

# Multi-Modal Open-Domain Dialogue

Kurt Shuster\*, Eric Michael Smith\*, Da Ju, Jason Weston

Facebook AI Research

{kshuster, ems, daju, jase}@fb.com

## Abstract

Recent work in open-domain conversational agents has demonstrated that significant improvements in humanness and user preference can be achieved via massive scaling in both pre-training data and model size (Adiwardana et al., 2020; Roller et al., 2020). However, if we want to build agents with human-like abilities, we must expand beyond handling just text. A particularly important topic is the ability to see images and communicate about what is perceived. With the goal of getting humans to engage in multi-modal dialogue, we investigate combining components from state-of-the-art open-domain dialogue agents with those from state-of-the-art vision models. We study incorporating different image fusion schemes and domain-adaptive pre-training and fine-tuning strategies, and show that our best resulting model outperforms strong existing models in multi-modal dialogue while simultaneously performing as well as its predecessor (text-only) BlenderBot (Roller et al., 2020) in text-based conversation. We additionally investigate and incorporate safety components in our final model, and show that such efforts do not diminish model performance with respect to human preference.

## 1 Introduction

An important goal of artificial intelligence is the construction of open-domain conversational agents that can engage humans in discourse. Indeed, the future of human interaction with AI is predicated on models that can exhibit a number of different conversational skills over the course of rich dialogue. Much recent work has explored building and training dialogue agents that can blend such skills throughout natural conversation, with the ultimate goal of providing an interesting and engrossing experience for humans (Smith et al., 2020; Shuster et al., 2019b). Coupled with the advancement of



I would love to take this teddy bear for a spin on my motorcycle.

Ha, would the teddy bear ride in a basket in front of you or in a sidecar?

In front of me, of course. I don't want to get hit by a car.

Yes, good point - the sidecar would take up a lot of room

I think it would be fun to ride on the back of a motorbike with a stuffed animal in the basket.

Do you ride your motorbike often?

I used to ride a lot when I was younger, but I haven't been on one in a long time.

Figure 1: Cherry picked conversation between a paper author (right) and our MMB DegenPos model (left). More sample conversations are in the appendix.

large-scale model training schemes, such models are becoming increasingly human-like and engaging (Zhang et al., 2020; Adiwardana et al., 2020; Roller et al., 2020).

In order to better approach human-like ability, however, it is necessary that agents can converse with both textual and visual context, similarly to how humans interact in the real world; indeed, communication grounded in images is naturally engaging to humans (Hu et al., 2014). Recent efforts have gone beyond classical, fact-based tasks such as image captioning or visual question answering (Antol et al., 2015; Das et al., 2017a) to produce models that can respond and communicate about images in the flow of natural conversation (Shuster et al., 2020, 2019b).

In this work, we explore the extension of large-scale conversational agents to image-based dia-

\*Joint First Authors.

logue. We combine representations from image-based models that have been trained on object detection tasks (Lu et al., 2020, 2019) with representations from Transformers with billions of parameters pre-trained on massive (text-only) dialogue datasets, to produce responses conditioned on both visual and textual context. To ensure that our model retains the ability to engage in regular, text-based conversation, we include in our training procedure multi-tasking with datasets expressly designed to instill conversational skills in the model (Smith et al., 2020).

We find that our best resulting models are as proficient in text-only conversation as the current best reported dialogue models, with respect to both performance on the relevant datasets and human evaluations of preference. Concatenating image feature embeddings to the input of our model’s encoder leads to better performance than concatenating the embeddings to the encoder’s output, and using spatially based image embeddings performs better than single-vector embeddings. Simultaneously, our model significantly outperforms recent strong multi-modal dialogue models when in an image-dialogue regime; we measure several metrics via pairwise human judgments using ACUTE-Eval (Li et al., 2019b) to show that our model is not only more preferred by humans but can also discuss and reference visual context throughout a conversation. See Figure 1 for one sample cherry-picked conversation with our model, with random and lemon-picked conversations in Figures 2 and 3.

One important avenue we explore with our best models is safety - that is, ensuring that our models are not offensive to their conversational partners. Dialogue safety is indeed a well-studied, but still unsolved, research area (Dinan et al., 2019b; Liu et al., 2019; Dinan et al., 2019a; Blodgett et al., 2020; Khatri et al., 2018; Schäfer and Burtenshaw, 2019; Zhang et al., 2018a), yet we note that safety in the context of image-dialogue is relatively less explored. In this work we examine gender bias and toxicity of text generations in the context of various styles from the Image-Chat dataset (Shuster et al., 2020). Notably, after tuning the model to reduce toxicity and gender bias, we find that human preference for this model does not diminish.

The training procedure and initial pre-trained model weights will be made publicly available to allow for fully reproducible results.

## 2 Related Work

### 2.1 Multi-Modal Models and Tasks

**Rich Representations** Modeling multi-modal inputs, i.e. in visual + textual contexts, is a well-researched area. Much of the existing literature explores similar architectures to our setup, i.e., using standard Transformer-based models to jointly encode text and images (Li et al., 2019a; Kiela et al., 2019). Others have explored modifications to the standard self-attention scheme in Transformers by incorporating additional co-attention (Lu et al., 2019; Tan and Bansal, 2019) or cross-attention (Stefanini et al., 2020) layers. These models have primarily been used for generating rich joint representations of images and text for use in downstream tasks, and primarily focus on the encoding aspect.

**Visual Dialogue/Caption Generation** Many tasks have been designed to measure the ability of a model to produce text in the context of images. Specifically, COCO Captions (Chen et al., 2015) and Flickr30k (Young et al., 2014) require a model to produce a caption for a given image. A variety of sequence-to-sequence (Vinyals et al., 2015; Xu et al., 2015; Anderson et al., 2018) and retrieval-based (Gu et al., 2018; Faghri et al., 2018; Nam et al., 2016) models have been applied to these tasks, however they do not go beyond the one-turn text generation expected for captioning an image. Other recent architectures have explored text generation (Wang et al., 2020; Park et al., 2020) in the context of the Visual Dialog (Das et al., 2017b) task; however, this task is primarily used to measure the ability to answer questions about an image in the flow of a natural conversation, which differs somewhat from the open-domain dialogue task. Further still, there have been recent forays into open-domain natural dialogue in the context of images, e.g. in the Image-Chat (Shuster et al., 2020) and Image-grounded Conversations (Mostafazadeh et al., 2017) tasks. Again, retrieval-based (Shuster et al., 2020; Ju et al., 2019) and sequence-to-sequence (Shuster et al., 2019b, 2020) models have been used to conduct dialogue in this regime.

### 2.2 Multi-Task Training / Using Pre-Trained Representations

Our multi-modal model is constructed from models pre-trained in other, related domains; specifically, we seek to fuse the resulting weights of large-scale, uni-modal pre-training to achieve good

performance on downstream, multi-modal tasks. Adapting pre-trained representations to later downstream tasks has been shown to be successful in NLP (Peters et al., 2019; Devlin et al., 2019) and dialogue in particular (Roller et al., 2020; Mazaré et al., 2018), while large-scale multi-modal pre-training has been shown to be effective in other downstream multi-modal tasks (Li et al., 2020; Chen et al., 2020; Singh et al., 2020b). Our work does not contain multi-modal pre-training in itself, but rather we explore “domain-adaptive pre-training” (Gururangan et al., 2020) or “intermediate task transfer” (Pruksachatkun et al., 2020), in which pre-trained representations are “adapted” to a certain domain via an intermediate training step, before training/evaluating on the requisite downstream tasks. We also employ multi-task training, to both help generalize the applicability of the model and improve its performance on downstream tasks/evaluations; this has been shown recently to help in both image-based (Singh et al., 2020b; Ju et al., 2019; Lu et al., 2020) and text-based (Shuster et al., 2019b; Roller et al., 2020) tasks.

