# A Face-to-Face Neural Conversation Model

Hang Chu[1,2] Daiqing Li[1] Sanja Fidler[1,2]
[1]University of Toronto [2]Vector Institute
{chuhang1122, daiqing, fidler}@cs.toronto.edu

## Abstract

*Neural networks have recently become good at engaging in dialog. However, current approaches are based solely on verbal text, lacking the richness of a real face-to-face conversation. We propose a neural conversation model that aims to read and generate facial gestures alongside with text. This allows our model to adapt its response based on the "mood" of the conversation. In particular, we introduce an RNN encoder-decoder that exploits the movement of facial muscles, as well as the verbal conversation. The decoder consists of two layers, where the lower layer aims at generating the verbal response and coarse facial expressions, while the second layer fills in the subtle gestures, making the generated output more smooth and natural. We train our neural network by having it "watch" 250 movies. We showcase our joint face-text model in generating more natural conversations through automatic metrics and a human study. We demonstrate an example application with a face-to-face chatting avatar.*

## 1. Introduction

We make conversations everyday. We talk to our family, friends, colleagues, and sometimes we also chat with robots. Several online services employ robot agents to direct customers to the service they are looking for. Question-answering systems like Apple Siri and Amazon Alexa have also gained many recent attentions. However, while most of these automatic systems feature a human voice, they are far from acting like human beings. They lack in expressivity, and are typically emotionless.

Language alone can often be ambiguous with respect to the person's mood, unless indicative sentiment words are being used. In real life, people make gestures and read other people's gestures when they communicate. Whether someone is smiling, crying, shouting, or frowning when saying "thank you" can indicate various feelings from gratitude to irony. People also form their response depending on this context, not only in what they say but also in how they say it. We aim at developing a more natural conversation model that jointly models text and gestures, in order to act and converse in a more natural way.

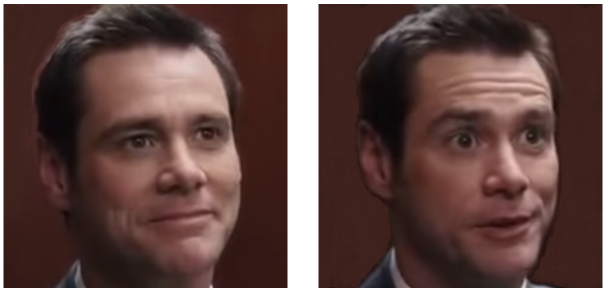


Figure 1: Facial gestures convey sentiment information. Words have different meanings with different facial gestures. Saying *he's a nice person* with different gestures could be either confirmative, or ironic. Therefore, different responses should be triggered.

Recently, neural networks have been shown to be good conversationalists [31, 15], which typically makes use of an RNN encoder which represents the history of the verbal conversation and an RNN decoder that generates a response. [16] furthered this idea, trying to personalize the model by adapting conversations to a particular user. However, all these approaches are based solely on text, lacking the richness of a real face-to-face conversation.

In this paper, we introduce a neural conversation model that reads and generates both a verbal response (text) and facial gestures. We exploit movies as a rich resource of such information. Movies show a variety of social situations with diverse emotions, reactions, and topics of conversation, making them well suited for our task. Movies are also multi-modal, allowing us to exploit both visual as well as dialogue information. However, the data itself is also extremely challenging due to many characters that appear on-screen at any given time, as well as large variance in pose, scale, and recording style.

Our model adopts the encoder-decoder architecture and adds gesture information in both the encoder as well as the decoder. We exploit the FACS representation [8] of ges-

Demo&data: http://www.cs.toronto.edu/face2face

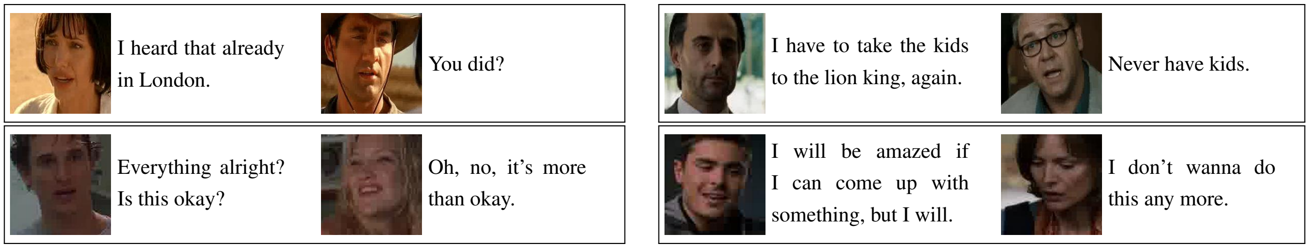

I heard that already in London.
You did?
I have to take the kids to the lion king, again.
Never have kids.
Everything alright? Is this okay?
Oh, no, it's more than okay.
I will be amazed if I can come up with something, but I will.
I don't wanna do this any more.

Figure 2: Example conversations from our MovieChat dataset. Each row shows two examples, left shows query face and text, right shows target face and text. Our dataset has various conversation scenarios, such as simple conversations shown in the first and second rows on the left, as well as more challenging cases shown on the right.

tures, which allows us to effectively encode and synthesize facial gestures. Our decoder is composed of two levels, one generating the verbal response as well as coarse gesture information, and another level that fills in the details, making the generated expressions more natural. We train our model using reinforcement learning that exploits a trained discriminator to provide the reward. We show that our model generates more appropriate responses compared to multiple strong baselines, on a large scale movie dataset. We further showcase NeuralHank, an expression-enabled 3D chatting avatar driven by our proposed model.

The rest of the paper is organized as follows. Sec. 2 reviews the related work. In Sec. 3 we introduce our dataset to facilitate face-to-face conversation modeling. In Sec. 4 we describe our approach. Sec. 5 provides extensive evaluation and introduces our chit-chatting avatar.

