

# Learning Multimodal Cues of Children’s Uncertainty

Qi Cheng<sup>1\*</sup>, Mert İnan<sup>4\*</sup>, Rahma Mbarki<sup>3\*</sup>, Grace Grmek<sup>2\*</sup>, Theresa Choi<sup>1\*</sup>, Yiming Sun<sup>3</sup>, Kimele Persaud<sup>3</sup>, Jenny Wang<sup>3</sup>, Malihe Alikhani<sup>4</sup>

<sup>1</sup> University of Pittsburgh, PA, USA <sup>2</sup> Harvard Medical School, MA, USA

<sup>3</sup> Rutgers University, NJ, USA <sup>4</sup> Northeastern University, MA, USA

{qic69, tec63}@pitt.edu, ggrmek@mgh.harvard.edu,

{rm1218, kjg117, jinjing.jenny.wang}@rutgers.edu,

{inan.m, alikhani.m}@northeastern.edu

## Abstract

Understanding uncertainty plays a critical role in achieving common ground (Clark et al., 1983). This is especially important for AI systems that communicate with their human interlocutors in a multimodal way, whether collaborating to solve a problem or guiding the user through a challenging concept. One particularly interesting facet of this is uncertainty. In this work, for the first time, we present a dataset annotated in collaboration with developmental and cognitive psychologists for the purpose of studying nonverbal cues of uncertainty. We then present an analysis of the data, studying different roles of uncertainty and its relationship with task difficulty and performance. Lastly, we present a multimodal machine learning model that can predict uncertainty given a real-time video clip of a participant and outperforms a baseline multimodal transformer model. This work informs research on cognitive coordination between human-human and human-AI interaction and has broad implications for the field of cognitive science. The anonymized version of our data and code will be publicly available upon the completion of the required consent forms and data sheets.

## 1 Introduction

Recognizing uncertainty in interlocutors plays a crucial role in successful face-to-face communication, and it is key to achieving common ground (Clark et al., 1983). An educational system detecting uncertainty in students in a classroom could alert educators to potential windows of intervention, while a conversational agent detecting uncertainty in their interlocutor might then work to re-establish common ground. As a signal that is both communicative in nature but generally non-verbalized, identifying uncertainty is an important and challenging task for AI systems. To accurately capture uncertainty signals in others, human listeners learn to identify facial expressions, hand gestures, vocal

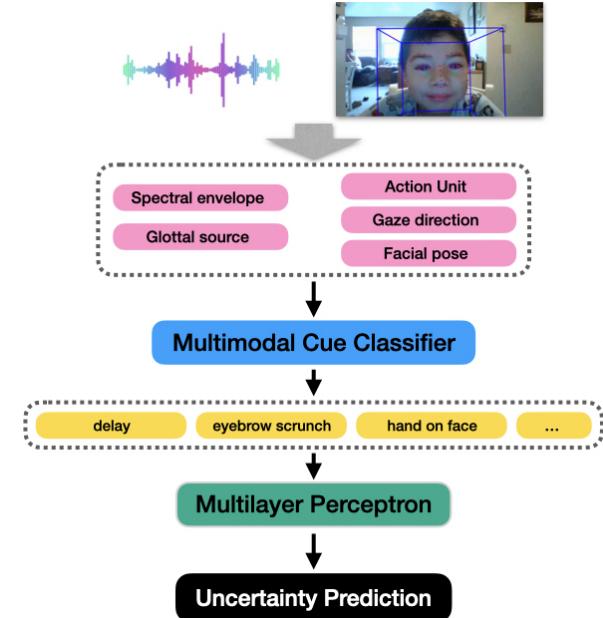


Figure 1: A diagram of our multimodal machine learning model. After identifying uncertainty cues in the multimodal transformer, the model passes the cues onto a final multilayer perceptron classifier to output whether the child is expressing uncertainty or not.

dynamics, or silence. AI systems that aim to collaborate and coordinate with users in a human-like manner also need to understand these signals of uncertainty. In this paper, we study this problem in the context of a tutoring system. The goal of the system is to teach children, ages 4 to 5, counting.

Therefore, rather than relying on a user to identify and explicitly communicate to systems when they are uncertain, it is important to understand the different ways and situations in which uncertainty may be communicated.

In this work, for the first time, we consider the task of modeling uncertainty in young children performing a number sense task. We work with children in ages 4-5, as this is a particularly important task to address in children. The task of detecting uncertainty is especially challenging in young chil-

dren. Although young children may experience more uncertainty and learn faster, they are also more limited in the ability to express themselves verbally (Blanco and Sloutsky, 2021).

We first identify potential cues of uncertainty presented in different modalities (e.g., spontaneous verbal responses, hand gestures, facial expressions, hesitation) and specifically examine the relationship between task difficulty, task performance, and exhibited levels of uncertainty. We then use these cues to inform an ensemble model that first identifies these cues from multimodal data and then predicts uncertainty from these cues (Figure 1).

This work informs research on cognitive coordination between human-human and human-AI collaboration. In addition to contributing a novel annotated multimodal dataset of uncertainty in children, we address the following questions. How do children display uncertainty beyond explicit verbal expression? When does uncertainty happen? Can a multimodal transformer model use the above knowledge to detect uncertainty in real-time reliably? Through what way are young children expressing their uncertainty the most? Is it possible for a multimodal transformer model to effectively identify uncertainty in real-time?

The structure of the paper is as follows. Section 2 presents a brief background of relevant literature on uncertainty. Section 3 provides details about the novel dataset and annotation protocol, as well as analyses of the dataset. Sections 4 and 5 discuss the computational experiments and their results. Section 6 presents a case study on the ways children express uncertainty based on their age.