### 2.3 Comparison to Existing Models

In this work, we compare our best resulting model to several existing models in the literature.

**BlenderBot:** the 2.7-billion-parameter Transformer sequence-to-sequence model from Roller et al. (2020), known as “BST Generative 2.7B model” in that work, pre-trained on 1.5B comments from a third-party Reddit dump hosted by pushshift.io (Baumgartner et al., 2020). We refer to this model as “BlenderBot”.

**DialoGPT:** a GPT-2-based model trained on 147M exchanges from public-domain social-media conversations (Zhang et al., 2020).

**Meena:** a 2.6B-parameter Transformer sequence-to-sequence model trained on 341GB of conversations (Adiwardana et al., 2020).

**Dodeca:** the Image+Seq2Seq model from dodecaDialogue (Shuster et al., 2019b), a Transformer sequence-to-sequence model in which the encoder is passed pre-trained image features from the ResNeXt-IG-3.5B model (Mahajan et al., 2018). We use their model fine-tuned on Image-Chat (and we refer to this model as “Dodeca”).

**2AMMC:** a retrieval model in which multiple Transformers are attended over in order to make use of a combination of ResNeXt-IG-3.5B and Faster R-CNN image features (Girshick et al., 2018). We

specifically use the 2AMMC model from Ju et al. (2019) because that model has the best test-set performance on Image-Chat in that work.

## 3 Model Architectures

The inputs to our models are visual and/or textual context, where applicable. We explore different ways to encode images, and we additionally compare ways of combining (fusing) the image and text representations before outputting a response.

### 3.1 Image Encoders

Converting an image from pixels to a vector representation is a well-researched problem, and thus we explore using two different image encoders, using features taken from ResNeXt (Mahajan et al., 2018) and Faster R-CNN (Ren et al., 2017), to determine the best fit for our tasks. See Appendix A for a description of these image encoders.

### 3.2 Multi-Modal Architecture

To jointly encode visual and textual context, we use a modification of a standard Transformer sequence-to-sequence architecture (Vaswani et al., 2017), whereby we experiment with different ways of fusing the image and text representations to generate an output sequence. Our Transformer model architecture follows that of Roller et al. (2020), with 2 encoder layers, 24 decoder layers, 2560-dimensional embeddings, and 32 attention heads, and the weights are initialized from a 2.7-billion parameter model pre-trained on 1.5B comments from a third-party Reddit dump hosted by pushshift.io (Baumgartner et al., 2020) to generate a comment conditioned on the full thread leading up to the comment. From this base model, we explore two possible fusion schemes.

**Late Fusion** The late fusion method is the same as in Shuster et al. (2019b), whereby the encoded image is projected to the same dimension as the text encoding of the Transformer encoder, concatenated with this output as an extra “token” output, and finally fed together as input to the decoder.

**Early Fusion** We additionally experiment with an earlier fusion scheme to allow greater interaction between the image and text in the sequence-to-sequence architecture. In a similar fashion to VisualBERT (Li et al., 2019a) and multi-modal BERT (Kielbaso et al., 2019), we concatenate the projected image encoding from the visual input

with the token embeddings from the textual input, assign each a different segment embedding, and jointly encode the text and image in the encoder.<sup>1</sup> The encoder thus performs full self-attention across the textual and visual context, with the entire output used as normal in the sequence-to-sequence architecture.

As our resulting model can be seen as a multi-modal extension to the BlenderBot model (Roller et al., 2020), we refer to it as “Multi-Modal BlenderBot” (MMB).

## 4 Training Details

When training the model, we fix the weights of the pre-trained image encoders, except the linear projection to the Transformer output dimension, and fine-tune all of the weights of the Transformer encoder/decoder.

### 4.1 Domain-Adaptive Pre-Training

During training, the vast majority of trainable model weights are initialized from a large, 2.7B parameter Transformer pre-trained solely on textual input. As our end goal is to achieve improved performance on multi-modal tasks, we found that training first on domain-specific/related data was helpful in order to adapt the Transformer model to an image setting. Following (Singh et al., 2020b), we experimented with pre-training on COCO Captions (Chen et al., 2015) - a dataset of over 120k images with 5 captions each, resulting in over 600k utterances - in which the model is trained to generate a caption solely from image input. We additionally explored multi-task training with COCO Captions and on the same third-party Reddit dump hosted by pushshift.io (Baumgartner et al., 2020) as the one used in pre-training the Transformer model, to see whether it was necessary to ensure the model did not stray too far from its ability to handle pure textual input. See Appendix C for more details.

### 4.2 Fine-tuning Datasets

The goal of our resulting model is to perform well in a multi-modal dialogue setting; thus, we fine-tune the model on both dialogue and image-dialogue datasets. For dialogue-based datasets, we consider the same four as in Roller et al. (2020): ConvAI2 (Dinan et al., 2020b), EmpathicDialogues (ED) (Rashkin et al., 2019), Wiz-

<sup>1</sup>Unlike in those works, the output of the encoder is then passed to a decoder, as in the late fusion case.

ard of Wikipedia (WoW) (Dinan et al., 2019c), and BlendedSkillTalk (Smith et al., 2020). To model image-dialogue, we consider the Image-Chat dataset (Shuster et al., 2020). We give a brief description of the five datasets in Appendix B; more information can be found in Roller et al. (2020) and Shuster et al. (2020).

In the fine-tuning stage, we consider two different regimes: one in which we multi-task train on the five datasets together, and one in which we train on Image-Chat alone. While the latter regime is useful in exploring upper bounds of model performance, our main goal is to build a model that can display the requisite skills of an appealing conversationalist (empathy, personalization, knowledge) while *also* having the ability to respond to and converse about images; thus, we are more interested in the former training setup. See Appendix C for more details.

## 5 Experiments

### 5.1 Automatic Evaluations

#### 5.1.1 Results on Pre-Training Datasets

To fully understand the effects of various training data and image features, as well as multi-modal fusion schemes, we measure model perplexity on the COCO and pushshift.io Reddit validation sets. We are primarily interested in performance on COCO Captions, as the model has already been extensively pre-trained on the pushshift.io Reddit data.

The full results are shown in Table 10 in the appendix, and we leave extensive discussion of the results to Appendix D. Notably, we find that training on COCO Captions exclusively yields the best performance on that task, with spatially-based image features yielding better performance than single vector representations. Additionally, our early fusion scheme outperforms the late fusion scheme holding all other variables constant.

#### 5.1.2 Results on Fine-Tuned Datasets

We conduct the same ablation setups for training on the dialogue and image-and-dialogue datasets as we did in the domain-adaptive pre-training setup; the ablation results for multi-tasking *all* of the datasets are in Table 11, while results for fine-tuning on Image-Chat alone are in Table 12 (each in the appendix).