## 2. Related Work

Dialogue systems have been explored since the 60', with systems like ELIZA [32] and PARRY [5] already capable of engaging in relatively complex conversations. These approaches have mainly been based on hand-coded rules, thus were not able to adapt to users and topics, and usually seemed unnatural. In [19], the authors formulated the problem as statistical machine translation, where the goal was to "translate" the query posts in blogs into a response. This problem setting is typically harder than traditional translation from one language to another, since the space of possible responses is more diverse.

Conversation modeling has recently been gaining interest due to the powerful language models learned by neural networks [31, 15, 16]. [31] was the first to propose a neural conversation model, which exploited the encoder-decoder architecture. An LSTM encoder was used to represent the query sentence while the decoder LSTM generated a response, one word at a time. The model was trained on a large corpus of movie subtitles, by using each sentence as a query and the following sentence as a target during training. Qualitative results showed that meaningful responses were formed for a variety of queries. In parallel, the Skip-Thought model [12, 35] adopted a similar architecture, and was demonstrated to be effective in a variety of NLP tasks as well as image-based story-telling.

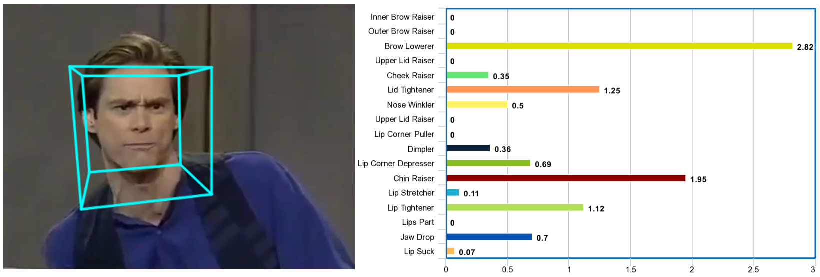
Since neural conversation models typically produce short and more generic sentences, the authors in [15] proposed an improved objective function that encouraged diversity in the generator. In [21], the authors exploited a hierarchical encoder-decoder, where one GRU layer was used to model the history of the conversation at the sentence level, and the second level GRU was responsible for modeling each sentence at the word level. This model was extended in [23] by adding latent variables aiming to capture different topics of conversation, allowing the model to achieve a higher diversity in its response.

An interesting extension was proposed in [16] which aimed at personalizing conversations. The model learned a separate embedding for each person conversationalist, jointly with dialogue. The purpose of the embedding was to bias the decoder when generating the response. This allowed for a more natural human-like chit-chat, where the model was able to adapt to the person it was speaking to.

Most of these works are based solely on language. However, humans often use body gestures as an additional means to convey information in a conversation. An interesting approach was proposed in [14, 13] which aimed at synthesizing body language animations conditioned on speech using a HMM. This approach required motion capture data recorded during several conversation sessions.

Face capture has been a long-studied problem in computer vision, with many sophisticated methods such as [2, 24, 10]. The FirstImpression dataset [3] was collected to facilitate the need of data in gesture recognition. Face synthesis has been widely studied in both vision and graphics communities. [27] proposed a reconstruction algorithm that captures a person's physical appearance and persona behavior. [30] transfers facial gesture from a source video to a target video to achieve realistic reenactment. [28] further transformed audio speech signal into a talking avatar using an RNN-based model.

In our approach, we aim to both encode and generate facial gestures jointly with language, by exploiting a large corpora of movies. Movies feature diverse conversations and interactions, and allow us to use both visual as well as dialogue information.



(a) (b)

| query | target | # of examples |
|---|---|---|
| text | text | $40,200,261$ |
| text+face | text | $48,475$ |
| text | text+face | $48,475$ |
| text+face | text+face | $24,727$ |

(c)

Figure 3: Overview of our MovieChat database. (a) and (b) show an example frame with 3D face detection and detected FACS intensities. We obtain detections using the off-the-shelf OpenFace [2] package. (c) shows the scale of our MovieChat database. Our database is by far the largest language-face conversation video dataset.

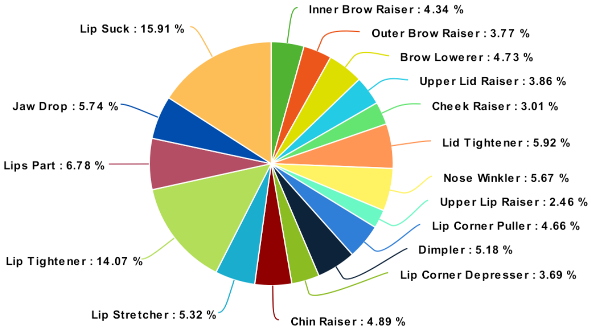


Figure 4: List of gestures recorded in the MovieChat dataset, and percentage of frames where each gesture is dominant.

## 3. The MovieChat Dataset

Datasets of considerable size are key to successfully training neural networks. In our work, we seek a dataset containing people engaging in diverse conversations, that contains both video as well as transcribed dialogues.

Towards this goal, we build the MovieChat dataset. We take advantage of the large movie collection of MovieQA [29], which contains clips from 250 movies, covering more than half of each movie in duration. To track 3D faces and detect facial gestures, we use the off-the-shelf OpenFace [2] package. Tracking and detection runs in real-time while maintaining good accuracy. This makes processing of such a large volume of video data possible.

However, even the best automatic face detector occasionally fails. Certain recording styles, such as the shaky and free-cam clips, make our processing more challenging. To address these problems and improve the quality of our dataset, we further divide all movies into short, single sentence clips by exploiting the time stamps stored in their subtitles file. We only keep clips where a single face is detected across all of its frames, and discard the rest of the clips. This is to avoid ambiguous dialog-face association when multiple characters appear in a single shot. Additionally, we remove fast-cut clips where the speaker's face is not fully visible throughout the clip. Finally, we also remove clips in which tracks are extremely shaky, which often suggests tracking failure. We observe significant quality improvement after these filtering steps, with only rare failure cases.

