (Banerjee et al., 2011) and maintain positive relationships with their fellow classmates (Fink et al., 2015). Moreover, case-control studies have revealed a higher propensity for mental health issues among individuals with impaired mindreading abilities (Cotter et al., 2018). Given these compelling findings, it becomes paramount to identify an accurate, effective, and reliable method for evaluating the mindreading capacities of children during their middle childhood and early adolescence.

[1]Our code is available at https://github.com/manic-dolphin/emnlp2023-unifm-mindreading.

Presently, standardized open-ended psychological tests are available to assess children's mindreading ability. These tests typically involve children responding to specific questions based on meticulously crafted vignettes prepared by experts (Happé, 1994; Banerjee et al., 2011). Alternatively, assessments may involve the use of short clips extracted from comedic sources (Devine and Hughes, 2013) or animated content (Castelli et al., 2000).

However, evaluating children's responses to these test questions necessitates manual assessment by extensively trained experts. This process is both labor-intensive and costly. Some work (Kovatchev et al., 2020, 2021) have attempted to tackle this challenge by integrating the aforementioned psychological tests and developing automated scoring systems to evaluate children's mind-reading ability. They constructed the MIND-CA dataset (Kovatchev et al., 2020), subsequently performed fine-tuning on a Transformer-based pre-trained model using the MIND-CA dataset, leading to noteworthy advancements in automated assessment of children's mind-reading abilities.

Despite previous research has achieved promising results, the methods employed still exhibit certain limitations. These models only considered the test's questions and children's responses as input, which brings two drawbacks. Firstly, different psychological tests may share identical questions, using this method leads to unavoidable errors during evaluation of the model, which restricts the model's generalizability to other related psychological test datasets, limiting its applicability. Secondly, disregarding the background information from psychological tests hampers the potential improvement in the model's evaluation performance. Hence, it is crucial to acknowledge the relevance and utility of this contextual information. Figure 1 illustrates an example that explains why our framework has the ability to extend to other related psychological test datasets.

To address the aforementioned issues, we firstly incorporate the psychological tests' background information into our model's input. This background information includes the complete story text that is read to the children during the psychological tests, as well as the silent comedy clips that are shown to the children. In the current context, the input may involve multiple modalities, such as the combination of clips and question-answer pairs. Moreover, considering the need for effective interaction between the background information and the question-answer pairs, we secondly introduce some effective methods from machine reading comprehension. This allows the model to not only fully utilize the background information but also seamlessly unify different input types within a single framework.

Overall, our proposed method **UniFM** (**Uni**fied **F**ramework for Measuring Children's **M**indreading Ability) can effectively addresses the issues encountered by previous models. Experimental results demonstrate that our model outperforms previous approaches, and the incorporation of multimodality information significantly enhances its performance. Moreover, in comparison to previous methods, **UniFM** exhibits the capability to handle a broader spectrum of psychological testing scenarios, thereby showcasing its potential for development into an automated, cost-free, online scoring system for evaluating children's mindreading ability.

2 Related Work

This section provides an overview of the relevant literature that forms the foundation of our research. Firstly, we present a concise introduction to the concept of mindreading and discuss two standardized psychological tests that were employed to construct the training dataset for automated scoring systems. Secondly, we provide a brief review of studies that have explored the integration of NLP techniques with psychology, highlighting the research endeavors focused on developing automated scoring systems. Throughout this paper, our primary objective is to compare our model with these existing systems. Finally, we present an overview of methods employed in the Machine Reading Comprehension task, incorporating techniques proposed in prior studies to construct our framework.

2.1 Standardized Psychological Tests for Assessing Children's Mindreading Ablility

Mindreading is a concept in psychology (Hughes and Devine, 2015) which usually is used to describe someone's ability in understanding others' thoughts, feelings, and desires. For example, a man was arguing with his wife, and finally, she said: "Well, fine." If the man is good at mindreading, he should know that his wife was still angry. Currently, there are established and standardized psychological tests that are used to assess children's

mind-reading abilities. Our study involves two of these tests, we will provide a brief description of these two assessments. Happé (1994) orally present five brief narratives to children while displaying the corresponding story text on a sizable screen. These stories encompass diverse social scenarios, such as double bluffing, cheating, misunderstandings, and lying (these stories are referred to as "Strange Story"), and each narrative concludes with a specific open-ended question. Children are then prompted to provide responses to these questions, aiming to discern the inner mental states of the characters involved. Devine and Hughes (2013) involves children viewing a series of five brief silent film clips displayed on a sizable screen. These clips are carefully chosen from a renowned silent comedy, portraying various social scenarios such as deception and misunderstanding (these clips are referred to as "Silent Film"). In alignment with the selected clips, researchers devise specific questions for children to address. Following a single viewing of each clip, children are requested to provide written responses to the questions, which are read aloud by the researchers.