## 2 Related Work

We cover related works of uncertainty in three sections: uncertainty in children and existing uncertainty protocols, the Approximate Number System, and the usage of multimodal models in understanding human cognition.

### 2.1 Uncertainty in children and existing uncertainty protocols

Adults are generally more direct and communicate their uncertainty via explicit verbal cues. Children, however, lack this insight into their own uncertainty making the detection of it more difficult from an outsider’s perspective. How this uncertainty can be unraveled and made sense of remains a complex problem.

What has consistently been established, however, is that children consistently communicate their uncertainty through the use of various facial, auditory, and gestural cues. For example, Harris et al. (2017) found that children are very expressive when they are uncertain. In the presence of an adult, these expressions may be communicated via hand flaps, questions, and utterances of ignorance, such as “I don’t know.” However, when children are alone, these same signals can be representative of signals of uncertainty. In the past, researchers have attempted to codify behaviors associated with communicating uncertainty by parsing through these various cues and creating annotation protocols.

Previously, researchers Swerts and Krahmer (2005) aimed to detect uncertainty in audiovisual speech by coding for different audiovisual cues in both adults and children. Their protocol consisted of audio cues, such as speech fillers and speech delays, and facial movements, such as eyebrow movement and smiling. While the protocol included both audio and visual cues, the cues that were noted were limited. Another protocol developed by Mori and Pell (2019) looked at solely visual cues signaling uncertainty in speech communication. These cues included changes in gaze direction, facial expressions, and embarrassed expressions. An additional protocol developed by Ricci Bitti et al. (2014) looked at uncertainty through facial expression entirely.

However, while these protocols are indeed useful, they lack the specificity necessary for our goal of pinpointing various multimodal cues associated with uncertainty. There are other various protocols, but they are also limited, typically adhering to one modality. Consequently, we expanded upon these existing protocols and included other multimodal behaviors grounded in developmental and cognitive psychology, and presented multimodal machine learning models that can predict these cues and their association with uncertainty.

### 2.2 The Approximate Number System (ANS)

Children have an intuitive sense of numbers relying on the Approximate Number System (ANS). The ANS obeys Weber’s Law, where one’s ability to differentiate between two quantities depends on the ratios of those quantities (Dehaene, 2011; Odic and Starr, 2018). The smaller the ratio, the more difficult it is to discriminate between quantities and the more uncertainty there is in the participants’

internal representations. Previous research showed that children perform better on a numerical comparison task when given a scaffolded, Easy-First numerical task starting with easier trials (e.g., 10 vs. 5) and progressing to harder ones (e.g., 10 vs. 9), compared to children seeing the same exact trials in the reversed order (i.e., Hard-First), an effect termed “confidence hysteresis” (Odic et al., 2014). This implies that confidence is built by gradually working up to harder tasks, resulting in better performance, whereas starting out with more difficult tasks reduces confidence, resulting in worse performance.

Due to its effectiveness at generating confidence or lack thereof in participants, such a numerical comparison task would be the ideal method for measuring behaviors associated with uncertainty. As such, the present study aims to fulfill this objective by implementing this “confidence hysteresis” paradigm into the task children are given.

### 2.3 Usage of multiple modalities in understanding human cognition

Multimodal models have been shown to improve performance on certain tasks by grounding some aspects of the human condition with features beyond text. Leveraging multiple modalities is particularly applicable in cases where text may miss key insights, such as sarcasm detection (Castro et al., 2019).

This is true, especially for tasks surrounding human cognitive states, such as depression prediction (Morales et al., 2018), sentiment detection (Yang et al., 2021), emotion recognition (Morency et al., 2011), and persuasiveness prediction (Santos et al., 2016). Tasks involving such complex labels benefit from multiple modalities due to the richness of the data streams. In addition to what is being said, how it is being said (pitch, facial expression, body language, gesture) is also crucial (Beinborn et al., 2018).

### 2.4 Predicting uncertainty

For the task of predicting uncertainty in particular, while there are no multimodal attempts at the task, there have been attempts at predicting uncertainty from the audio modality through prosodic features. Dral et al. (2011) reported that prosodic features were successful in detecting speaker uncertainty in spoken dialogue with a 75% accuracy. Pon-Barry and Schieber (2011) had similar findings with prosodic features. Interestingly, Pon-Barry and

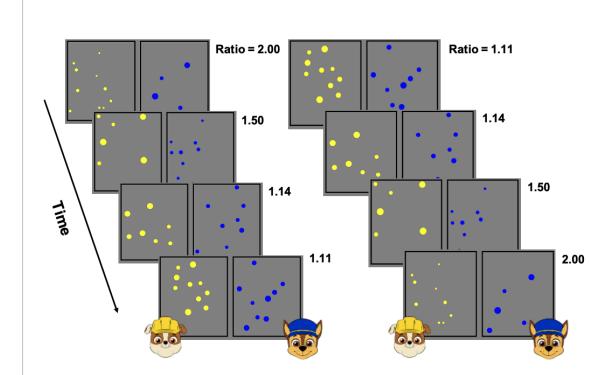


Figure 2: Schematic of experimental procedure depicting the Easy-First condition on the left and the Hard-First condition on the right. As time progresses throughout the task, the trials advance from easier ratios (2.0) to hard ratios (1.11) in the Easy-First condition. Whereas in the Hard-First condition, trials move in reverse order from hard ratios (1.11) to easy ratios (2.0) as time progresses.

Shieber also note that self-reported states of certainty and perceived states have strong mismatches. In this paper, we address this by controlling task difficulty to affect a participant’s level of uncertainty externally.