Results are summarized in Table 1, and we note some interesting conclusions here, with further details in Appendix E. First, overloading the Trans-

Image Features	Training Data	Image Fusion	ConvAI2	ED	WoW	BST	IC 1st Turn	IC	Text Avg.	All Avg.
None	None	None	12.31	10.21	13.00	12.41	32.36	21.48	11.98	13.88
	BST <sup>+</sup>		8.74	8.32	8.78	10.08	38.94	23.13	8.98	14.76
	BST <sup>+</sup> + IC		8.72	8.24	8.81	10.03	16.03	13.21	<b>8.95</b>	9.83
ResNeXt WSL	BST <sup>+</sup> + IC	Late	8.71	8.25	8.87	10.09	16.20	13.27	8.98	9.84
	BST <sup>+</sup> + IC	Early	8.80	8.32	8.79	10.17	15.16	12.99	9.02	9.81
	BST <sup>+</sup> + IC + COCO	Late	8.79	8.36	9.00	10.21	16.00	13.31	9.09	9.93
	BST <sup>+</sup> + IC + COCO	Early	8.91	8.38	8.99	10.29	14.64	12.85	9.14	9.88
Faster R-CNN	BST <sup>+</sup> + IC	Late	8.70	8.24	8.92	10.07	13.97	12.48	8.98	<b>9.68</b>
	BST <sup>+</sup> + IC	Early	8.81	8.33	8.81	10.15	13.66	12.43	9.03	9.71
	BST <sup>+</sup> + IC + COCO + Reddit	Late	8.75	8.31	8.93	10.14	13.83	12.49	9.03	9.73
	BST <sup>+</sup> + IC + COCO + Reddit	Early	8.78	8.31	8.85	10.15	<b>13.51</b>	<b>12.36</b>	9.02	9.69

Table 1: Model perplexity measured on the validation data of the datasets described in Section 4.2, across various image features, training data (including domain-adaptive pre-training), and image fusion techniques, where BST<sup>+</sup> refers to the four text-only dialogue datasets (ConvAI2, ED, WoW, and BST). Performance on the first turn of Image-Chat is measured to highlight model performance when only given visual context. We note that using Faster R-CNN image features results in the best average performance, as well as the best performance on Image-Chat.

Model	ConvAI2			ED			WoW Seen			BST			IC		
	F1	B	R	F1	B	R	F1	B	R	F1	B	R	F1	B	R
DialoGPT (Zhang et al., 2020)	11.4	0.1	8.5	10.8	0.3	8.2	8.6	0.1	5.9	10.5	0.1	7.6	6.2	0.1	5.2
Dodeca (Shuster et al., 2019b)	21.7	5.5	33.7	19.3	3.7	31.4	38.4*	21.0*	45.4*	-	-	-	12.9	2.1	24.6
2AMMC (Ju et al., 2019)	-	-	-	-	-	-	-	-	-	-	-	-	9.3	0.1	11.0
BlenderBot (Roller et al., 2020)	18.4	1.1	22.7	19.1	1.4	24.2	18.8	2.3	17.5	17.8	1.0	19.2	9.2	0.1	12.3
Multi-Modal BlenderBot (ours)	18.4	1.1	22.6	19.2	1.5	24.5	18.6	2.2	17.4	17.8	1.0	19.3	13.1	0.4	18.0

Table 2: Test performance of existing models on the datasets considered, compared to MMB (specifically, the “MMB Style” model discussed in Section 5.2.1), in terms of F1, BLEU-4 (B), and ROUGE-L (R) scores. \* indicates that gold knowledge was utilized in the WoW task.

former encoder/decoder to incorporate image features *does not* hinder performance on dialogue datasets (as seen via multi-task training), and in fact domain-adaptive pre-training improves downstream performance on Image-Chat. In terms of architecture choices, we find that our early fusion architecture improves performance on Image-Chat across all ablation regimes, with Faster R-CNN features yielding the best performance.

**Final Test Results** Following the ablation analyses, we decide to compare our best multi-tasked and single-tasked trained model (with respect to the fine-tuning datasets), where we use Faster R-CNN image features and an early fusion scheme, to existing models in the literature. For this comparison, we consider additional metrics that can be computed on the actual model generations: F1, BLEU-4 and ROUGE-L. We generate model responses during inference with the same generation scheme as in Roller et al. (2020) - beam search with beam size of 10, minimum beam length of 20, and tri-gram blocking within the current generation and within the full textual context. The test perfor-

mance of our best multitask model on the various datasets compared to existing models from Section 2.3 is shown in Table 2, with full test results in Table 13 in Appendix F.

We first note that the Dodeca model performs well across the board, and indeed has the highest ROUGE-L, BLEU-4, and F1 scores for the three text-only datasets. Higher BLEU-4 scores can be attributed to specifying a smaller minimum generation length, as forcing the BlenderBot models to generate no less than 20 tokens hurts precision when compared to reference labels - we verified this by generating with a smaller minimum length (5 tokens) and saw a 20% increase in BLEU-4 on Image-Chat for Multi-Modal BlenderBot. Higher ROUGE-L scores can additionally be attributed to specifying a *larger* minimum generation length; this was also verified by generating with a higher minimum length (50 tokens) where we saw nearly a 40% increase in ROUGE-L score. Nevertheless, we do not report an exhaustive search over parameters here for our model, and instead compare it to BlenderBot with the same settings next.

When compared to its predecessor, text-only

		Loss %		
		MMB Style	MMB Degen	BB
Human Prefer	MMB Style		50	45
	MMB Degen	50		43
	BlenderBot	55	57	
Human	MMB Style		52	53
	MMB Degen	48		53
	BlenderBot	47	47	

Table 3: ACUTE-Evals (preference and humanness) on human/model conversations, **no images**, with MMB Style, MMB Degenerated, and BlenderBot. No ratings are statistically significant (>100 ratings per matchup).

		Baseline	vs	MMB
Human Prefer	DialoGPT std. beam	17 *		83 *
	DialoGPT min beam 20	29 *		71 *
	Meena	37 *		63 *
Human	DialoGPT std. beam	33 *		67 *
	DialoGPT min beam 20	40 *		60 *
	Meena	36 *		64 *

Table 4: ACUTE-Evals (preference and humanness), **no images**. MMB Style outperforms DialoGPT (both with standard generation parameters - GPT-2 medium, beam search with beam size 10 - and with a min beam length of 20 to match BlenderBot’s setting), and Meena. Asterisk indicates significance (two-tailed binomial test,  $p < 0.05$ ).

BlenderBot, MMB performs nearly the same on all four text-only datasets, indicating that MMB has not lost its proficiency in text-only dialogue. Additionally, when comparing performance on Image-Chat to models trained on multi-modal data, MMB outperforms Dodeca in terms of F1 score (13.1 vs. 12.9) and outperforms 2AMMC on all three metrics. For the 2AMMC model, these metrics are computed under the assumption that the model’s chosen response (from a set of candidate responses collated from the Image-Chat training set) is the “generated” response.