2.2 Natural Language Processing for Psychology Research

Recently, many works in psychology are benefited from advanced natural language processing methods, including building chatbots to promote the mental health of their users (Tewari et al., 2021), making inferences about people's mental states from what they write on Facebook, Twitter, and other social media (Calvo et al., 2017), combining with computational algorithms to understand a suicidal patient's thoughts, such as suicide notes (Pestian et al., 2010).

Similarly, NLP techniques have also been applied to the research on building automated assessment of children's mind-reading abilities. Utilizing the aforementioned standardized psychological tests (Happé, 1994; Devine and Hughes, 2013), Kovatchev et al. (2020) created the MIND-CA dataset which consisting 11,311 question-answer pairs based on the responses. They trained a series of models (i.e., SVM, BiLSTM, Transformer) to build automated scoring systems, and obtained ideal results.

Following this work, Kovatchev et al. (2021) adopt some data augmentation methods to enhance the performance of the automated systems. The results showed that the model gained performance improvement both on the MIND-CA and the new dataset UK-MIND-20.

Despite the promising results, they only considered evaluating children's mindreading ability by using question-answer pairs as input. However, in different psychological tests, the quality of children's responses cannot be solely evaluated based on the question-answer pairs. It is crucial to integrate relevant test information in order for the model to handle situations where different psychological tests may contain the identical questions.

2.3 Machine Reading Comprehension

Machine reading comprehension (MRC) is a challenging task in NLP field (Liu et al., 2019; Zeng et al., 2020; Zhang et al., 2019). It can be divided into four categories: cloze style, multiple-choice, span prediction, and free-form answer.

For the multiple-choice task, inspired by the human's transposition thinking process of handling MRC problem, Zhu et al. (2021) proposes the Dual Multi-head Co-Attention(DUMA) to calculate the attention score between passage and question-answer pair. It boosts the model's performance on the DREAM (Sun et al., 2019) and RACE (Lai et al., 2017) datasets.

3 Method

In this section, we provide a detailed description of our proposed method. As mentioned earlier, our framework requires the training of two distinct models to accommodate different input scenarios. Hence our model is designed to handle two subtasks: the *Strange Stories Task* and the *Silent Film Task*.

Different tasks corresponds to different inputs. For the *Strange Stories Task*, we utilize the source text of the "Strange Storie" and the corresponding question-response pairs as input to our model. On the other hand, for the *Silent Film Task*, we employ the clips from the "Silent Films" and the related question-response pairs as input.

Given the inclusion of information from both the language modality and the visual modality in the input, we employ specific pre-trained models to encode this information. These pre-trained models serve as robust encoders that extract meaningful representations from the input data.

Subsequently, in order to facilitate the interaction between the representations of the text (or