## 3 Data

**Participants** A group of 68 children between the ages of 4 and 5 years old ( $M_{age} = 5; 0$ ;  $SD_{age} = 6.88$  months; 28 females) was recruited through Lookit, an online platform for developmental studies (Scott and Schulz, 2017). Thirty-six parents identified their child as White, six as Asian, three as from Hispanic, Latino, or Spanish origins, and the rest identified as multi-racial. All but three parents reported having a college degree or higher level of education. After completing the study, compensation was sent in the form of a \$5 gift card. Each child participates in 30 trials which are, on average, 8 seconds long. In total, our data is composed of 16,320 seconds of video data.

**Task** Participants were given an Approximate Number System manipulation task adapted from Wang, Halberda, & Feigenson (2021) designed to impact children’s certainty about numerical quantities. Children were presented with two arrays of dots paired with two cartoon characters (Figure 2) and asked to play a guessing game of who has more dots.

Both characters and their array of dots appeared for 2500 ms before disappearing. This short display

duration was chosen to ensure that children did not have sufficient time to count. Children were then asked to click on the side of the screen that shows the greater amount of dots. Children were given immediate audio feedback for each trial once they chose their response.

Children completed 30 trials with the following number pairs: 10:9 dots (1.11 ratio), 8:7 (1.25 ratio), 14:12 (1.17 ratio), 10:8 (1.13 ratio), 9:6 (1.5 ratio), and 10:5 (2 ratio). In half of the trials, arrays with a greater number of dots had a greater, congruent cumulative area. In the other half of the trials, arrays with the greater number of dots had a smaller, incongruent cumulative area.

Children were randomly assigned to either the Easy-First or Hard-First conditions. In the Easy-First condition, trials advanced from the easier trials (e.g., 10:5) to the harder trials (e.g., 10:9) in a staircasing order following the design of Wang et al. (2021). Whereas in the Hard-First condition, trials move in reverse order from hard ratios (e.g., 10:9) to easy ratios (e.g., 10:5).

**Annotation Procedure** Annotators first watched the video muted so as not to be influenced by the vocal feedback from the task since whether or not the child answered right or wrong may lead them to over/under-interpret certain cues. During this first watch, they marked all present physical cues as indicated by the protocol. On their second watch, annotators unmuted the video, and marked all verbal cues. If the cue was not present, the corresponding cell was left empty.

### 3.1 Annotation Protocol

In collaboration with a team of developmental psychologists and cognitive psychologists, we have collected and designed a protocol that aims to study uncertainty, particularly expressed non-verbally. Through pilot studies observing and annotating our data, we iteratively defined our protocol and constructed a list of signals that we observed as signs of uncertainty. In order to account for a broad spectrum of possible behaviors, our protocol spans multiple modalities and includes facial, gestural, and auditory cues. The protocol can be found in Table 1 with supplemental example images in Figure 4. The Rutgers University Institutional Review Board approved the research, and all parents of children provided verbal consent prior to children’s participation. However, only some of the parents agreed to allow the video and voice recordings of

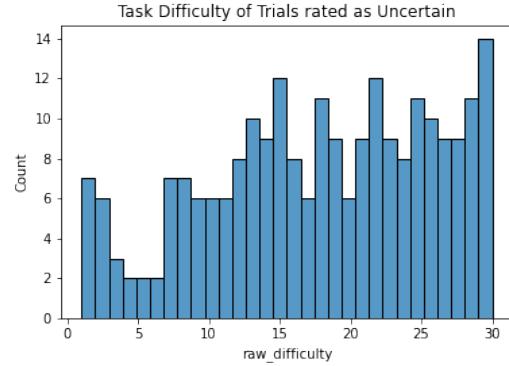


Figure 3: The distribution of uncertain trials with task difficulty on a scale of 1 (easiest) to 30 (hardest). Uncertainty shows a strong correlation with task difficulty ( $r(58) = -.927, p < .01$ ).

children to be shared publicly.

### 3.2 Analysis

In this section, we analyze our annotations, identify any significant cues that contribute to detecting uncertainty, and explore when different cues occur.

#### When were children annotated as uncertain?

The annotations are split 13.8/5.3/80.9 between the labels *uncertain/unclear/non-uncertain*. Of all the annotated trials, 79.3% were correct, of which 82.4% were rated as having no uncertainty. In other words, most of the trials are within the children’s capability and confidence. While there is a significant class imbalance between uncertain trials and non-uncertain trials, the distribution is realistic.

#### Are uncertainty and task difficulty related?

As shown in Figure 3, uncertainty was found to be highly correlated with task difficulty,  $r(58) = -.927, p < .01$ . Both ratio of dot size and size control factor into the difficulty of a trial (with a smaller ratio and the presence of size control both indicating a harder trial).

#### Are uncertainty and task performance related?

Contrarily, there was not a substantial correlation between uncertainty and task correctness,  $r(58) = .290, p > .4$ . This makes sense as the accuracy and ease of a task are not necessarily intertwined; a participant may make a mistake on an easy trial or get lucky on a difficult trial.

#### How demographics affect uncertainty?

We analyze participant demographics in terms of age and gender. In terms of frequency of expressing uncertainty, we found that the average participant age of the uncertain trials is younger than that of the non-uncertain trials, though they are not sig-

Cue	Description
<b>Delay</b>	The participant delayed their decision-making with a pronounced pause
<b>Eyebrow raise</b>	The participant raised their eyebrows
<b>Eyebrow scrunch</b>	The participant markedly scrunched their eyebrows or squinted their eyes
<b>Filled pause</b>	Utterances such as “umm,” “hmm,” or “uh.”
<b>Frustrated noise</b>	Sounds of verbal frustration, such as sighing, groaning, and growling
<b>Funny face</b>	The participant grimaced or made an unconventional facial expression
<b>Hand on face</b>	Any kind of movement that includes a participant putting a hand on their face
<b>Head tilt</b>	The participant tilted their head to either side while making their decision
<b>Look away</b>	The participant was distracted and not paying attention to the task
<b>Look to adult</b>	The participant looked towards their parent when making their decision
<b>Shoulder movement</b>	The participant made a pronounced shoulder movement, such as shrugging
<b>Smile</b>	The participant smiled
<b>Verbal cues</b>	Any spoken words

Table 1: Categories in our annotation protocol.