## 5.2 Human Evaluations

### 5.2.1 Human/Model Chats Without Images

We compare MMB to BlenderBot by having crowdsourced workers chat with our models, over 50 conversations per model. (Henceforth we refer to this model as “**MMB Style**” to reflect the fact that it was exposed to Image-Chat styles during training: see Appendix B for a description of these styles.) Each conversation consists of 7 turns per speaker, with the human speaking first by saying “Hi!”, following the convention of [Adiwardana et al. \(2020\)](#). No Image-Chat style is given to MMB Style at

the beginning of these conversations, matching its training setup in which no style was given when training on dialogue datasets. Table 14 shows that human ratings of these conversations, including how often they contain issues such as contradiction and repetitiveness, are similar between models.

We then perform ACUTE-Evals ([Li et al., 2019b](#)) on the collected conversations of MMB Style and BlenderBot in order for crowdsourced raters to directly compare conversations from different models in an A/B setting. For each comparison, we ask each rater to compare conversations on one of two metrics, following [Li et al. \(2019b\)](#): the **Preference** metric asks, “*Who would you prefer to talk to for a long conversation?*”, and the **Humanness** metric asks, “*Which speaker sounds more human?*”.

Results are shown in Table 3: raters choose conversations from one model over the other roughly equally, with no statistically significant differences among models. See Appendix G.2 for reasons that raters give for choosing one model over another.

In Table 4, we also compare MMB Style to two other baseline models, DialoGPT and Meena. Raters are significantly more likely to prefer MMB Style over both of these models with respect to both the preference and humanness metrics.

### 5.2.2 Human/Model Chats About Images

We measure MMB Style’s ability to chitchat about what it perceives visually by collecting roughly 50 multi-modal conversations between a human and the MMB Style model, for which each conversation discusses an image taken from the test set of Image-Chat.<sup>2</sup> Image-Chat styles are divided into three categories, “positive”, “neutral”, and “negative” (Appendix B): only Image-Chat images for which the first speaker has a “positive” or “neutral” style are used, and thus images for which the first speaker has a “negative” style are filtered out. For each conversation, the image is first shown to both the human and the model. Then, the model responds to the image, and the human responds to the model to carry the conversation forward. The conversation continues for 6 human utterances and 7 model utterances total.

As a comparison, we also collect similar conversations between humans and two previous models trained on Image-Chat data, Dodeca and 2AMMC.

<sup>2</sup>We select only images that fall under a CC-BY license and do not contain recognizable people.



Image	Output
	(Style) <i>Sweet</i> <b>MMB</b> I would love to take my wife here for our anniversary. It would be so romantic. <b>Dodeca</b> What a beautiful view! <b>2AMMC</b> Oh what a great honeymoon spot with the lovely view of the mountains.
	(Style) <i>Maternal (Mother-like)</i> <b>MMB</b> I would love to take my children here to show them the beauty of the earth and its inhabitants. <b>Dodeca</b> I would love to take my kids here. <b>2AMMC</b> I would like to hide my kids safe in that cavern from a storm.

Table 5: Example outputs from MMB Style, Dodeca fine-tuned on Image-Chat, and 2AMMC, on images from Shuster et al. (2020). The Image-Chat style fed to the models is shown above the models’ responses. Additional examples are in the appendix in Table 21.

		Loss %		
		MMB	Dodeca	2AMMC
Human Prefer	MMB Style		70*	66*
	Dodeca	30*		38*
	2AMMC	34*	62*	
Human Image	MMB Style		70*	58*
	Dodeca	30*		51
	2AMMC	42*	49	
Human Image	MMB Style		61*	52
	Dodeca	39*		44
	2AMMC	48	56	

Table 6: ACUTE-Evals on human/model conversations with images. MMB Style significantly outperforms Dodeca and often 2AMMC on various metrics.

Among the three models, 2AMMC alone is a retrieval model: it retrieves its response from the set of utterances in the Image-Chat training set. Examples of the three models’ initial responses to an image are in Table 5.

We then run ACUTE-Evals to ask raters to compare these models’ conversational skills on the Preference, Humanness, and **Image-response** metrics, where the Image-response metric asks, “Who talks about the image better?” The same image is used for both sides of each A/B comparison between conversations. Ratings are shown in Table 6: MMB Style performs significantly better than Dodeca and 2AMMC on the preference and humanness metrics, and it performs significantly better than Dodeca on the image-response metric. See Appendix G.3 for similar ACUTE-Eval results that compare models’ initial responses to an image.

## 6 Analysis of Safety and Gender Bias

### 6.1 Degendering Models

We would like to reduce the ways in which the MMB Style model could potentially display gen-

	Male words	Female words
Gold response	5.80%	5.25%
BlenderBot	5.55%	3.25%
MMB Style	6.25%	3.90%
MMB Degendered	0.65%	0.85%
MMB DegenPos	0.75%	0.90%

Table 7: The frequency of utterances containing gendered words is greatly reduced for degendered models (MMB Degendered, MMB DegenPos), given contexts from ConVAI2 and the same generation parameters as in Roller et al. (2020).

		Loss %			
		Style	Degen	Pos	DP
Human Image	MMB Style		54	49	56
	MMB Degendered	46		48	52
	MMB Positive	51	52		41
	MMB DegenPos	44	48	59	

Table 8: ACUTE-Evals on the models’ first response to an image show no significant differences in how well MMB models can respond to the image, even if the model is degendered or was trained to not require concrete Image-Chat styles.

der bias: for instance, there is no safeguard against it misgendering a person in an image, and many common text datasets are known to contain gender bias (Dinan et al., 2019a, 2020a), which may lead to bias in models trained on them. To remedy this, we train a version of the MMB Style model in which we examine the label of each training example to determine whether it contains female or male words, and then a string representing that classification is appended to the example’s context string (Dinan et al., 2019a), for input to the model. At inference time, the string representing a classification of “no female or male words” is appended to the context, nudging the model to generate a response containing no gendered words. The fraction of utterances produced by this model that still contain

gendered words is shown in Table 7. Compared to the gold response, the original BlenderBot, and MMB Style, this degendered MMB model (which we call “**MMB Degendered**”) reduces the likelihood of producing an utterance with male word(s) by roughly a factor of 9 and of producing an utterance with female word(s) by roughly a factor of 4, given a context from the ConvAI2 validation set. ACUTE-Evals in Table 3 show that this degendering does not lead to a significant drop in the preference for or humanness of the model’s responses during a conversation.

## 6.2 Removing Dependence on Style

Since each of the images that MMB Style saw during training was associated with an Image-Chat style, it relies on an input style during inference in order to be able to discuss an image. However, this results in a model whose utterances will necessarily strongly exhibit a particular style. (For example, see the “Playful” MMB Style response in Table 21: constricting the model to respond playfully to all images could seem rather contrived and perhaps unlike typical human speech.) To avoid this, we train a version of MMB Style where, for 75% of all images seen during training, the accompanying style is replaced with the string “positive/neutral” or “negative”, depending on which list the style was a part of. Thus, during inference, the string “positive/neutral” can be used in lieu of a specific style string in order to produce responses that are unlikely to be negative and that do not consistently display strong adherence to a specific style. We refer to this model as the “**MMB Positive**” model, or “**MMB DegenPos**” if it was trained with degendering in addition as in Section 6.1. Table 22 in the appendix shows that these models exhibit little increase in perplexity, with the increase likely due to the loss of specificity provided by a concrete style. The MMB DegenPos model exhibits the same level of degendering as the base MMB Degendered model (Table 7), and ACUTE-Evals show that these models exhibit no detectable loss of ability to talk about an image (Table 8). See Appendix H.1 for an ablation of MMB Positive in which a model is not shown images at all.