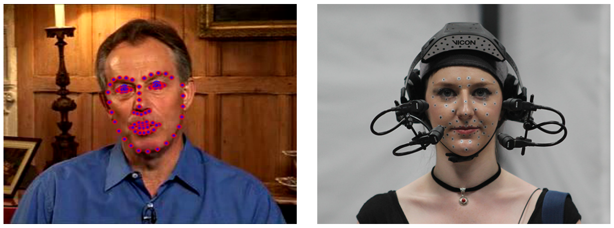


Figure 5: Facial Landmark (FL) systems. Left shows an example of "in the wild" landmarks [6, 34, 2], which fails to capture subtle gesture information. Right shows invasive motion capture landmarks [1].

We build our final dataset with the remaining clips. We record image frames, time stamps, 3D face poses, facial gestures, and transcribed dialogues. Fig. 3 shows an example, and provides statistics summarizing our dataset. Fig. 4 shows the recorded gestures and their statistics in our MovieChat data.

## 4. Face-to-Face Neural Conversation Model

We first explain our facial gestures representation using Facial Action Coding System (FACS) [8]. We then describe our proposed model.

### 4.1. FACS Gesture Representation

Various approaches are available for representing gesture numerically, e.g. Six Universal Expressions (SUE) [4], Facial Landmarks (FL) [6, 34], and FACS [8].

SUE [4] categorizes gesture into six emotions: anger, disgust, fear, happiness, sadness and surprise. It is effective in encoding high-level emotion, but it is overly abstract to describe detailed gestures. Each emotion involves a combination of up to 6 muscle movements, making it difficult for face synthesis and animation.

FL [6, 34] represents gesture using landmark points. Typically, 68 points are used to track corner-edge keypoint positions of the face. Compared to SUE, FL carries more details. However, FL has two disadvantanges. First, FL does not contain complete gesture information. The cheek

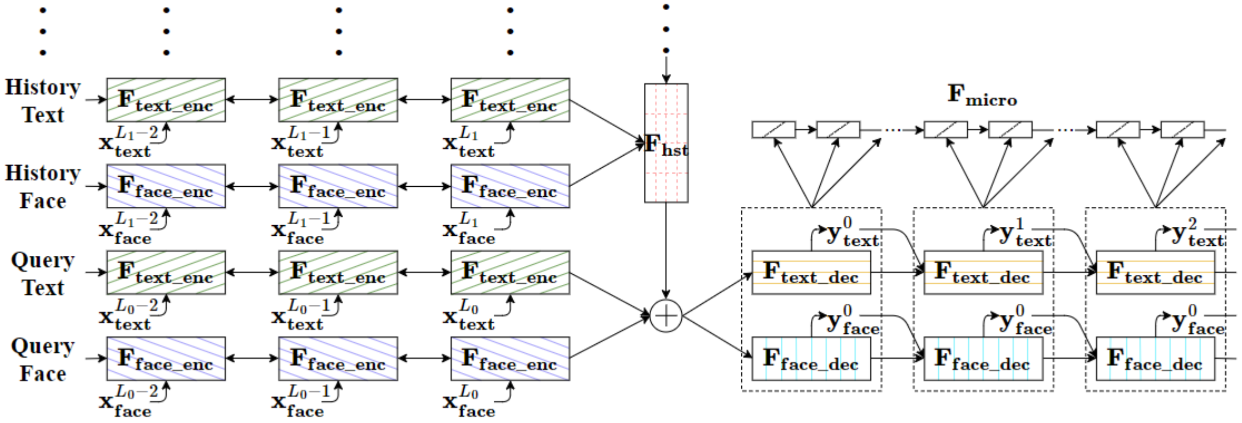


Figure 6: Our face-to-face conversation model. Our model consists of 6 RNNs shown in different colors. First, $\mathbf{F_{text\_enc}}$ and $\mathbf{F_{face\_enc}}$ encode $N$ previous text-face sentences in the chat history (only one history sentence is depicted). History sentence encodings are further encoded by $\mathbf{F_{hst}}$. Next, query and history encodings are added to form the context vector. $\mathbf{F_{text\_dec}}$ and $\mathbf{F_{face\_dec}}$ generate response sentences conditioning on the context vector. Finally, $\mathbf{F_{micro}}$ generates frame-level animation controls based on the word-level decodings.

and forehead regions, which are texture-less but contain many muscles, are usually missing. Second, FL is anatomically redundant. 5 landmarks are used to outline one brow, while its underlying motion is lower dimensional that involves 2 muscle intensity values. Therefore, FL is less desirable for our task. It should be noted there are variations of FL that places landmarks across all muscles evenly. They are widely used in motion capture, e.g. Cara [1] in Figure 5. This FL system requires visible marks on the character face, thus making large scale data collection difficult.

We adopt FACS [8] in this paper. Particularly, we use 18 action unit each controls a face muscle, as well as 3 dimensions to represent the 3D head pose. Compared to the SUE and FL, FACS not only captures subtle detail gestures, but also produces highly interpretable gesture representation which makes animation simple and straight-forward. We detect FACS from images using the off-the-shelf OpenFace [2].

## 4.2. Face-to-Face Conversation Model

Following previous work on conversation modeling [23, 26], we adopt the RNN encoder-decoder architecture, but adapt it to our face-to-face conversation task. Our proposed model consists of 6 RNN modules that capture and generate information across different modalities and resolutions. Fig. 6 provides an overview of the model. Overall, our model is an encoder-decoder framework that is trained with RL and GAN.

**Notation.** Our algorithm takes a series of paired text and gesture sequences as input. These represent the query sequence as well as a recent conversation history of $N$ sequences. Let $\{L_0, L_1, ...L_N\}$ denote the word lengths of each sentence, where $L_0$ is the word length of query (source), $L_1$ and $L_N$ are word lengths for the most recent and earliest sequences in the history. The reverse-order $\ell$-th word in the $n$-th history sentence ($n$=0 for query sentence) is therefore a pair of $(\mathbf{x}^{L_n-\ell}_{\textbf{text}}, \mathbf{x}^{L_n-\ell}_{\textbf{face}})$, where $n \in [0, N]$ and $\ell \in [0, L_n]$. Similarly, we denote the ground truth answer (target) sequence as $(\hat{\mathbf{y}}^{L_*-\ell}_{\textbf{text}}, \hat{\mathbf{y}}^{L_*-\ell}_{\textbf{face}})$, where $\ell \in [0, L_*]$, $L_*$ being the answer word length.