Figure 4: Examples of (from left to right) *eyebrow raise*, *eyebrow scrunch*, *hand on face*, *funny face*, and *smile*

nificantly different. Similarly, female participants have slightly more uncertain trials, and male participants have slightly more non-uncertain trials, but the results are not significant.

In terms of the choice of cues, however, there were some gender differences. Female participants exhibited more of the *filled pause* cue, while male participants exhibited more of the *smile* and *shoulder movement* cues. The full table of comparisons can be found in Appendix Table 5.

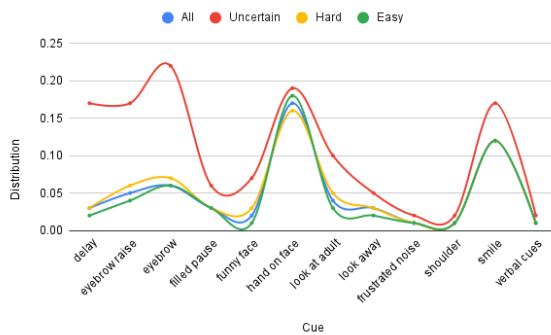


Figure 5: Distribution of uncertainty cues across all/uncertain, difficult/easy. We can see that *delay*, *eyebrow raise*, and *eyebrow scrunch* are significantly more frequent in uncertain trials.

### Which cues occur the most?

The percentage that each cue appears in all uncertain trials can be found in Figure 5. We can see that in general, *hand on face* and *smile* are the most common cues, appearing in 17% and 12% of all trials, respectively.

Interestingly, when looking only at uncertain trials, while *hand on face* and *smile* remain common, other cues also begin to appear more frequently. In particular, *eyebrow scrunch*, *eyebrow raise* and *delay* are now equally, if not more common, appearing in 22% and 17% of all uncertain trials. This is promising, as cues that frequently appear in uncertain trials but are not so common throughout all trials can be valuable indicators of uncertainty.

### Which cues occur in difficult trials as opposed to easy trials?

The percentage that each cue appears in hard and easy trials can also be found in Figure 5. Hard trials are defined as half of the trials with a more difficult ratio (1.11 to 1.17), and easy trials are half with ratios of 1.25 to 2.

We find that if the trial is hard, the participant is slightly more likely to make more cues in general. In particular, they’re more likely to exhibit the *look*

*at adult* or *funny face* cues. Meanwhile, if the trial is easy, they might instead show *hand on face*.

This is interesting as it shows support for the potential to differentiate between stages of uncertainty. Namely, if a child or student expresses uncertainty at an easier task, this could be out of a lack of confidence (*I'm generally familiar with this and have an idea on how to do it, but I need a little help.*) or another factor that may entail minor assistance. Meanwhile, if facing a more challenging or perhaps even a completely new task, they may feel a more difficult uncertainty (*I don't know where to start.*) that requires more involved guidance.

This possible distinction in stages of uncertainty may lead to possibilities for a more precise intervention in the context of education and tutoring systems. For instance, if notified about a student exhibiting the former uncertainty, the teacher might engage with small hints and encouragement to maximize the student's learning. However, if a student exhibits the latter uncertainty, the teacher can offer hands-on guidance, such as checking foundational concepts.

## 4 Approach

With the goal of predicting uncertainty from multimodal signals, we conducted experiments with three approaches: learning uncertainty from proposed cues, a multimodal transformer-based model, and an ensemble learning approach that first predicts cues from the multimodal features and then predicts uncertainty from those cues.

### 4.1 Multimodal Features

For video features, we take facial action units, gaze direction, and facial pose from the OpenFace toolkit (Baltrušaitis et al., 2018). Action units (AUs) are coded for facial muscle movements, which are indicative of various facial expressions (Tian et al., 2001). For audio features, we extracted glottal source and spectral envelope features using Degottex (2014)(v1.4.2). For text features, we then passed GloVe embeddings of each trial's annotated transcription to the model (Pennington et al., 2014). It should be noted that as a task that does not ask the participant to speak, most trials contain no text.

### 4.2 MuLT Model

Given the cost of annotation, an ideal uncertainty prediction system would take multimodal data of the participant as features and be able to make

real-time predictions on the participant's level of uncertainty. To test this goal, we first experiment with an end-to-end model. Specifically, we use the Multimodal Transformer proposed in Tsai et al. (2019) on audio, video, and text data from videos of the participants. This model is uncertainty cue-agnostic, as it contains no information about our annotated cue categories.

### 4.3 Contrastive Learning

Our dataset has high-dimensional features: 710 dimensionalities for each video frame, 71 for each second of corresponding audio, and 30 for each word in the corresponding text. Training a prediction model end-to-end in a high-dimensional feature space focuses the model on local differences in the latent space instead of the global relationships between classes. To overcome this challenge, we also tested a contrastive learning procedure with a custom loss function. This method encourages the model to learn representations that are close for positive pairs and far apart for negative pairs and better discriminates between different classes. The details of the loss function and our weighted sampling strategy are given in Appendix B.1.