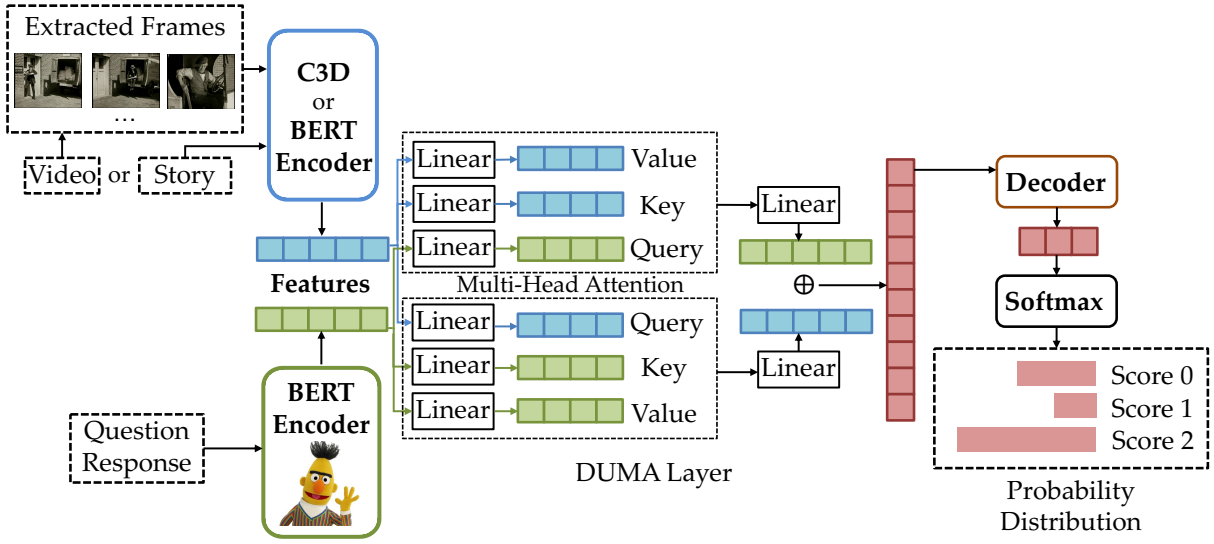

Figure 2: The architecture of our model. For inputs from different psychological test backgrounds, we need to train two models with only differences in the encoder architecture. Specifically, if the input consists of stories and question-answer pairs, we employ BERT for encoding to obtain hidden representations. If the input includes videos and question-answer pairs, we use C3D to encode the video. Once we obtain the hidden representations of the input, we utilize a dual-tower attention mechanism to interact and align these features, which is a unified process. After aligning the features, we concatenate and aggregate them, and finally decode them into a probability distribution of scores, completing the automated assessment of children's mental interpretation ability.

clips) and the representations of the question-answer pairs, we employ the "dual-tower" architecture (DUMA layer) that borrows from DUMA method (Zhu et al., 2021).

The DUMA mechanism enables our model to achieve effective alignment between different pieces of information, resulting in an integrated representation that captures the essential information of the passage (or the clips) and the associated question-answer pairs. This integrated representation is subsequently mapped into a probability distribution, serving as a measure of the model's confidence or certainty in assessing mindreading abilities. Figure 2 provides an detailed overview of the structure of our framework.

In the following subsections, we discuss in more detail our method. In Section 3.1, we describe two sub-tasks, and in Section 3.2, we provide a detailed description of the model architecture.

3.1 Task Definition

In this Section, we give the definition of the two sub-tasks of our model, for these two distinct sub-tasks, we design corresponding model architectures and unifiy them within our **UniFM** framework.

3.1.1 Strange Stories Task

In the *Strange Stories Task*, we formalize the input of the model as a triplet: (S, Q, R), where S refers to the source text of the "Strange Stories" (5 in total), and each story corresponding to a fixed question Q. We concatenate the question Q and children's response R as $Q \oplus R$. Moreover, for each triplet (S, Q, R), the ground truth of the children's mindreading ability score is defined as s, where $s \in Score$, and $Score = \{0, 1, 2\}$, which describes all the possible score of children's mindreading ability.

Given N data instances $\{S_i, Q_i \oplus R_i, s_i\}_{i=1}^{N}$ in the train set $\mathcal{D}_{train}$, model's output is defined as the $p(s|S, Q \oplus R)$. We train our model to maximize the probability: $\prod_{i=1}^{N} p(s_i|S_i, Q_i \oplus R_i)$. Toward this goal, our model must acquire the ability to accurately map the $(S, Q \oplus R)$ input into the corresponding ground truth score.

3.1.2 Silent Film Task

Similar to the *Strange Stories Task*, in the *Silent Film Task*, we formalize the input of our model as: (C, Q, R), where C denotes the clips (5 in total) of the "Silent Film", and each clip is also corresponding to a fixed question.[2] $Score$ still represents the

[2]Except for Silent Film 1, which has two questions.

ground truth of the score of children's mindreading ability. Given N data $\{C_i, Q_i \oplus R_i, s_i\}_{i=1}^{N}$, our objective aligns with the goal of the *Strange Stories Task*, which is to maximize the probability: $\prod_{i=1}^{N} p(s_i|C_i, Q_i \oplus R_i)$.

3.2 Model Architecture

In this section, we provide a comprehensive overview of the model's architecture for the two distinct sub-tasks. We discuss the design of the encoder, DUMA layer, and decoder, and elucidate the reasons behind their integration into our framework.

3.2.1 Model for the Strange Stories Task

Encoder. We use the BERT (Devlin et al., 2018) encoder to generate the representations of the source text of the "Strange Story" and the source text of question-response pairs. The pre-trained BERT model is capable of providing a powerful hidden representation of the input text. This greatly facilitates the subsequent interaction among diverse information within the model.

Let $S = [S_1, S_2, \ldots, S_m]$ and $Q \oplus R = [Q_1, \ldots, Q_n, R_1, \ldots, R_p]$ represent the sequences of the story and the question-response pair respectively, where S_i, Q_i, R_i are tokens. We denote $BERT(\cdot)$ as the BERT encoder, the encoded representations of S and $Q \oplus R$ are defined as $B_S = BERT(S)$, and $B_{QR} = BERT(Q \oplus R)$ respectively.