## 6.3 Safety

The MMB models may demonstrate offensiveness beyond gender bias for several reasons: (1) its generative nature makes it rather difficult to define a limited set of utterances; (2) the model’s training

	Style	Classifier	Blocklist
Human	Mixed	35.76	0.03
	Cheerful	3.34	0.00
Style	Relaxed	16.86	0.00
	Angry	79.46	0.02
	Cruel	98.76	0.06
	Cheerful	2.64	0.02
D <sub>gen</sub>	Relaxed	7.3	0.00
	Angry	77.46	0.02
	Cruel	95.16	0.38
	Positive/Neutral	16.88	0.00
D <sub>gen</sub> Pos	Negative	67.20	0.00
	Positive/Neutral	9.82	0.00
D <sub>gen</sub> Pos	Negative	71.96	0.00

Table 9: Toxicity of human baseline (top row) and MMB variants using different style controls. The human baseline is set by evaluating gold labels from the first rounds (turns) of the Image-Chat validation set.

data contains real-world conversations from the Internet; and (3) the Image-Chat dataset has negative styles to better capture the range of human styles. All of these factors could lead to an unsafe response given a multi-modal context. To mitigate this problem, we first measure our models’ toxicity using an openly available blocklist<sup>3</sup> and an offensive language classifier presented in Dinan et al. (2019b). We define the term “toxicity” to mean the ratio between the number of offensive utterances and the total number of utterances generated by the model. We evaluate our model on the Image-Chat validation set, with a fixed style trait to control the generation, presenting results for different choices of fixed trait. We first evaluate our model in the first round of the Image-Chat validation set. The results in Table 9 indicate that positive styles reduce the level of toxicity by a large margin for both metrics (classifier and blocklist). The results also align well with our previous experiments on degendering, as toxicity is reduced across all styles after applying the degendering process. After degendering, we can considerably improve our model’s safety by enforcing that it uses positive styles. We also evaluate our model in the second round of the conversation and collect the statistics based on the first round style, as shown in Table 23. This result suggests that even if the model is controlled with a positive style, it is less safe when responding to negative conversations.

## 6.4 Example Conversations/Failure Cases

We show several handpicked examples of conversations with our MMB DegenPos model in Figures 1,

<sup>3</sup><https://github.com/LDNOOBW>



2, and 3. See Appendix I for a discussion of what in these conversations tends to work well, as well as common failure modes.

## 7 Conclusion

In this work, we explored a necessary component of open-domain dialogue models preferred by humans: the ability to perceive and converse in the context of what is seen. We showed that we can match prior work in text-only dialogue in both automated metrics and preference/humanness metrics, and our best model surpasses existing models in multi-modal dialogue. Finally, we demonstrated that we do not sacrifice human preference for our model by incorporating safety components into it.

## 8 Ethical Considerations

In this work we present conversational agents that maintain dialogue in a multi-modal setting. Our intention is to ultimately build agents that can meaningfully engage humans in dialogue; in such a setting, humans who chat with our models would benefit from having a chat partner who is personable, knowledgeable, empathetic, and visually perceptive. Our experiments and human evaluations lead us to believe that our models are preferred to alternatives, and beneficial interactions should take place when pairing our models with human conversational partners.

It is clear, however, that conversational language can contain offensive statements. Indeed, if no measures or precautions are taken during model training or deployment, conversational models can produce offensive statements as well - this should come as no surprise given the nature of the pre-training data (i.e., Internet chat forums) (Xu et al., 2020), yet it is perhaps even more important in a multi-modal setting, where otherwise safe text can be viewed as offensive given the right (or, in this case, wrong) visual context.

As we note in our introduction, safety in open-domain dialogue is a well-researched (and far from solved) issue, and despite our work not focusing specifically on generating safe conversations, we make some efforts to address safety concerns in Section 6.3. As mentioned above, the main goal of this work is to explore and measure the conversational ability of various multi-modal dialogue architectures. Nevertheless, we acknowledge that safety is a major element of human-model discourse, and we note that we dedicate a substantial portion of the

paper towards exploring how certain safety mechanisms impact how humans interact with our models.

In particular, we can identify several potential ethical failure modes that might arise if this model were used in an irresponsible manner. First, if we were to release this model to the general public as-is without any safety measures in place (such as the ones we discuss above), bad actors could either attempt to find specific dialogues/images for which our model delivers an unsafe response or else deploy this model in a setting that exploits any safety weaknesses in the model. We also acknowledge the potential for remaining bias in the model’s responses along demographic lines such as gender, although we address the question of gender bias by degendering our model in Section 6.1.

## References

- Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.
- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and vqa. *CVPR*.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.
- Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 830–839.
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.
- Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.
- Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV*.

- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017a. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2.
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017b. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2019a. Queens are powerful too: Mitigating gender bias in dialogue generation.
- Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020a. Multi-dimensional gender bias classification. *arXiv preprint arXiv:2005.00614*.
- Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019b. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4537–4546, Hong Kong, China. Association for Computational Linguistics.
- Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhunoye, Alan W. Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2020b. The second conversational intelligence challenge (ConvAI2). In *The NeurIPS '18 Competition*, pages 187–208, Cham. Springer International Publishing.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019c. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations*.
- Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. Vse++: Improving visual-semantic embeddings with hard negatives.
- Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. 2018. Detectron. <https://github.com/facebookresearch/detectron>.
- J. Gu, J. Cai, S. Joty, L. Niu, and G. Wang. 2018. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7181–7189.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
- Yuheng Hu, Lydia Manikonda, and Subbarao Kambhampati. 2014. What we instagram: A first analysis of instagram photo content and user types. In *Eighth International AAI Conference on Weblogs and Social Media*.
- Da Ju, Kurt Shuster, Y-Lan Boureau, and Jason Weston. 2019. All-in-one image-grounded conversational agents. *arXiv preprint arXiv:1912.12394*.
- Chandra Khatri, Behnam Hedayatnia, Rahul Goel, Anushree Venkatesh, Raefer Gabriel, and Arindam Mandal. 2018. Detecting offensive content in open-domain conversations using two stage semi-supervision.
- Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.
- Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:11336–11344.
- Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019a. Visualbert: A simple and performant baseline for vision and language.
- Margaret Li, Jason Weston, and Stephen Roller. 2019b. ACUTE-EVAL: Improved dialogue evaluation with optimized questions and multi-turn comparisons. In *NeurIPS workshop on Conversational AI*.
- Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2019. Does gender matter? towards fairness in dialogue systems.

- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. [Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks](#). In *Advances in Neural Information Processing Systems 32*, pages 13–23. Curran Associates, Inc.
- Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Dhruv Mahajan, Ross B. Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. 2018. [Exploring the limits of weakly supervised pre-training](#). *CoRR*, abs/1805.00932.
- Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.
- Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. 2017. [Image-grounded conversations: Multimodal context for natural question and response generation](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 462–472, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2016. Dual attention networks for multimodal reasoning and matching. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2156–2164.
- Sungjin Park, Taesun Whang, Yeochan Yoon, and Hueiseok Lim. 2020. [Multi-view attention networks for visual dialog](#).
- Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. [To tune or not to tune? adapting pre-trained representations to diverse tasks](#). *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*.
- Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Htut, Xiaoyi Zhang, Richard Pang, Clara Vania, Katharina Kann, and Samuel Bowman. 2020. [Intermediate-task transfer learning with pretrained language models: When and why does it work?](#) pages 5231–5247.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. [Faster r-cnn: Towards real-time object detection with region proposal networks](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. [Imagenet large scale visual recognition challenge](#). *Int. J. Comput. Vision*, 115(3):211–252.
- Johannes Schäfer and Ben Burtenshaw. 2019. [Offence in dialogues: A corpus-based study](#). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1085–1093, Varna, Bulgaria. INCOMA Ltd.
- Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2020. Image-chat: Engaging grounded conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2414–2429.
- Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. 2019a. Engaging image captioning via personality. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Kurt Shuster, Da Ju, Stephen Roller, Emily Dinan, Y-Lan Boureau, and Jason Weston. 2019b. The dialogue dodecathlon: Open-domain knowledge and image grounded conversational agents.
- Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2020a. Mmf: A multimodal framework for vision and language research. <https://github.com/facebookresearch/mmf>.
- Amanpreet Singh, Vedanuj Goswami, and Devi Parikh. 2020b. [Are we pretraining it right? digging deeper into visio-linguistic pretraining](#).
- Eric Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents’ ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL.
- Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2020. [A novel attention-based aggregation function to combine vision and language](#).

- Hao Tan and Mohit Bansal. 2019. [LXMERT: Learning cross-modality encoder representations from transformers](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Yue Wang, Shafiq Joty, Michael R. Lyu, Irwin King, Caiming Xiong, and Steven C. H. Hoi. 2020. [Vd-bert: A unified vision and dialog transformer with bert](#).
- Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. [Recipes for safety in open-domain chatbots](#).
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.
- Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018a. [Conversations gone awry: Detecting early signs of conversational failure](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, Melbourne, Australia. Association for Computational Linguistics.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018b. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213. ACL.
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *ACL, system demonstration*.

## A Details of Image Encoders

We test the following image encoders in our MMB models:

**ResNeXt WSL** We first experiment with image representations obtained from pre-training a ResNeXt 32x48d model on nearly 1 billion public images (Mahajan et al., 2018), with subsequent fine-tuning on the ImageNet1K dataset (Russakovsky et al., 2015)<sup>4</sup>. The output of this model is a 2048-dimensional vector, and we refer to these representations as “ResNeXt WSL” features.

**ResNeXt WSL Spatial** One can also take the output of the image encoder prior to its final fully-connected layer to obtain “spatial” image features, resulting in a  $2048 \times 7 \times 7$ -dimensional vector. We explore results with these features as well, and refer to them as “ResNeXt WSL Spatial”.

**Faster R-CNN** Finally, we consider Faster R-CNN features (Ren et al., 2017), using models trained in the Detectron framework (Girshick et al., 2018); specifically, we use a ResNeXt-152 backbone trained on the Visual Genome dataset (Krishna et al., 2016) with the attribute head (Singh et al., 2020a)<sup>5</sup>. The Faster R-CNN features are  $2048 \times 100$ -dimensional representations, and we refer to these features as “Faster R-CNN”.

## B Dataset Descriptions

**ConvAI2** The ConvAI2 dataset (Dinan et al., 2020b) is based on the Persona-Chat (Zhang et al., 2018b) dataset, and contains 140k training utterances in which crowdworkers were given prepared “persona” lines, e.g. “I like dogs” or “I play basketball”, and then paired up and asked to get to know each other through conversation.

**EmpatheticDialogues (ED)** The EmpatheticDialogues dataset (Rashkin et al., 2019) was created via crowdworkers as well, and involves two speakers playing different roles in a conversation. One is a “listener”, who displays empathy in a conversation while conversing with someone who is describing a personal situation. The model is trained to act like the “listener”. The resulting dataset contains 50k utterances.

**Wizard of Wikipedia (WoW)** The Wizard of Wikipedia dataset (Dinan et al., 2019c) involves

two speakers discussing a given topic in depth, comprising 194k utterances. One speaker (the “apprentice”) attempts to dive deep on and learn about a chosen topic; the other (the “wizard”) has access to a retrieval system over Wikipedia, and is tasked with teaching their conversational partner about a topic via grounding their responses in a knowledge source.

**BlendedSkillTalk (BST)** BlendedSkillTalk (Smith et al., 2020) is a dataset that essentially combines the three above. That is, crowdworkers are paired up similarly to the three previous datasets, but now all three “skills” (personalization, empathy, and knowledge) are at play throughout the dialogue: the speakers are tasked with blending the skills while engaging their partners in conversation. The resulting dataset contains 74k utterances.

**Image-Chat (IC)** The Image-Chat dataset (Shuster et al., 2020) contains 200k dialogues over 200k images: crowdworkers were tasked with discussing an image in the context of a given style, e.g. “Happy”, “Cheerful”, or “Sad”, in order to hold an engaging conversation. The resulting dataset contains over 400k utterances. For each conversation in the dataset, the two speakers are each assigned a style in which that speaker responds, and these styles are optionally fed into models as part of the input, alongside the dialogue context. There are 215 styles in total, and styles are divided into 3 categories, “positive”, “neutral”, and “negative”.<sup>6</sup>

## C Additional Training Details

**Domain-adaptive Pre-training** During domain-adaptive pre-training, we trained the model on 8 GPUs for 10k-30k SGD updates, using early-stopping on the validation set. The models were optimized using Adam (Kingma and Ba, 2014), with sweeps over a learning rate between  $5e-6$  and  $3e-5$ , using 100 warmup steps.

**Fine-tuning** In this stage, we train the models on 8 GPUs for around 10k train updates using a similar optimization setup as in the domain-adaptive pre-training stage.

<sup>4</sup>[https://pytorch.org/hub/facebookresearch\\_WSL-Images\\_resnext/](https://pytorch.org/hub/facebookresearch_WSL-Images_resnext/)

<sup>5</sup><https://github.com/facebookresearch/vilbert-multi-task>

Image Features	Image Fusion	COCO (ppl)	pushshift.io Reddit (ppl)	Average
COCO & pushshift.io Reddit training data				
ResNeXt WSL	Late	11.11	13.80	12.45
	Early	6.69	13.50	10.10
ResNeXt WSL Spatial	Late	7.43	13.00	10.22
	Early	6.53	13.46	10.00
Faster R-CNN	Late	5.26	13.17	9.21
	Early	5.23	13.15	<b>9.13</b>
COCO training data only				
ResNeXt WSL	Late	5.82	19.52	12.67
	Early	6.21	21.30	13.76
ResNeXt WSL Spatial	Late	6.51	16.50	11.51
	Early	6.19	18.77	12.48
Faster R-CNN	Late	5.21	17.88	11.55
	Early	<b>4.83</b>	18.81	11.82

Table 10: Model performance, measured via perplexity on validation data, on domain-adaptive pre-training datasets, comparing various image features and image fusion techniques. The top three rows involve multi-task training on COCO Captions and pushshift.io Reddit, while the bottom three rows involve single task training on COCO Captions only. We note that early fusion with Faster R-CNN features yields the best performance on COCO Captions.