### 4.4 Annotation-based Ensemble learning

We further propose an ensemble learning approach illustrated in Figure 6, that first predicts each of the annotator cues found to be significantly correlated to the annotator prediction of uncertainty and then predicts uncertainty using the trained classifier. We compare this proposed model to the previous end-to-end multimodal transformer, as well as the unimodal transformers. We choose only to predict the cues that were used in the decision tree classifier (i.e., *delay*, *eyebrow raise*, *eyebrow scrunch*, *look at adult*, and *hand on face*). **Multimodal transformer model** is used as the classifier for uncertainty.

## 5 Experimental Evaluation

In order to determine the viability of a multimodal uncertainty prediction model, we detail the results from each of our computational experiments and models. The implementation details of the experiments and models are given in the Appendix C.

**Baselines** We employed two baseline models for comparison. The first is a simple multi-modal neural network that separately processes video, audio, and text inputs and combines the features for

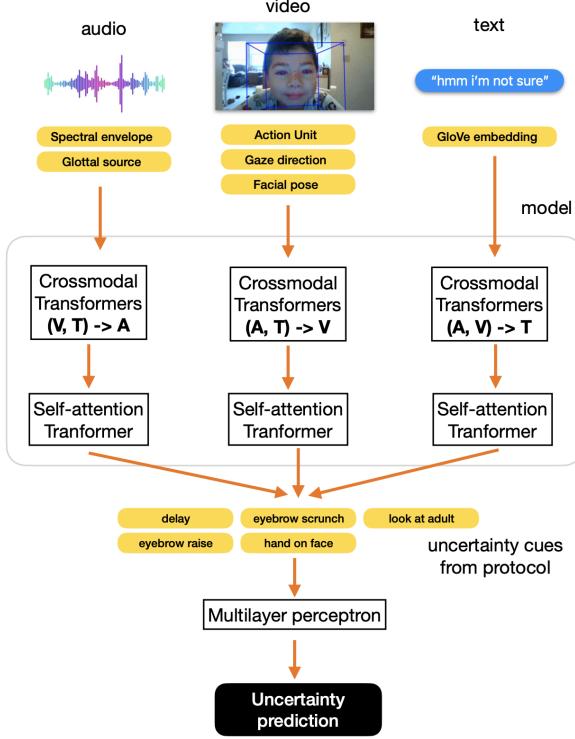


Figure 6: This figure shows the architecture of the ensemble learning model. First, cross-modal transformers learn the attention across the features of each modality with each of the other two modalities' low-level features. Then, using the fused features, self-attention transformers predict the present uncertainty cues, which are then passed into a multilayer perceptron to output the final prediction of whether or not uncertainty is present.

a three-class softmax classification. The second is a detection model trained with adult data that takes visual information (Jahoda et al., 2018). This second baseline is a traditional machine-learning approach using SVMs and LBP descriptors.

**Metrics** In our experiment, we utilized three key metrics to evaluate the performance of our model. The weighted F1 score was employed to account for any class imbalance and provide a more comprehensive assessment of the model's precision and recall. Mean Absolute Error (MAE) was used to measure the average magnitude of the errors in our predictions, illustrating the model's ability to minimize deviations from the true values. Lastly, the R-square statistic was employed to quantify the proportion of variance in the dependent variable explained by the model, offering insight into the overall goodness-of-fit and the model's explanatory power.

	Model	F1	MAE	R <sup>2</sup>
No Cue	Basic	.7011	.3603	-.2419
	Adult	.6397	.8553	-2.7119
	MulT	.8120	.2307	-.1391
	CL + W	.8216	.3076	-.5412
Cue	Ensemble	<b>.8366</b>	<b>.2222</b>	-.1250

Table 2: This table shows the results of different models. The "Basic" refers to the MLP baseline, and the "Adult" refers to the adult uncertainty baseline. "MulT" refers to the Multimodal Transformer, and "CL + W" is the MulT model with contrastive learning and weighted sampling. "Ensemble" refers to the cue-based ensemble model. No Cue and Cue indicate whether the model uses the identified cues as intermediate features.

## 5.1 Results

We report a weighted F1 score of .8216 and a mean absolute error of .3076 on the cue-agnostic end-to-end model with reweighted class labels, as seen in Table 2. Full results for each modality can also be found in the same table. The cue-aware ensemble model shows improvements in both weighted F1 and MAE over the multimodal transformer model. Contrastive learning and weighted sampling improve the performance but are subpar compared to the cue-based ensemble method. The intermediate prediction of cues like *delay* that have been shown to be a key indicator but may be difficult to learn in an end-to-end model may play a role in this performance.

**Modality Ablations** After doing an ablation study on the modalities for the cue-agnostic models, we find that the text and audio modality report the best scores overall, as shown in 3. This is unexpected due to participant speech being scarce. However, when participants do talk, it is usually to express their feeling about the task. For instance, participants may either say "This is easy!" or "I don't know," tell the adult if the trial is hard or begin counting. As a result, the text modality could be less noisy than the video and audio modalities. We note that the particular task does not request verbal responses from the participants. Thus, we expect that with a task that entices a verbal response, such as question answering, there may be more contribution from the text and audio modalities.

Model	F1	MAE	$R^2$
CL + W	<b>.8216</b>	.3076	-.5412
Video only	.7991	.2820	-.4072
Text only	<b>.8056</b>	<b>.2564</b>	-.2731
Audio only	<b>.8056</b>	<b>.2564</b>	-.2731

Table 3: F1, MAE, and  $R^2$  results for the best performing cue-agnostic model (weighted) with the ablation studies for all the modalities. The model performs the best with all the components, but the most influential modality is text/audio.