**Word-level encoders.** We synchronize text and gesture at word level. $\mathbf{x}^{L_n-\ell}_{\textbf{text}}$ is naturally an one-hot encoding of the current word. To keep the representation consistency and simplify the multi-dimensional gesture data, we set $\mathbf{x}^{L_n-\ell}_{\textbf{face}}$ as a similar one-hot encoding of the closest gesture template. Gesture templates are obtained from k-means ($k = 200$) clustering all gestures in the training set. We define our word-level encoders as bidirectional RNNs, i.e.

$$\begin{aligned}\mathbf{h}^{L_n-\ell}_{\textbf{text}} &= \mathbf{F_{text\_enc}}(\mathbf{x}^{L_n-\ell}_{\textbf{text}} \mid \mathbf{h}^{L_n-\ell-1}_{\textbf{text}}, \mathbf{h}^{L_n-\ell+1}_{\textbf{text}}) \\ \mathbf{h}^{L_n-\ell}_{\textbf{face}} &= \mathbf{F_{face\_enc}}(\mathbf{x}^{L_n-\ell}_{\textbf{face}} \mid \mathbf{h}^{L_n-\ell-1}_{\textbf{face}}, \mathbf{h}^{L_n-\ell+1}_{\textbf{face}})\end{aligned} \tag{1}$$

where $\mathbf{h}$ denotes the hidden encoding with zero padding at sequence boundaries, and $\mathbf{F}$ denotes the RNN cell function.

**Sentence-level encoder.** The sentence-level encoder takes a sequence of history (excluding query) sentence encodings, and summarizes the conversation history into a single encoding vector. We model the history with a sentence-level bidirectional RNN on both text and gesture, i.e.

$$\mathbf{h}^{n}_{\textbf{hst}} = \mathbf{F_{hst}}(\mathbf{h}^{L_n}_{\textbf{text}} \oplus \mathbf{h}^{L_n}_{\textbf{face}} \mid \mathbf{h}^{n-1}_{\textbf{hst}}, \mathbf{h}^{n+1}_{\textbf{hst}}) \tag{2}$$

where $n \in [1, N]$, $\oplus$ denotes vector concatenation, and boundaries are zero-padded.

**Word-level decoders.** We use two decoders that generate the target sentence in both text and gesture. The output follows the same one-hot encoding as source and history sentences. We condition the target sentence generation on the joint text-face-history context vector, which is obtained by the summation of source text encoding, source face encoding, and history encoding. Concretely,

$$\mathbf{h} = \mathbf{h}_{\mathbf{text}}^{L_0} + \mathbf{h}_{\mathbf{face}}^{L_0} + \mathbf{h}_{\mathbf{hst}}^{1} \tag{3}$$

where $\mathbf{h}$ is the final encoding that we condition our generation decoders on. We use two single-directional RNN decoders, i.e.

$$\begin{aligned} \mathbf{h}_{\mathbf{text\_dec}}^{\ell} &= \mathbf{F}_{\mathbf{text\_dec}}(\mathbf{y}_{\mathbf{text}}^{\ell-1} \mid \mathbf{h}, \mathbf{h}_{\mathbf{text\_dec}}^{\ell-1}) \\ \mathbf{y}_{\mathbf{text}}^{\ell} &= \operatorname*{argmax}_{\mathbf{y}} P(\mathbf{y} \mid \mathbf{h}_{\mathbf{text\_dec}}^{\ell}) \end{aligned} \tag{4}$$

where we pad the beginning with the `begin_of_sent` token and terminates the decoding at the `end_of_sent` token. The face decoder follows the same definition.

**Micro-gesture generator.** The output of our gesture decoder is a highly summarized gesture template. We observe that the although templates are sufficient for representing semantics, it is insufficient for synthesizing vivid, high framerate animations. We use the micro-gesture module to fill in this resolution gap. It is defined as a frame-level RNN. At the $t$-th frame, we have

$$\mathbf{h}_{\mathbf{micro}}^{t} = \mathbf{F}_{\mathbf{micro}}(\mathbf{y}_{\mathbf{text}}^{t} \oplus \mathbf{y}_{\mathbf{face}}^{t} \mid \mathbf{h}_{\mathbf{micro}}^{t-1}) \tag{5}$$

with zero boundary padding. Finally, we obtain the frame-level gesture by linearly regressing $\mathbf{h}_{\mathbf{micro}}^{t}$ to each individual gesture dimension. These gesture values directly controls muscle intensities that drive the 3D avatar.

**Policy gradient.** The default cross-entropy training suffers from exposure bias, as the model is only exposed to ground truth samples during training. For our decoder networks, we alleviate this problem using policy gradient training. In this setting, *policy* takes form of a decoder RNN, and an *action* is a complete sentence sampled from the policy denoted by $\tilde{\mathbf{y}}_{\mathbf{text}}$. Our goal is to achieve higher *reward* under the metric of choice evaluated at the end of sentence, conditioned on the context. This is denoted by $\mathbf{R}(\tilde{\mathbf{y}}_{\mathbf{text}}|\mathbf{h})$. The text policy gradient is given by (same for face)

$$\nabla J_{\mathbf{text}} = [\mathbf{R}(\tilde{\mathbf{y}}_{\mathbf{text}}|\mathbf{h}) - \mathbf{b}]\nabla \log \mathbf{F}_{\mathbf{text\_dec}}(\tilde{\mathbf{y}}_{\mathbf{text}}|\mathbf{h}) \tag{6}$$

where $J$ is the objective function, $\mathbf{b}$ is the baseline evaluated on greedy sequence sampling conditioned on the same $\mathbf{h}$, and $\tilde{\mathbf{y}}_{\mathbf{text}}$ are monte-carlo sampled from the generator.