DUMA Layer. We utilize the DUMA layer (Zhu et al., 2021) to capture the interaction amoung the encoded information. Compared with the vanilla multi-head attention architecture (Vaswani et al., 2017), the DUMA layer adopts the dual-tower architecture, which is inspired by the cognitive processes employed by humans during reading comprehension. The distinctive architecture enables the model to fully understand the content of psychological tests and achieve alignment between both unimodal and multimodal information. As a result, it enhances the automated evaluation process, leading to improved performance.

Given the representation B_S and B_{QR}, we calculate the attention score in the following way: (1) B_S as $Query$, B_{QR} as Key and $Value$; (2) B_{QR} as $Query$, B_S as Key and $Value$. Here, the terms $Query$, Key, and $Value$ have the same meaning as Q, K, and V in the vanilla multi-head attention respectively. We denote the output of the DUMA layer as $\mathrm{DUMA}(\cdot)$. The complete computation process in the DUMA layer can be found in the Appendix A.

Decoder. We use the Multi-Layer Perceptron (MLP) to decode the output of the DUMA layer into the representation named O:

$$O = \mathrm{MLP}(\mathrm{DUMA}(B_S, B_{QR})) \tag{1}$$

Here, $O \in \mathbb{R}^l$, l denotes the number of the score (3 in total). For each data (S, Q, R, s) in the train set, the objective probability can be calculated in the same way as the Softmax function:

$$p(s|S, Q \oplus R) = \frac{\exp(O^t)}{\sum_{k=1}^{l} \exp(O^k)} \tag{2}$$

where O^t denotes the element's value in O that matches the ground truth score, O^k denotes the k-th element's value in O.

Given N data $\{S_i, Q_i \oplus R_i, s_i\}_{i=1}^{N}$ sampled in the train set, we define the loss function loss_{SS} as:

$$\mathrm{loss}_{SS} = -\sum_{i=1}^{N} \log p(s_i|S_i, Q_i \oplus R_i) \tag{3}$$

3.2.2 Model for the Silent Film Task

Encoder. Different from the *Strange Stories Task*, in the *Silent Film Task*, model's input is (C, Q, R), where C denotes a short video. For the question-answer pair, we still use the BERT encoder to obtain the output representation $B_{QR} = BERT(Q \oplus R)$. For the clip, we utilize 3D ConvNet (C3D) (Tran et al., 2014) to extract information and obtain the hidden representation $F = C3D(C)$. Simultaneously, we will ensure that the output hidden representations of the clip and text remain consistent across all dimensions.

There are two reasons why we do not utilize more advanced and larger models as video encoders. Firstly, we considered that C3D is a relatively mature and classic model with stable training performance. Secondly, our dataset contains a relatively limited variety of videos, and the need to encode highly complex information is not prominent. Therefore, C3D is sufficient for efficient information extraction. However, in the future, if our framework necessitates application in more complex scenarios, we do not exclude the possibility of employing state-of-the-art video encoders with larger parameter sizes.

DUMA Layer. Besides the hidden representation B_S of the story in the input, now it becomes the hidden representation F of the clip, the computation process in the DUMA layer remains unchanged.

Therefore, we unify the two tasks into one framework, which opens up the possibility of training the model and extending its applicability to a broader range of scenarios. Moreover, to our surprise, we found that the DUMA layer can effectively handle the interaction between multimodal information, even without adopting more advanced Video Question Answer-based methods. This validates the rationale of applying machine reading comprehension techniques to address multimodal alignment issues in our task scenario.

Decoder. The decoding process here is completely identical to the *Strange Stories Task*, so we directly provide the loss function loss_{SF}:

$$\text{loss}_{SF} = -\sum_{i=1}^{N} \log p(s_i|C_i, Q_i \oplus R_i) \quad (4)$$

4 Experiments

Firstly, we reorganized the MIND-CA dataset and obtained two datasets $\mathcal{D}_{SS}$, $\mathcal{D}_{SF}$ for the *Strange Stories Task* and the *Silent Film Task* respectively. For different tasks, we train our model on the corresponding dataset. In addition, we conduct ablation study to explore the effect of some hyperparameters and model's architecture. Finally, we perform a case study to test our model's performance in the real world.

4.1 Dataset

Based on the *Strange Stories Task* and the *Silent Film Task*, we divided the MIND-CA dataset into two parts: $\mathcal{D}_{SS}$ and $\mathcal{D}_{SF}$ respectively. The dataset $\mathcal{D}_{SS}$ consist of all the question-answer pairs associated with the "Strange Strories", with each question-answer pair accompanied by the corresponding story. Similarly, the dataset $\mathcal{D}_{SF}$ consists of all the question-answer pairs associated with the "Silent Film", with each pair accompanied with the corresponding clip. The specific construction process of the MIND-CA dataset and the organization format of the dataset can be found in Appendix B.