## 6 Conventions of Uncertainty: A Case Study in Different Age Groups

It is observed that for older children ( $> 2150$  days old), less parental guidance is present, faster decision-making is observed, less diverse facial expressions are present, and more verbal cues are present, which increases the performance of the models for 5-year-olds Table 4. In addition, certain behavior patterns that children of different age groups display are context-dependent, convention-oriented, and personality-specific which makes it difficult to identify only through visual and textual modalities. Some of these behaviors that we investigate here are nail-biting, pointing, and social facilitation (see Figure 7).

Model	4 year old			5 year old		
	F1	MAE	$R^2$	F1	MAE	$R^2$
Basic	.69	.32	-.24	.73	.33	-.21
Adult	.57	<b>1.06</b>	<b>-4.40</b>	.62	1.0	<b>-8.13</b>
MuLT	.78	.26	-.39	.81	.26	-.27
CL + W	<b>.79</b>	<b>.22</b>	<b>-.15</b>	<b>.81</b>	<b>.23</b>	<b>-.14</b>

Table 4: This table shows the results of the models between different age groups. There are slight inference differences between the 4 and 5-year-old groups. These performance changes are dependent on the conventions of uncertainty and age.

In the developmental psychology literature, nail-biting is either found to be an acquired habit or related to states of nervousness (Gilleard et al., 1988; Silber and Haynes, 1992; Ghanizadeh, 2008; McClanahan, 1995; Wells et al., 1998). This behavior is hard to classify as stress-induced or uncertainty-induced. Hence, a context-dependent analysis of the person using skeletal features can be helpful in deciding.

Pointing (see Figure 7) is another context-

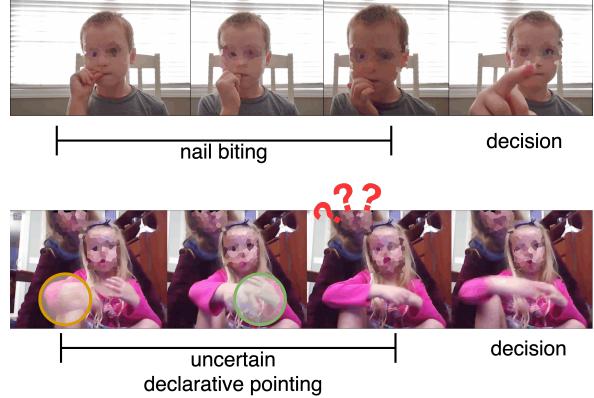


Figure 7: This figure shows two complex behavior patterns by children: nail-biting and uncertain declarative pointing. The top sequence belongs to the oldest male, and the bottom sequence belongs to the youngest female.

dependent occurrence. When the child is uncertain, the pointing to options also becomes ambiguous, and the parent needs to ask a follow-up grounding clarification question, such as "Which one?". This behavior pattern involves spatial placement of options and understanding the boundaries between them. Younger children prefer ambiguous pointing gestures in comparison to conventional and visible cues of uncertainty. This type of declarative pointing is observed to be a way of engaging with the parent, pointing to a theory of mind (ToM) understanding by the children (Cochet et al., 2017). Skeletal and ToM modeling can be helpful in making prediction performance better.

Another behavior pattern is social facilitation. Children that are younger prefer to be together with their parents while solving tasks. Children that are older follow verbal conventions and reduce the vividness of their facial expressions, while younger children exaggerate their facial expressions and rely more on social facilitation factors. Similar behavior patterns happen in adults when they are in a competitive atmosphere where social facilitation has different effects on an individual's facial expressions (Buck et al., 1992; Katembu et al., 2022). ToM and multi-party dialogue modeling can increase the performance of uncertainty understanding models.

Uncertainty is context-dependent – some children are naturally more fidgety, or shy. So predicting uncertainty on an isolated trial basis may lead to less accurate results. As a result, one interesting question is how to incorporate contextual features about the participant's personality and recent cog-

nitive states to make more informed predictions.

## 7 Conclusion

In this paper, we explored the task of predicting uncertainty from our annotated dataset with a multimodal transformer-based model. We first validated our protocol and identified key cues such as delay and eyebrow movement as promising indicators of uncertainty. We also discover that demographic and trial difficulty can affect the frequency of certain cues. Moreover, trial difficulty strongly correlates with uncertainty, but trial performance interestingly does not. The emergence of this dataset and protocol can encourage further exploration of the task of identifying uncertainty for both better human-human and human-AI dialogue.

## 8 Ethics

Due to the sensitive nature of the video data of children and their privacy, we are only making some portion of the data publicly available with the consent of the parents of the children. All the images used in this paper are from the videos that are from children and parents who have given consent to share their video data publicly.