**Adversarial discriminator.** Conversation models suffers from dull responses [17], while diverse dialogues are preferred in practical scenarios. We address this problem by using a sequential GAN setup [33]. The *generator* is our decoder network, while the *discriminator* is another network that distingtuishes whether the resulting sequence is generated. We can formulate this as a minmax problem, i.e.

$$\min_{\theta} \max_{\eta} J_{\mathbf{gan}}(\mathbf{F}^{\theta}, \mathbf{D}^{\eta} \mid \mathbf{h}) \tag{7}$$

where $\mathbf{F}$ is either the text or face generator, $\mathbf{D}$ being its corresponding discriminator, $\theta$ and $\eta$ being parameters of the generator and discriminator. Specifically, the discriminator objective $J_{\mathbf{gan}}$ is defined as

$$\mathbb{E}_{\mathbf{y}\sim\hat{\mathbf{y}}_{\mathbf{h}}}[\log\mathbf{D}^{\eta}(\mathbf{h},\mathbf{y})] + \mathbb{E}[\log(1-\mathbf{D}^{\eta}(\mathbf{h},\mathbf{F}^{\theta}(\mathbf{h})))] \tag{8}$$

which learns to distinguish whether a sentence is real or generated. This can be viewed as an extended version of policy gradient, where the reward function is replaced by a discriminator that is simultaneously trained. Inspired by [7], we also add mis-matched context-truth pairs to improve the generation's semantic relevance.

**Implementation details.** All $\mathbf{F}$ functions are substantiated as a 1024-d LSTM [9] cell on top of a 512-d embedding layer, followed by a linear layer with hyperbolic tangent non-linearity to compute the final encoding. Our GAN discriminator is implemented as an 3-layer, 512-d MLP that takes sentence encoding and context vector as inputs.

Nested hierachical neural networks are difficult to train from scratch in an end-to-end fashion. We observe the same for our model. To train our model successfully, we first pre-train our text and face encoders on single sequence corpora. Then we freeze the encoder modules and use them to generate setence-level encodings, which is used to pre-train our history model. Similarly, we pre-train decoders to make them familiar with the context. Finally, we jointly fine-tune the entire framework from end to end.

To train the model with policy gradient and adversarial discriminators, we adopt the MIXER [18] strategy. We initialize the policy network with MLE training. Then we gradually anneal MLE steps and blend in RL steps temporally. We keep this process until all time steps are replaced by RL. We use greedy inference as baseline to reduce reward variance, as described in Eq.(6). To train our discriminator, we mix same ratio of sampled generation, ground truth, and unrelated ground truth. In both PG and GAN training, we observe that balanced positive and negative reward in samples are important to training success. In our case, we find knocking out random samples by setting zero reward until balanced average reward particularly effective. We use clipped gradient descent in our pre-training steps, and Adam [11] in all other training steps.

We pre-train our micro-gesture module on the FirstImpression dataset [3], which contains close-up talking videos that allows high precision micro-gesture tracking. To synchronize words and gestures at frame level, we perform speech recognition with Bluemix, and callibrate the noisy recognition with transcription using Smith-Waterman alignment [25]. We reduce the jittering effect of our generation using an online Savitzky-Golay filter [20].

|  |  | beam=1 |  |  | beam=3 |  |  | beam=5 |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
|  | *perp.* | *pre. %* | *rec. %* | *f1* | *pre. %* | *rec. %* | *f1* | *pre. %* | *rec. %* | *f1* |
| Text [12, 26] | 32.53 | 23.18 | 15.58 | 17.12 | **25.00** | 17.13 | 18.62 | **24.70** | 16.91 | 18.34 |
| Text+RandFace | 32.65 | 22.92 | 15.99 | 17.27 | 24.74 | 17.32 | 18.57 | **24.71** | 17.82 | 18.84 |
| Text+Face | **30.17** | 24.25 | 17.52 | 18.69 | **24.78** | 18.60 | 19.40 | 24.34 | 18.74 | 19.37 |
| History-rnn [22] | 31.15 | 23.99 | 19.46 | 19.59 | 23.79 | 20.11 | 19.67 | 23.37 | **20.50** | 19.68 |
| History-fc | 30.39 | 24.49 | 19.61 | 19.88 | 24.38 | **20.50** | **20.14** | 23.70 | 20.45 | 19.91 |
| Ours-mle | **30.08** | 25.16 | 19.72 | 20.17 | 24.50 | **20.32** | **20.11** | 23.75 | **20.47** | 19.89 |
| Ours-pg | 31.91 | **25.16** | **20.24** | **20.42** | 24.48 | 20.26 | 20.02 | 24.06 | 20.33 | **19.96** |
| Ours-gan | 31.60 | **25.23** | **20.19** | **20.44** | 24.56 | 20.31 | 20.08 | 24.11 | 20.38 | **19.97** |

Table 1: The mind-reading text results on text. Second column lists word perplexity (lower the better). Third to last columns list unigram *precision*, *recall*, and *f1* score (higher the better) across different beam search size. For each column, we mark the **best** and **second best** results in red and blue color. We underscore the **overall best** result across all methods and all beam sizes.

|  |  | beam=1 |  |  | beam=3 |  |  | beam=5 |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
|  | *perp.* | *pre. %* | *rec. %* | *f1* | *pre. %* | *rec. %* | *f1* | *pre. %* | *rec. %* | *f1* |
| Face [26] | 18.98 | 26.48 | 9.83 | 12.96 | 22.41 | 8.18 | 10.82 | 20.74 | 7.55 | 10.02 |
| Face+RandText | 18.94 | 26.63 | 10.01 | 13.15 | 22.54 | 8.15 | 10.82 | 20.20 | 7.43 | 9.80 |
| Face+Text | 17.20 | 29.46 | 10.89 | 14.41 | 25.84 | 9.46 | 12.57 | 24.82 | 9.14 | 12.15 |
| History-rnn [22] | 20.30 | 20.84 | 7.33 | 9.81 | 20.84 | 7.33 | 9.81 | 20.84 | 7.33 | 9.81 |
| History-fc | 20.26 | 20.86 | 7.35 | 9.83 | 20.81 | 7.33 | 9.80 | 20.84 | 7.33 | 9.81 |
| Ours-mle | **17.18** | 35.81 | 13.74 | 18.07 | **30.44** | **11.43** | **15.10** | **28.25** | **10.58** | 13.49 |
| Ours-pg | 17.20 | **36.17** | **13.92** | **18.28** | 30.42 | **11.43** | **15.09** | **28.30** | **10.63** | **14.06** |
| Ours-gan | **17.19** | **36.06** | **13.85** | **18.20** | **30.43** | 11.38 | 15.05 | 28.12 | 10.52 | **13.92** |

Table 2: The mind-reading test results on gesture. Legend same as Table 1.