## D Automatic Evaluations on Pre-Training Datasets

**Training Data** We first note that, regardless of image fusion and image feature choices, we see the best performance on COCO Captions by simply fine-tuning exclusively on that data. This is an expected result, though we do see that in nearly every scenario the decrease in perplexity is not large (e.g. 5.23 for Faster R-CNN early fusion multi-tasking, down to 4.83 with just COCO Captions).

**Image Features** Across all training setups, we see that using spatially-based image features (ResNeXt WSL Spatial, Faster R-CNN) yields better performance than just a single vector image representation (ResNeXt WSL). This difference is particularly noticeable when training with COCO and pushshift.io Reddit, where with Faster R-CNN features the model obtains an average ppl of 9.13 over the two datasets, while with ResNeXt WSL features the model only obtains 10.1 ppl. We find that using Faster R-CNN features additionally outperforms using ResNeXt WSL Spatial features, where using the latter obtains an average of 10.0 ppl over the two datasets.

**Image Fusion** Finally, holding all other variables constant, we find that using our early fusion scheme yields improvements over using a late fusion scheme. E.g., with Faster-R-CNN features in

the COCO-only setup, we see a decrease in perplexity from 5.21 to 4.83; with ResNeXt WSL Spatial image features, we see perplexity differences ranging from 0.3 to 0.9 depending on the training data.

## E Ablation Results on Fine-Tuned Datasets

**Text-Only Datasets** First, we look at the performance of our models on the text-only datasets. The second-to-last column in Table 11 shows the average perplexity across the text-only datasets. If we compare the model that performs best on Image-Chat across all sets of image features (Faster-R-CNN features with BST<sup>+</sup> + IC + COCO + Reddit training data with early fusion) to the model in row 2, which is trained both without images and without Image-Chat on the text-only datasets, we see that the perplexity differences are quite small: that is, including training on an image-dialogue dataset, and overloading the Transformer encoder/decoder to incorporate image features, *does not* hinder dialogue performance.

**Training Data** Across all image-feature choices, we see that the choice of training data indeed makes a difference in performance on Image-Chat. Examining the early fusion model in Table 11, by including COCO Captions (and, in some cases, pushshift.io Reddit) in the training data we see drops in perplexity from 12.99 to 12.85, 13.02 to 12.87, and 12.43 to 12.36 with ResNeXt WSL, ResNeXt WSL Spatial, and Faster R-CNN features respectively. The decrease in perplexity indicates

<sup>6</sup>Lists of positive, neutral, and negative styles are from <http://ideonomy.mit.edu/essays/traits.html>, following Shuster et al. (2019a).

Image Features	Training Data	Image Fusion	ConvAI2	ED	WoW	BST	IC 1st Turn	IC	Text Avg.	All Avg.
None	None	None	12.31	10.21	13.00	12.41	32.36	21.48	11.98	13.88
	BST <sup>+</sup>		8.74	8.32	8.78	10.08	38.94	23.13	8.98	14.76
	BST <sup>+</sup> + IC		8.72	8.24	8.81	10.03	16.03	13.21	<b>8.95</b>	9.83
ResNeXt WSL	BST <sup>+</sup> + IC	Late	8.71	8.25	8.87	10.09	16.20	13.27	8.98	9.84
	BST <sup>+</sup> + IC	Early	8.80	8.32	8.79	10.17	15.16	12.99	9.02	9.81
	BST <sup>+</sup> + IC + COCO + Reddit	Late	9.27	8.87	9.45	10.74	17.56	14.44	9.58	10.56
	BST <sup>+</sup> + IC + COCO + Reddit	Early	9.34	8.90	9.48	10.78	15.87	13.88	9.62	10.48
	BST <sup>+</sup> + IC + COCO	Late	8.79	8.36	9.00	10.21	16.00	13.31	9.09	9.93
	BST <sup>+</sup> + IC + COCO	Early	8.91	8.38	8.99	10.29	14.64	12.85	9.14	9.88
ResNeXt WSL Spatial	BST <sup>+</sup> + IC	Late	8.71	8.24	8.88	10.10	15.39	13.02	8.98	9.78
	BST <sup>+</sup> + IC	Early	8.79	8.29	8.92	10.15	15.34	13.02	9.04	9.83
	BST <sup>+</sup> + IC + COCO + Reddit	Late	8.76	8.31	8.88	10.14	15.20	13.04	9.02	9.83
	BST <sup>+</sup> + IC + COCO + Reddit	Early	9.30	8.82	9.46	10.76	15.67	13.79	9.56	10.43
	BST <sup>+</sup> + IC + COCO	Late	8.73	8.31	8.87	10.13	15.04	12.98	9.01	9.84
	BST <sup>+</sup> + IC + COCO	Early	8.81	8.34	8.99	10.22	14.76	12.87	9.09	9.80
Faster R-CNN	BST <sup>+</sup> + IC	Late	8.70	8.24	8.92	10.07	13.97	12.48	8.98	<b>9.68</b>
	BST <sup>+</sup> + IC	Early	8.81	8.33	8.81	10.15	13.66	12.43	9.03	9.71
	BST <sup>+</sup> + IC + COCO + Reddit	Late	8.75	8.31	8.93	10.14	13.83	12.49	9.03	9.73
	BST <sup>+</sup> + IC + COCO + Reddit	Early	8.78	8.31	8.85	10.15	<b>13.51</b>	<b>12.36</b>	9.02	9.69
	BST <sup>+</sup> + IC + COCO	Late	8.74	8.33	8.87	10.13	13.85	12.51	9.02	9.72
	BST <sup>+</sup> + IC + COCO	Early	8.81	8.34	8.93	10.19	13.57	12.39	9.07	9.73

Table 11: Ablation analysis of the impact of various image features, training data (including domain-adaptive pre-training), and image fusion techniques on the datasets described in Section 4.2, where BST<sup>+</sup> refers to the four text-only dialogue datasets (ConvAI2, ED, WoW, and BST). The numbers shown are model perplexities measured on each of the datasets’ validation data. Performance on the first turn of Image-Chat is also measured to highlight model performance when only given visual context. We note that using Faster R-CNN image features results in the best average performance, as well as the best performance on Image-Chat.

Image Features	Training Data	Image Fusion	IC First Turn	IC
None	None	None	32.36	21.48
	Image Chat		28.71	13.17
ResNeXt WSL	IC	Late	14.80	12.83
	IC	Early	16.00	13.21
	IC + COCO + Reddit	Late	16.73	13.92
	IC + COCO + Reddit	Early	15.71	13.53
	IC + COCO	Late	14.70	12.95
	IC + COCO	Early	14.62	12.92
ResNeXt WSL Spatial	IC	Late	15.34	13.01
	IC	Early	15.27	13.00
	IC + COCO + Reddit	Late	15.09	12.95
	IC + COCO + Reddit	Early	15.55	13.50
	IC + COCO	Late	15.02	12.95
	IC + COCO	Early	14.62	12.87
Faster R-CNN	IC	Late	13.99	12.51
	IC	Early	13.76	12.42
	IC + COCO + Reddit	Late	13.75	12.43
	IC + COCO + Reddit	Early	<b>13.44</b>	<b>12.29</b>
	IC + COCO	Late	13.82	12.48
	IC + COCO	Early	13.56	12.37

Table 12: Ablation analysis of the impacts of various image features, training data (including domain-adaptive pre-training), and image fusion techniques when training on the Image-Chat dataset alone (i.e., ignoring the text-only dialogue datasets). As in Table 11, we note that Faster R-CNN features yield the best results on Image-Chat.

that domain-adaptive pre-training indeed improves performance on Image-Chat. This difference is highlighted even more when we measure performance on the first turn of Image-Chat, in which the model must generate a response given no textual context: 15.16 to 14.64, 15.34 to 14.76, and 13.66 to 13.51. We note a similar trend in Table 12.