4.2 Baseline Models

Kovatchev et al. (2020) fine-tuned a DistilBERT (Sanh et al., 2019) on the whole MIND-CA dataset. Their model follows the framework of the traditional text classification task. Their model takes two types of data as input: 1) children's response only; 2) question and children's response. We take their model as our baselines.

Model	Input	val-Acc	test-Acc
baseline (reported)	R only	$--$	89.00
baseline (reported)	Q+R	$--$	91.00
BERT$_{SS}$	Q+R	94.11	94.66 ± 0.85
BERT$_{SF}$	Q+R	93.63	92.01 ± 1.25
UniFM$_{SS}$	S+Q+R	94.76	$\mathbf{95.49}\pm0.52$
UniFM$_{SF}$	C+Q+R	93.35	$\mathbf{94.84}\pm0.86$

Table 1: The experimental results of our models and baselines. S denotes the source text of the "Strange Stories", C denotes the clips of the "Silent Film". Q, R denote the questions and responses respectively. The accuracy of the baseline models are excerpted from (Kovatchev et al., 2020). The subscript of the model represents on which dataset the model was trained and evaluated. We select different random seeds to train and test the model 10 times. Then, we calculate the average and standard deviation of the experimental results.

In addition, considering that we divided the MIND-CA dataset into two parts for our two different tasks, we fine-tuned two extra BERT models on the dataset $\mathcal{D}_{SS}$ and $\mathcal{D}_{SF}$ respectively. To be consistent with the two baselines, we only take question-answer pairs as the model's input. These two sets of experiments eliminated the potential impact of dataset partitioning on the experimental results. We denote these two models as BERT$_{SS}$ and BERT$_{SF}$.

4.3 Experimental Settings

In the *Strange Stories Task*, we train our model on the dataset $\mathcal{D}_{\mathcal{SS}}$, we use BERT$_{\text{base}}$ as the encoder, and use $k = 1$ layer of the DUMA Layer.

In the *Silent Film Task*, We train our model on the dataset $\mathcal{D}_{SF}$. We use BERT$_{\text{base}}$ to encode the question-answer pairs. For each clip, we firstly utilize the OpenCV[3] tools to extract a fixed number of frames. Typically we extract 32 frames. We implement a C3D model to encode these frames.

For both tasks, we adopt Adam (Kingma and Ba, 2014) as our optimizer with the $\text{learning rate} = 5 \times 10^{-5}$. We train our models on four NVIDIA Geforce RTX 2080Ti GPUs.

4.4 Main Results

Table 1 shows main experimental results. Compared with the baselines, our models show better performance. In the *Strange Stories Task*, our model gains 95.49% accuracy on $\mathcal{D}_{SS}$, which outperforms the baselines and the BERT$_{SS}$. In the *Silent Film Task*, our model gains 94.86% accuracy on $\mathcal{D}_{SF}$, outperforms the baselines and the

[3]https://github.com/opencv/opencv

Model	Input	test-Acc
UniFM$_{SS}$	S + Q + R	$\mathbf{95.49}\pm0.52$
UniFM$_{SS}$ w/o DUMA	S + Q + R	94.79 ± 0.74
BERT$_{SS}$	Q + R	94.66 ± 0.85

Table 2: Ablation study of the proposed model on the *Strange Stories Task*. S denotes a stroy from the "Strange Stories", Q and R denote the question, response respectively.

Model	Input	test-Acc
UniFM$_{SF}$	C + Q + R	$\mathbf{94.84}\pm0.86$
UniFM$_{SF}$ w/o DUMA	C + Q + R	91.78 ± 1.27
BERT$_{SF}$	Q + R	92.01 ± 1.25

Table 3: Ablation study of the proposed model on the *Silent Film Task*. C denotes a clip from the "Silent Film", Q and R denote the question, response respectively.

BERT$_{SF}$. This result gives us a strong proof that injecting rich background information from psychology tests into the model can indeed improve the ability of the automated scoring system.

The improvement in model performance can be attributed to two factors. Firstly, the incorporation of supplementary background information from psychological tests into the input has contributed to the improvement. Despite the limited diversity of stories and videos within the dataset, this inclusion has proven beneficial for enhancing the model's performance. Secondly, the introduction of DUMA allows for effective alignment between diverse pieces of information. This enables the model to gain a better understanding of psychological tests and leverage the additional information to enhance its evaluation capability.