## References

- Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. *Openface 2.0: Facial behavior analysis toolkit*. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66.
- Lisa Beinborn, Teresa Botschen, and Iryna Gurevych. 2018. *Multimodal grounding for language processing*. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2325–2339, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Nathaniel J. Blanco and Vladimir M. Sloutsky. 2021. *Systematic exploration and uncertainty dominate young children’s choices*. *Developmental Science*, 24(2):e13026.
- Ross Buck, Jeffrey I Losow, Mark M Murphy, and Paul Costanzo. 1992. Social facilitation and inhibition of emotional expression and communication. *Journal of Personality and Social Psychology*, 63(6):962.
- Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. *Towards multimodal sarcasm detection (an \_Obviously\_ perfect paper)*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy. Association for Computational Linguistics.
- Herbert H Clark, Robert Schreuder, and Samuel Buttrick. 1983. Common ground at the understanding of demonstrative reference. *Journal of verbal learning and verbal behavior*, 22(2):245–258.
- Hélène Cochet, Marianne Jover, Cécile Rizzo, and Jacques Vauclair. 2017. Relationships between declarative pointing and theory of mind abilities in 3-to 4-year-olds. *European Journal of Developmental Psychology*, 14:324–336.
- Gilles Degottex. 2014. Covarep – a collaborative voice analysis repository for speech technologies.
- Stanislas Dehaene. 2011. *The number sense: How the mind creates mathematics*. Oxford University press.
- Jeroen Dral, Dirk Heylen, and Rieks Akker. 2011. *Detecting Uncertainty in Spoken Dialogues: An Exploratory Research for the Automatic Detection of Speaker Uncertainty by Using Prosodic Markers*, pages 67–77.
- Ahmad Ghanizadeh. 2008. Association of nail biting and psychiatric disorders in children and their parents in a psychiatrically referred sample of children. *Child and adolescent psychiatry and mental health*, 2:1–7.
- Esen Gilleard, Mehmet Eskin, and Buğda Savaşır. 1988. Nailbiting and oral aggression in a turkish student population. *British journal of medical psychology*, 61(2):197–201.
- Paul L. Harris, Deborah T. Bartz, and Meredith L. Rowe. 2017. Young children communicate their ignorance and ask questions. *Proceedings of the National Academy of Sciences*, 114(30):7884–7891.
- Pavel Jahoda, Antonin Vobecky, Jan Cech, and Jiri Matas. 2018. Detecting decision ambiguity from facial images. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 499–503.
- Stephen Katembu, Qiang Xu, Hadiseh Nowparast Rostami, Guillermo Recio, and Werner Sommer. 2022. Effects of social context on deliberate facial expressions: Evidence from a stroop-like task. *Journal of Nonverbal Behavior*, 46(3):247–267.
- Terry Michael McClanahan. 1995. Operant learning (rs) principles applied to nail-biting. *Psychological reports*, 77(2):507–514.
- Michelle Morales, Stefan Scherer, and Rivka Levitan. 2018. A linguistically-informed fusion approach for multimodal depression detection. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 13–24, New Orleans, LA. Association for Computational Linguistics.

- Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. **Towards multimodal sentiment analysis: Harvesting opinions from the web**. In *Proceedings of the 13th International Conference on Multimodal Interfaces, ICMI ’11*, page 169–176, New York, NY, USA. Association for Computing Machinery.
- Yondu Mori and Marc D Pell. 2019. The look of (un) confidence: visual markers for inferring speaker confidence in speech. *Frontiers in Communication*, 4:63.
- Darko Odic, Howard Hock, and Justin Halberda. 2014. Hysteresis affects approximate number discrimination in young children. *Journal of Experimental Psychology: General*, 143(1):255.
- Darko Odic and Ariel Starr. 2018. **An introduction to the approximate number system**. *Child Development Perspectives*, 12(4):223–229.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. **GloVe: Global vectors for word representation**. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Heather Pon-Barry and Stuart Shieber. 2011. **Recognizing uncertainty in speech**. *EURASIP journal on advances in signal processing*, 2011.
- Pio E. Ricci Bitti, Luisa Bonfiglioli, Paolo Melani, Roberto Caterina, and Pierluigi Garotti. 2014. **Expression and communication of doubt/uncertainty through facial expression**. *Ricerche di Pedagogia e Didattica. Journal of Theories and Research in Education*, 9(1):159–177.
- Pedro Bispo Santos, Lisa Beinborn, and Iryna Gurevych. 2016. **A domain-agnostic approach for opinion prediction on speech**. In *Proceedings of the Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 163–172, Osaka, Japan. The COLING 2016 Organizing Committee.
- Kimberly Scott and Laura Schulz. 2017. **Lookit (Part 1): A New Online Platform for Developmental Research**. *Open Mind*, 1(1):4–14.
- Kevin P Silber and Clare E Haynes. 1992. Treating nail-biting: a comparative analysis of mild aversion and competing response therapies. *Behaviour research and therapy*, 30(1):15–22.
- Marc Swerts and Emiel Krahmer. 2005. Audiovisual prosody and feeling of knowing. *Journal of Memory and Language*, 53(1):81–94.
- Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. 2001. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115.
- Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Florence, Italy. Association for Computational Linguistics.
- Jennifer H Wells, Janet Haines, and Christopher L Williams. 1998. Severe morbid onychophagia: the classification as self-mutilation and a proposed model of maintenance. *Australian and New Zealand journal of psychiatry*, 32(4):534–545.
- Xiaocui Yang, Shi Feng, Yifei Zhang, and Daling Wang. 2021. **Multimodal sentiment detection based on multi-channel graph neural networks**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 328–339, Online. Association for Computational Linguistics.

## A Percentages of each cue

Here we present in Table 5, all the distribution of the uncertainty cues we found in the dataset. We also present more statistics between female and male participants in Figure 8.

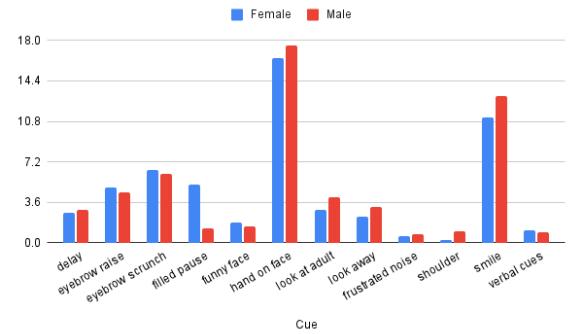


Figure 8: Female and male participants show small differences in frequency of displaying certain cues.