# 5. Experiments

We evaluate our model through automatic metrics with a "mind-reading" test, and through human study with a NeuralHank chat avatar driven by our model. We randomly split MovieChat into 4:1:1 train-validate-test in all experiments.

## 5.1. The Mind-Reading Test

In this experiment, we evaluate how well the model's generation matches with the ground truth target text and gesture. This reflects the models ability to produce appropriate, human-like responses. We note that this is an extremely challenging task. Due to the multi-modal nature of chit-chat conversations, there exists many plausible responses to the same query, and the ground truth only represents one mode among many. Therefore, we refer to this evaluation as the mind-reading test.

We evaluate the model at both word and sentence level. At the word level, we evaluate the *perplexity*, i.e. the likelihood of generating the correct next target word, given the source and correct previous words in target. By doing this this, we aim to measure coherence of the text and facial languages. At the sentence level, we evaluate the *precision*, *recall*, and *f1 score* between the words appear in generation and ground truth. Despite our method also achieves higher BLEU than baselines, we instead choose these metrics because we rarely observe grammar mistake in generation due to extensive language model pre-training, and the proportion of times certain keywords are correctly guessed becomes a better reflection of mind-reading capability.

We compare with five baselines. 1. Text(Face): The classic Seq2Seq [12, 26] method that uses text(face)-only single query sentence. 2. Text+Face(Face+Text): Two encoders for both text and face query sentences without history. 3. Text+RandFace(Face+RandText): Same model as previous but trained with randomized face(text) query sentences. 4. History-rnn: Modelling conversation history as well as query text(face) using a hierachical RNN, which is similar to to [22]. 5. History-fc: Same as previous but directly connects history sentences to the decoder with fully connected layers. This exploit the potential in conversation history, at the cost of inflexibility to history length $N$ and significantly heavier models. For our model, we compare Ours-mle, Ours-pg, and Ours-gan. We use beam search with varying sizes for all methods.

From Table 1, 2, it can be seen that our method achieves the best performance. Due to non-overlapping conversation scenes between data splits, the improvement of our methods is meaningful and generalizable. Therefore, our experiments quantitatively prove the common intuition that seeing the face makes understanding conversation easier and better, backing the main argument of this paper. Our base model can be further improved using reinforcement and adversarial training. GANs do not achieve better automatic metric score than directly setting metric reward for PG. However, GANs are able to generate more diverse and interesting responses, as we will later show in Sec. 5.2. This finding is in accordance with image captioning [7].

**The role of gesture**. In Table 1, Text+Face outperforms the text-only method, indicating gesture information helps text understanding. Text+RandFace is unable to achieve any improvement, which verifies the improvement of Text+Face is not due to the additional model capacity of the additional encoder network, and further confirms the usefulness of gesture information. Our method outperforms both History-

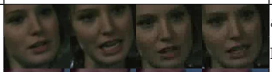
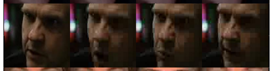
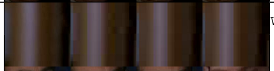
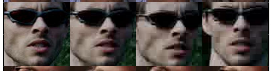
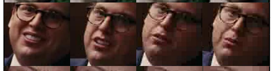
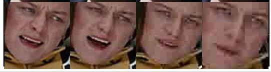
| source text | source face sequence | true target text | text only [12, 26] | text+face |
|---|---|---|---|---|
| we went to the hickory stick, we had a drink, two drinks. |  | and then? and then i went home alone. | we drank a bottle of champagne. | and then i went to bed. |
| she doesn't know where he is. | | i don't know where he is. | i'm sorry. | i don't know where she is. |
| and he sleeps only one hour a night. |  | he's a great man. | he sleeps in the same bed. | he's a good man. |
| a night that marked the opening of a new chapter. | | in world history. | for the future. | in the history of the world. |
| i hope you're not a hothead like sonny. | | he's a good kid. | he's got a lot of something. | he's a good kid. |
| i guess they was worried they wouldn't find a vein in my arm. |  | what's that number? | what's that? | i don't think so. |
| oh, he's so cute. | | oh, my god. | oh, my god | he's so cute. |
| can you hear me? i'm still here. scott. stop. |  | i'm here. scott. stop. | i'm sorry. | what the f*** are you doing here? |
| so i don't really remember, yeah. |  | yeah, right. stupid. | yeah, yeah, yeah. | well, you know what? i'm sorry. |
| i can't feel my legs. |  | i can't feel my legs. | and i can't breathe. | it's too much. |

Figure 7: Success and failure cases of using face along with text. Top five rows show successful examples where adding facial gesture information produces sentences closer to the ground truth. Bottom five rows show failure modes, including face detection failure in the sixth row, and detecting another face that does not belong to the speaker in the seventh row.

rnn and History-fc. This shows the compatiblity between gesture and history information.

**The role of text**. Similarly, from Table 2 it can be seen that in understanding and generating face gestures, text information is helpful. This confirms the mutual benefits between text and gesture information.

**The role of history**. In both text and face, History-fc outperforms History-rnn. This indicates there is still room for further improvement for better history encoders. However, History-rnn remains the preferable solution, for its smaller model size and flexibility to varying history length, which are important in practical scenarios. Interestingly, history methods outperforms Text+Face in text mind reading, indicating that multiple sentences of text history is more helpful than a query face sequence. It is the opposite in gesture mind reading, indicating when guessing facial gesture, seeing the source face and react accordingly can be more helpful than knowing a series of text-only history sentences. This can be also partially due to the nature of movie data, where both source and target can be said by the same character.