5 Ablation Study

In this section, we will dive further into the impact of the DUMA layer on our experimental results in the *Strange Stroies Task*. Additionally, we will explore the significance of incorporating visual information in improving model performance in the *Silent Film Task*. We will also investigate the influence of the DUMA layer on aligning multimodality information.

To investigate whether DUMA laer would benefit the model's performance, in the *Strange Stories Task*, we remove the DUMA layer in our model, train and test it on the dataset D_{SS}.

The results presented in the Table 2 demonstrate that the removal of the DUMA layer from **UniFM**$_{SS}$ leads to a decline in the model's performance. Moreover, **UniFM**$_{SS}$ without the DUMA layer achieves better performance compared to the BERT$_{SS}$ model trained exclusively on question-answer pairs, as discussed previously. This proves that additional background information can enhance the performance of the model. Moreover, the DUMA layer facilitates a better understanding of the background information, leading to optimal model performance.

Similar to the *Strange Stroies Task*, in the *Silent Film Task*, we train a **UniFM**$_{SF}$ model without DUMA layer on the dataset D_{SF}. Results in the Table 3 shows that, removing the DUMA layer extremely decreases the model's performance. This confirms that simply adding videos to the input does not directly improve the model's performance. We also need to employ effective alignment techniques. The DUMA layer allows the model to better understand the background information, achieve alignment between multimodal information, and attain optimal performance.

Pair	P-Value
(**UniFM**$_{SS}$, BERT$_{SS}$)	0.01%
(**UniFM**$_{SS}$, **UniFM**$_{SS}$ w/o DUMA)	0.02%
(**UniFM**$_{SF}$, BERT$_{SF}$))	0.02%
(**UniFM**$_{SF}$, **UniFM**$_{SF}$ w/o DUMA)	0.01%

Table 4: Paired t-test results. Each pair consists of two models that need to be compared. We conducted multiple experiments and calculated the corresponding p-value for each pair. A lower p-value indicates a more significant performance difference between the two models in that pair.

6 Statistical Significance Analysis

To confirm that our method indeed improves the model's performance, we conduct the paired t-test on the *Strange Story Task* and the *Silent Film Task*. Specifically, each pair consists of two models. We train and test these two models multiple times, each time with a fixed random seed shared by both models. We record the results of the models on the test set and then calculate the p-value based on the results obtained from multiple tests. Table 4 presents the experimental results. These results demonstrate that the differences in performance between each pair of models are statistically significant, which essentially confirms the effectiveness of our model.

7 Quantitative Analysis

7.1 Number of DUMA Layers

We investigate the influence of the number of DUMA Layers. We only change the number of DUMA Layers and calculate the accuracy. The results are shown in figure 3. In the *Strange Stories Task*, our model show good performance both when $k = 1$ and $k = 4$. When $k = 6$, model's performance drop dramatically. It is not strange that increasing the number of DUMA Layers does not bring promising improvement: there are only five different types of stories in the *Strange Stories Task*, interacting the text of stories and question-response pairs once is enough to capture the hidden information. Stacking too many layers makes the model harder to train and may lead to the overfitting problem. These analyses are also applicable to the *Silent Film Task*.

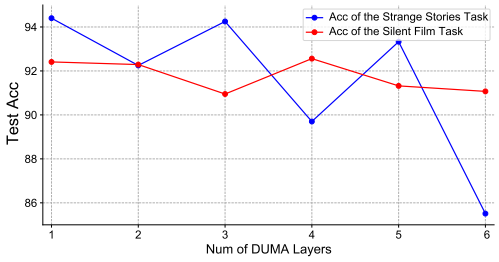

Figure 3: The performance change under different the number of DUMA Layers. The blue line shows the test accuracy on $\mathcal{D}_{SA}$ and the red line shows the test accuracy on $\mathcal{D}_{SF}$.

7.2 Number of Frames and Epochs

The silent film contains 5 clips, clips last 27.6 seconds on average. For each clip, we need to extract a fixed number of frames and convert the frames into tensor form. In this section, we will investigate the influence of the number of frames in the *Silent Film Task*. Figure 4 shows the results. We found that extracting 32 frames enables our model to achieve the best performance. Due to the limitation of the video types, we think 32 frames are sufficient for the model to obtain useful information.

In the *Silent Film Task*, we have not utilized some pre-trained models while the training data is limited, instead we train a C3D from scratch. However, training too much will lead to overfitting while the lack of training may weaken the advantage of C3D. In this section, we seek to find the trade-off by changing the number of training epochs. Figure 4 shows the results.

We found that when we trained our model for 4 epochs, our model perform the best. However, different epochs have not brought a significant effect on our model. We believe that this is due to the lack of a sufficient number of diverse videos in the dataset.