## B Contrastive Learning Details

### B.1 Problem Statement

In this study, we aim to predict the uncertainty of a child based on multimodal inputs, including video, transcripts, and audio. Given the dataset of instances, each containing video (V), transcripts (T), and audio (A) data, our goal is to develop a model that can accurately predict whether a child is uncertain, unclear, or not uncertain. We represent this problem as a function F that maps the input features (V, T, A) to the binary output variable

Cue	All Pct.	Uncertain Pct.	Hard Pct.	Easy Pct.	Female Pct.	Male Pct.
delay	0.03	<b>0.17</b>	0.03	0.02	2.7	2.93
eyebrow raise	0.05	<b>0.17</b>	<b>0.06</b>	0.04	4.94	4.48
eyebrow scrunch	0.06	<b>0.22</b>	0.07	0.06	6.49	6.15
filled pause	0.03	0.06	0.03	0.03	<b>5.17</b>	1.26
funny face	0.02	0.07	<b>0.03</b>	0.01	1.84	1.49
hand on face	<b>0.17</b>	<b>0.19</b>	<b>0.16</b>	<b>0.18</b>	16.44	17.53
look at adult	0.04	0.1	<b>0.05</b>	0.03	2.93	4.02
look away	0.03	0.05	0.03	0.02	2.36	3.16
frustrated noise	0.01	0.02	0.01	0.01	0.57	0.8
shoulder	0.01	0.02	0.01	0.01	0.29	<b>1.03</b>
smile	<b>0.12</b>	<b>0.17</b>	<b>0.12</b>	<b>0.12</b>	11.15	<b>13.05</b>
verbal cues	0.01	0.02	0.01	0.01	1.09	0.92

Table 5: Distribution of uncertainty cues across all/uncertain, difficult/easy, and female/male trials. We can see that *delay*, *eyebrow raise*, and *eyebrow scrunch* are significantly more frequent in uncertain trials. Meanwhile, if the trial is hard, participants are more likely to look at an adult or make a funny face. If the trial is easy, the participant may display *hand on face* instead. Female and male participants also show mild differences in the frequency of displaying certain cues.

$y \in \{0, 0.5, 1\}$ , where 0 denotes not uncertain, 0.5 denotes unclear, and 1 denotes uncertain:

To achieve this, we design a multimodal transformer model that leverages the complementary information available in the video, transcripts, and audio data to make predictions. The model is trained on a dataset of labeled examples  $(V_i, T_i, A_i, y_i)$ , where  $i \in \{1, \dots, N\}$  and  $N$  is the total number of instances. Our objective is to minimize the cross-entropy loss. By minimizing this loss, our model will learn to make accurate predictions of a child’s uncertainty level based on the provided multimodal inputs.

The specific contrastive learning loss function,  $L(X_1, X_2, L_1, L_2)$ , that we are focusing on here is defined as the following:

$$L(X, L) = \frac{1}{N_1 * N_2} \sum_i \sum_j W_{ij} * (m - S_{ij})_+^2 \quad (1)$$

This function captures the relationship between pairs of data points  $X_1$  and  $X_2$ , with associated labels  $L_1$  and  $L_2$ . The cosine similarity,  $S_{ij}$ , is used to measure the similarity between the data points, and the weighing factor,  $W_{ij}$ , is used to differentiate between positive and negative pairs. The weighting factor is determined using the Kronecker delta function,  $\delta(L_{1i}, L_{2j})$ . The  $e^{\alpha * \delta(L_{1i}, L_{2j})}$  coefficient ensures that the positive pairs have a higher influence on the learning process where  $m$  is a threshold to separate positive and negative pairs

and  $\alpha$ , is a scaling factor. Lastly,  $(x)_+$  ensures that only the non-negative values of  $x$  are considered.

To further improve our model’s performance, we employed a weighted sampling method that uses the inverse square root of the class frequencies. Given a dataset with classes 0 being certain, 0.5 being unclear, and 1 being uncertain, we calculate the weights for each class sample as follows:

$$w_i = \frac{1}{\sqrt{N_i}}, \quad \text{where } i \in \{0, 0.5, 1\}. \quad (2)$$

This weighting scheme assigns higher weights to underrepresented classes, the class of 0.5 (unclear) and 1 (uncertain) in our case, which helps balance class sampling probabilities. The inverse square root function is particularly useful as it provides a smooth topology that is less sensitive to small changes in class frequencies compared to other weighing functions.

## C Implementation Details

### C.1 Experimental setup

The argmax of the label probabilities was taken as the output layer. All networks were trained for 40 epochs with a batch size of 24. Both raw and weighted cross entropy loss were used to train two versions of the model. Class weights were set based on the distribution of train set samples to mitigate class imbalance issues.

We employ a 75-10-15 training-dev-test split. For each result, we report the average across 3

different seeds. We also run the model on every single modality to provide unimodal baselines.

Additionally, we have age information for each participant, so we divided the dataset into two age groups: 4-year-olds and 5-year-olds. This division will allow us to investigate potential differences between the two age groups as shown in the 4-year-old and 5-year-old tabs in Table 2.

**Transformer Model Details.** Our multimodal transformer model is based on a modified version of the Transformer architecture. It consists of five layers, each equipped with five attention heads to capture various contextual relationships within the input data. We used the Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of 0.001 to train our model. To enhance convergence and overall performance, we employed the ReduceLROnPlateau learning rate scheduler, which adjusts the learning rate when the validation loss ceases to improve. We set the reduction factor to 0.1 and patience of 5 epochs for monitoring improvements. Our model was trained on an NVIDIA RTX 4090 GPU, using a batch size of 1. We trained the model for 100 epochs. For models with contrastive learning, we trained the model with additional 10 epochs before real training.