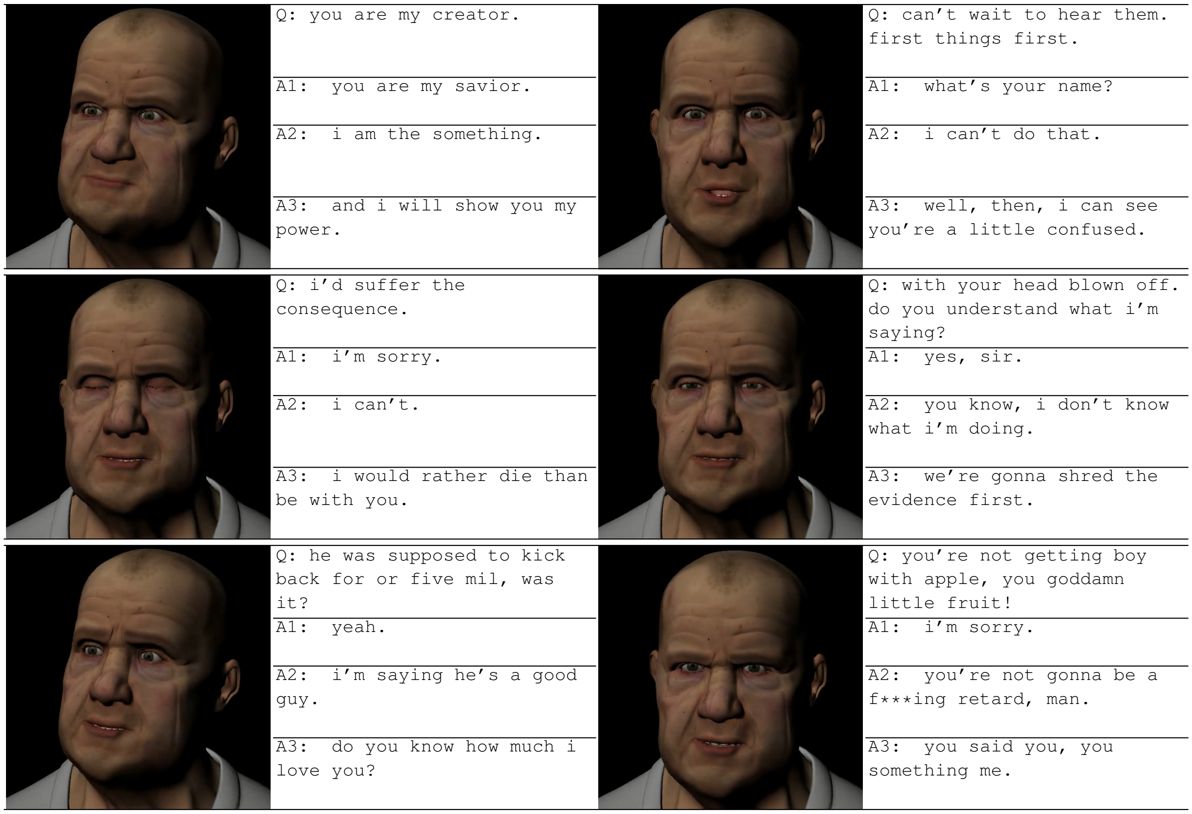
## 5.2. The NeuralHank Chatbot

In this experiment, we test how our model performs in the eyes of real human users. We create a virtual chatbot named NeuralHank, that is driven by our model. By this experiment, we aim to demonstrate the potential of our model and provide a pilot study towards new applications, e.g. AI assistant and gaming/HCI.

In NeuralHank, we ensemble a series of off-the-shelf packages to convert our model's generation into a real talking avartar. We use Microsoft Speech API to render text as audio, while also keeping record of viseme time tags. We then render FACS gesture using Maya's Facial Animation Toolset, with its default character Hank. For Hank's lip motion, we simply use the viseme event record with tangent interpolation over time. On training, we continue for a few more epochs after the early stopping point until training loss is below a certain threshold. We found this makes the model's generation more particular, which is helpful for building a lively avatar.

We compare three methods: 1. *nomicro-nogan* that does not use GANs, and only word-level face decoder without micro-gesture RNN. 2. *micro-nogan* that uses the micro-gesture RNN but not GANs. 3. *ours* as our full model that uses both.

We conduct human study via Amazon Mechanical Turk (AMT), by asking paticipants to rate Hank's responses using different methods on the same query. We ask participants to choose the more interesting and natural response, in terms of text, gesture, and overall. To keep the query subjectivity, we randomly choose 65 query sentences from our held-out test set and run all methods using them as input. Most AMT participants are not well-trained experts. It is important to make our study easy to follow. To achieve this, we only display a pair of different methods' genera-



Q: you are my creator.
A1: you are my savior.
A2: i am the something.
A3: and i will show you my power.

Q: can't wait to hear them. first things first.
A1: what's your name?
A2: i can't do that.
A3: well, then, i can see you're a little confused.

Q: i'd suffer the consequence.
A1: i'm sorry.
A2: i can't.
A3: i would rather die than be with you.

Q: with your head blown off. do you understand what i'm saying?
A1: yes, sir.
A2: you know, i don't know what i'm doing.
A3: we're gonna shred the evidence first.

Q: he was supposed to kick back for or five mil, was it?
A1: yeah.
A2: i'm saying he's a good guy.
A3: do you know how much i love you?

Q: you're not getting boy with apple, you goddamn little fruit!
A1: i'm sorry.
A2: you're not gonna be a f***ing retard, man.
A3: you said you, you something me.