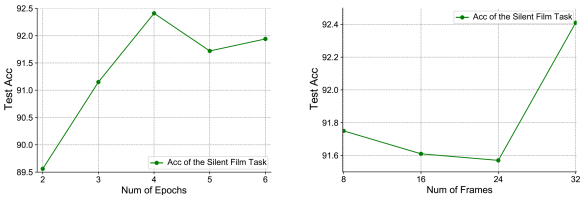

Figure 4: Performance change under different hyper-parameter choices. We investigate it in the *Silent Film Task*. The left shows the influence of different epochs and the right shows the influence of different frames we extract from one cilp.

7.3 Case Study

We conduct a case study to verify our system's performance in the real world. We selected several well-educated English-speaking children aged between 10 and 11. Under the guidance of our detailed instructions, they completed the mindreading tests (include the *Strange Stories Task* and the *Silent Film Task*.) accompanied by their guardians. We collected their responses and hired experts to score the responses manually. These scores were used as the ground truth. Then we tested our model on these new data.

We utilize the Pearson correlation coefficient to measure the difference between the model's predictions and the ground truth. The complete calculation process can be found in Appendix C. Let r represent the Pearson correlation coefficient. In the *Strange Stories Task* we got $r = \mathbf{0.83}$ and in the *Silent Film Task* we got $r = \mathbf{0.77}$, which indicates that our models' evaluation have strong correlation with the experts' assessment. Therefore, our model has the potential for being applied to other related scenarios.

8 Conclusions

In this paper, we propose a new framework **UniFM** for automated scoring children's mindreading ability. We fuse additional multimodality information from mindreading tests into the model's input. Different from previous methods, this novel frame-

work can be used to train on more than one psychological test dataset. Experimental results have shown that our model achieves great performance on different tasks and outperforms the previous methods. In addition, the case study indicate that our models have the potential to be transferred in other scenarios.

Limitations

The MIND-CA dataset only has limited question source and question-response pairs, which is not enough for training a DNN model with a large scale of parameters. It would be intriguing to see the performance of our model under a larger dataset which contains a variety of stories and videos.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments. This work was supported by National Natural Science Foundation of China (No. 62076068), and Shanghai Municipal Science and Technology Project (No. 21511102800).

References

Robin Banerjee, Dawn Watling, and Marcella Caputi. 2011. Peer relations and the understanding of faux pas: longitudinal evidence for bidirectional associations. *Child Development*.

Rafael A. Calvo, David Milne, M. Sazzad Hussain, and Helen Christensen. 2017. Natural language processing in mental health applications using non-clinical texts†. *Natural Language Engineering*.

Fulvia Castelli, Francesca Happé, Uta Frith, and Chris D. Frith. 2000. Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. *NeuroImage*.

Jack Cotter, Kiri Granger, Rosa Backx, Matthew Hobbs, Chung Yen Looi, and Jennifer H. Barnett. 2018. Social cognitive dysfunction as a clinical marker: A systematic review of meta-analyses across 30 clinical conditions. *Neuroscience & Biobehavioral Reviews*.

Rory T. Devine and Claire Hughes. 2013. Silent films and strange stories: theory of mind, gender, and social experiences in middle childhood. *Child Development*.

Rory T. Devine and Claire Hughes. 2016. Measuring theory of mind across middle childhood: Reliability and validity of the silent films and strange stories tasks. *Journal of Experimental Child Psychology*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *north american chapter of the association for computational linguistics*.

Elian Fink, Sander Begeer, Candida C. Peterson, Virginia Slaughter, and Marc de Rosnay. 2015. Friendlessness and theory of mind: A prospective longitudinal study. *British Journal of Development Psychology*.

Francesca Happé. 1994. An advanced test of theory of mind: understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*.

Claire Hughes and Rory T. Devine. 2015. Individual differences in theory of mind from preschool to adolescence: Achievements and directions. *Child Development Perspectives*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv: Learning*.

Venelin Kovatchev, Phillip Smith, Mark Lee, and Rory T. Devine. 2021. Can vectors read minds better than experts? comparing data augmentation strategies for the automated scoring of children's mindreading ability. *meeting of the association for computational linguistics*.

Venelin Kovatchev, Phillip Smith, Mark Lee, Imogen Grumley Traynor, Irene Luque Aguilera, and Rory T. Devine. 2020. "what is on your mind?" automated scoring of mindreading in childhood and early adolescence. *arXiv: Computation and Language*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *empirical methods in natural language processing*.

Shanshan Liu, Xin Zhang, Sheng Zhang, Hui Wang, and Weiming Zhang. 2019. Neural machine reading comprehension: Methods and trends. *Applied Sciences*, 9(18):3698.