Figure 8: NeuralHank examples. `Q` is the query text. `A1`, `A2`, and `A3` are generated by nomicro-beam, micro-nogan, and ours, respectively. We also show the face image generated by our method. First two rows show that our GAN-based model generates more diverse and interesting responses. Last row shows examples where our method fails and generates confusing responses. Please refer to our supplemental material for example videos.

| | *text %* | *face %* | *overall %* | |
|---|---|---|---|---|
| nomicro-nogan | 48.8 | 39.0 | 46.2 | *pairwise* |
| micro-nogan | **51.2** | **61.0** | **53.8** | |
| nomicro-nogan | 44.8 | 35.3 | 42.4 | |
| ours | **55.2** | **64.7** | **57.6** | |
| micro-nogan | 46.1 | 48.8 | 46.7 | |
| ours | **53.9** | **51.2** | **53.3** | |
| nomicro-nogan | 31.5 | 25.0 | 29.8 | *accumu.* |
| micro-nogan | 32.5 | 36.8 | 33.5 | |
| ours | **36.0** | **38.2** | **36.6** | |

Table 3: AMT user study on interestingness and naturalness. The evaluation is conducted in form of pairwise comparison. We further accumulate number of votes for different methods.

tions in random order, instead of showing all three together. We also intentionally set query text as the most important information, and set query gesture and history as zero. This makes our task easy to understand, while not affecting the fairness of comparison because the methods only differ on the decoder side.

We request 10 evaluations for each sample. This results in 5850 answers from 37 unique participants. Due to the novelty of our study and the rush tendency of users, our evaluation concept is not well understood by all participants. To address this, we stress our goal of identifying *interesting* and *natural* responses in task description. We further use exam questions to filter out rushing participants. The questions are verified samples where one answer is obviously better, e.g. a spot on, grammar error free, and fun sentence, versus a simple and boring yes/no answer.

It can be seen from Table 3 that micro-gesture significantly improves gesture quality. Our full model with adversarial training achieves the best user rating among from all three perspectives. Compared to non-GAN methods that tends to produce universally correct but less interesting responses, GAN methods produces generally more diverse and interesting responses. However, GAN methods also suffer from occational confusing or offensive responses. Fig. 8 shows generated samples.

## 6. Conclusion

We proposed a face-to-face neural conversation model, an encoder-decoder neural architecture trained with RL and GAN. Our approach used both textual and facial information to generate more appropriate responses for the conversation. We trained our model by exploiting rich video data in form of movies. We evaluated our model through a mind-reading test as well as a virtual chatting avatar.

## References

[1] https://www.vicon.com/products/camera-systems/cara-1. 3, 4

[2] T. Baltrusaitis, P. Robinson, and L.-P. Morency. Openface: an open source facial behavior analysis toolkit. In *WACV*, pages 1–10, 2016. 2, 3, 4

[3] J.-I. Biel and D. Gatica-Perez. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Trans. on Multimedia*, 15(1):41–55, 2013. 2, 5

[4] M. J. Black and Y. Yacoob. Recognizing facial expressions in image sequences using local parameterized models of image motion. *IJCV*, 25(1):23–48, 1997. 3

[5] K. M. Colby. Modeling a paranoid mind. *Behavioral and Brain Sciences*, 4:515–534, 1981. 2

[6] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *TPAMI*, 23(6):681–685, 2001. 3

[7] B. Dai, D. Lin, R. Urtasun, and S. Fidler. Towards diverse and natural image descriptions via a conditional gan. In *ICCV*, 2017. 5, 6

[8] P. Ekman, W. V. Freisen, and S. Ancoli. Facial signs of emotional experience. *Journal of personality and social psychology*, 39(6):1125, 1980. 1, 3, 4

[9] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 5

[10] L. Hu, S. Saito, L. Wei, K. Nagano, J. Seo, J. Fursund, I. Sadeghi, C. Sun, Y.-C. Chen, and H. Li. Avatar digitization from a single image for real-time rendering. In *SIGGRAPH Asia*, 2017. 2

[11] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[12] R. Kiros, Y. Zhu, R. Salakhutdinov, R. Zemel, A. Torralba, R. Urtasun, and S. Fidler. Skip-thought vectors. *NIPS*, 2015. 2, 6, 7

[13] S. Levine, P. KrShenbYhl, S. Thrun, and V. Koltun. Gesture controllers. *ACM Trans. on Graphics*, 29(4), 2010. 2

[14] S. Levine, C. Theobalt, and V. Koltun. Real-time prosody-driven synthesis of body language. *ACM Trans. on Graphics*, 28(5), 2009. 2

[15] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. A diversity-promoting objective function for neural conversation models. In *arXiv:1510.03055*, 2015. 1, 2

[16] J. Li, M. Galley, C. Brockett, G. P. Spithourakis, J. Gao, and B. Dolan. A persona-based neural conversation model. In *arXiv:1603.06155*, 2016. 1, 2

[17] J. Li, W. Monroe, A. Ritter, M. Galley, J. Gao, and D. Jurafsky. Deep reinforcement learning for dialogue generation. *arXiv:1606.01541*, 2016. 5

[18] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba. Sequence level training with recurrent neural networks. *arXiv:1511.06732*, 2015. 5

[19] A. Ritter, C. Cherry, and W. B. Dolan. Data-driven response generation in social media. In *EMNLP*, 2011. 2

[20] A. Savitzky and M. J. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964. 5

[21] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *arXiv:1507.04808*, 2015. 2

[22] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. *arXiv:1507.04808*, 2015. 6

[23] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *arXiv:1605.06069*, 2016. 2, 4

[24] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 2

[25] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981. 5

[26] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014. 4, 6, 7

[27] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. What makes tom hanks look like tom hanks. In *ICCV*, 2015. 2

[28] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. In *SIGGRAPH*, 2017. 2

[29] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, 2016. 3

[30] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niesner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, 2016. 2

[31] O. Vinyals and Q. Le. A neural conversational model. In *arXiv:1506.05869*, 2015. 1, 2

[32] J. Weizenbaum. Eliza, a computer program for the study of natural language communication between man and machine. *ACM*, 9(1):36–45, 1966. 2

[33] L. Yu, W. Zhang, J. Wang, and Y. Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, 2017. 5

[34] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012. 3

[35] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. In *ICCV*, 2015. 2