John Pestian, Henry Nasrallah, Pawel Matykiewicz, Aurora J. Bennett, and Antoon Leenaars. 2010. Suicide note classification using natural language processing: A content analysis. *Biomedical Informatics Insights*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv: Computation and Language*.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Abha Tewari, Amit Chhabria, Ajay Singh Khalsa, Sanket Chaudhary, and Harshita Kanal. 2021. A survey of mental health chatbots using nlp. *Social Science Research Network*.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2014. Learning spatiotemporal features with 3d convolutional networks. *international conference on computer vision*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *neural information processing systems*.

Changchang Zeng, Shaobo Li, Qin Li, Jie Hu, and Jianjun Hu. 2020. A survey on machine reading comprehension—tasks, evaluation metrics and benchmark datasets. *Applied Sciences*, 10(21):7640.

Xin Zhang, An Yang, Sujian Li, and Yizhong Wang. 2019. Machine reading comprehension: a literature review. *arXiv preprint arXiv:1907.01686*.

Pengfei Zhu, Zhuosheng Zhang, Hai Zhao, and Xiaoguang Li. 2021. Duma: Reading comprehension with transposition thinking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:269–279.

A DUMA Layer

In this section, we present the complete process of computing attention and obtaining the final output in the DUMA layer.

Given the representation B_S and B_{QR}, we calculate the attention score in the following way: (1) B_S as $Query$, B_{QR} as Key and $Value$; (2) B_{QR} as $Query$, B_S as Key and $Value$.

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{Q(K^T)}{\sqrt{d_k}})V$$

$$\text{head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

$$\text{MHA}(Q,K,V) = \text{Concat}(\text{head}_1,\ldots,\text{head}_\text{h})W^O$$

$$\text{MHA}^{(1)} = \text{MHA}(B_S, B_{QR}, B_{QR})$$

$$\text{MHA}^{(2)} = \text{MHA}(B_{QR}, B_S, B_S)$$

$$\text{DUMA}(B_S, B_{QR}) = \text{FUSE}(\text{MHA}^{(1)}, \text{MHA}^{(2)}) \quad (5)$$

where d_q, d_k, d_v denote the dimension of the $Query$, Key, $Value$. $W_i^Q \in \mathbb{R}^{d_{model} \times d_q}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$, h denotes the number of the attention heads, and d_{model} denotes the fixed dimension of model. $W^O \in \mathbb{R}^{hd_v \times d_{model}}$, $\text{MHA}^{(1)}$, and $\text{MHA}^{(2)}$ represent the aforementioned two kinds of attention representations respectively. $\text{FUSE}(\cdot)$ refers to the concatenation operation. The output of the DUMA Layer is denoted as $\text{DUMA}(\cdot)$.

B Dataset Details

To create the MIND-CA dataset, Kovatchev et al. (2020) recruited 1,066 English-speaking children aged between 7.25 and 13.53, from 46 different classrooms in 13 primary and 4 secondary schools in England between 2014 and 2019. The children took part in a whole-class testing session lasting approximately 1 hour, led by a trained research assistant using a scripted protocol (Devine and Hughes, 2016). Each child answered 11 questions - five in the Strange Story Task and six in the Silent Film Task. They obtained a total of 11,726 question-answer pairs. The paper test booklets were digitalized and manually scored by two postgraduate research assistants and the test developer.

Our dataset is constructed in the format shown in Table 5. In the D_{SS}, each data instance consists of a story , a corresponding question, the child's response to the question, and an expert-labeled score representing the child's mentalizing ability. In the D_{SF}, each data instance consists of a clip , a corresponding question, the child's response to the question, and a score.

Dataset	Col-1	Col-2	Col-3	Col-4
$\mathcal{D}_{SF}$	Story	Question	Response	Score
$\mathcal{D}_{SS}$	Clip	Question	Response	Score

Table 5: The format of dataset $\mathcal{D}_{SS}$ and $\mathcal{D}_{SF}$.

C Compute the Pearson Correlation Coefficient

Let $P = \{p_1, \ldots, p_n\}$, $T = \{t_1, \ldots, t_n\}$ denote the predicted results and true labels respectively, where p_i, t_i represent the children's mindreading ability score, n denotes the total number of the responses. We calculate the Pearson correlation coefficient in the following way:

$$r = \frac{\sum_{i=1}^{n}(p_i - \overline{P})(t_i - \overline{T})}{\sqrt{\sum_{i=1}^{n}(p_i - \overline{P})^2}\sqrt{\sum_{i=1}^{n}(t_i - \overline{T})^2}} \quad (6)$$