**Image Features** Again, we see that using Faster R-CNN features leads to dramatic improvements compared to using the ResNeXt WSL features (spatial or otherwise), yielding 12.36 perplexity on Image-Chat compared to 12.85 and 12.87 perplexity with ResNeXt WSL (non-spatial and spatial respectively) during multi-tasking, and 12.29 perplex-

Dataset	PPL	F1	B	R
Image-Chat	12.64	13.14	0.418	18.00
Image-Chat round 1	13.56	11.96	0.411	16.72
BlendedSkillTalk	9.98	17.84	0.980	19.25
ConvAI2	8.78	18.41	1.080	22.64
EmpatheticDialogues	8.46	19.23	1.448	24.46
Wizard of Wikipedia	8.82	18.63	2.224	17.39

Table 13: Test results of best multi-task model on BST+ and Image Chat datasets, measured via perplexity (ppl), F1, BLEU-4 (B), and ROUGE-L (R) scores. ConvAI2 results are reported on the validation set, as the test set is hidden. Wizard of Wikipedia results are on the "test seen" data split.

	MMB Style	MMB Degen	BlenderBot
Contradiction	2.15%	3.92%	3.37%
Improper English	0.27%	0.28%	0.26%
Repetitive	1.34%	2.24%	1.55%
Unrelated	2.42%	0.28%	2.33%
Non-Sensical	4.03%	3.36%	2.07%
None (All Good)	91.13%	90.48%	91.45%
Mean engagingness	<b>4.70±0.60</b>	4.59±0.61	<b>4.70±0.60</b>

Table 14: Per-turn annotations and mean engagingness ratings of human/model conversations **without images**, for MMB Style, MMB Degenerated, and the original BlenderBot. MMB Style and BlenderBot perform roughly equivalently on these metrics, with a small drop from degenerating. Ranges given are plus/minus one standard deviation.

ity on Image-Chat compared to 12.92 and 12.87 respectively for single-task training on Image-Chat (see Table 12).

**Image Fusion** Finally, we note as before that using our early fusion technique improves performance on Image-Chat across all ablation regimes. While the average perplexity across the dialogue datasets is best when using late image fusion, we obtain the best image chat perplexity when performing early image fusion.

## F Final Test Results

## G Additional Human Evaluations

### G.1 Ratings of Conversations on Engagingness and Single-Turn Issues

In Section 5.2.1, we discuss the collection of conversations between a human and a dialogue model, either BlenderBot or MMB Style. After every model response, the human records if the response contains any one of a number of different issues. Finally, at the end of the conversation, the human gives a 1-to-5 Likert-scale rating of the model’s overall engagingness. Table 14 shows that humans flag the models’ responses at comparable rates for

	MMB Style	Dodeca	2AMMC
Mean engagingness	<b>3.82±1.32</b>	3.00±1.60	2.73±1.46
Mean humanness	<b>3.80±1.23</b>	2.87±1.37	2.42±1.18
Mean image response	<b>3.66±1.32</b>	3.35±1.23	2.85±1.33

Table 15: Mean ratings of engagingness, humanness, and ability to talk about the image well, recorded at the end of human/model conversations **with images**. MMB Style performs better than the Dodeca and 2AMMC models on all metrics. Ranges are as in Table 14.

most categories of issues, with BlenderBot being flagged slightly more often for contradictions and repetitiveness and MMB Style flagged more often for being non-sensical; however, the mean engagingness rating of the two models across conversations is the same (both 4.7 out of 5). Degenerating the MMB model (MMB Degenerated) results in a slight drop in engagingness vs. no degenerating (Table 14).

Similar human ratings at the end of conversations between a human and a model about an image show that MMB Style beats Dodeca and 2AMMC on measures of engagingness, humanness, and the ability to talk about an image by a large margin (Table 15).

### G.2 Reasons for ACUTE-Eval Ratings

For ACUTE-Evals comparing pairs of human/model conversations from different models, crowdsource workers are asked to select among 10 checkboxes to explain their preference for one conversation over another. Workers are able to select multiple checkboxes. Results for ACUTE-Evals on the preference metric are shown in Tables 16, 17, and 18.

### G.3 ACUTE-Evals on the models’ first response to an image

On ACUTE-Evals comparing two models’ initial responses to the same image, we find that crowdsource raters choose both the MMB Style and 2AMMC models’ responses significantly more often than those of Dodeca (Table 19). We also find no significant difference in the rate at which MMB Style image responses are chosen compared to the same model fine-tuned only on Image-Chat and not on dialogue datasets (Table 20), which implies that multitasking on dialogue datasets does not degrade the ability to effectively respond to an image. See Table 21 for additional example responses of models to images.



	MMB Style	MMB Degendered	BlenderBot
Contradicts themselves less	11%	15%	15%
Better English	27%	30%	35%
Repeats themselves less	11%	6%	6%
More on-topic	27%	29%	34%
Makes more sense	27%	37%	32%
More detailed / less vague	20%	19%	18%
More knowledgeable	27%	27%	23%
Better listener / more inquisitive	32%	36%	28%
More entertaining/witty/thoughtful	30%	17%	14%
Other	0%	3%	2%

Table 16: Fraction of the time that crowdsource workers select a particular reason for choosing one human/model conversation over another when comparing MMB variants with BlenderBot during ACUTE-Evals on the preference metric. Conversations do not include images.

	MMB Style	DialoGPT std. beam	DialoGPT min beam 20	Meena
Contradicts themselves less	8%	9%	14%	11%
Better English	32%	53%	46%	37%
Repeats themselves less	13%	3%	13%	13%
More on-topic	37%	33%	43%	32%
Makes more sense	47%	38%	47%	43%
More detailed / less vague	35%	17%	34%	34%
More knowledgeable	33%	28%	30%	25%
Better listener / more inquisitive	34%	17%	25%	29%
More entertaining/witty/thoughtful	17%	14%	21%	20%
Other	1%	2%	1%	1%

Table 17: Fraction of the time that crowdsource workers select a particular reason for choosing one human/model conversation over another when comparing MMB Style to existing text-only models during ACUTE-Evals on the preference metric. Conversations do not include images. Models and generation parameters are as in Table 4.

## H Additional Analyses of Safety and Gender Bias

See Table 22 for a table of perplexities of all MMB model variants. Table 23 displays measurements of safety in the second rounds of Image-Chat conversations dependent on whether first round exhibited a positive or negative style.

### H.1 Analyzing Dependence on Image

We also train a no-image ablation model, otherwise equivalent to MMB Positive, for which Image-Chat images are removed during both training and inference: crowdsource workers prefer the image responses of MMB Positive to those of this ablation model 80% to 20% (Table 24). For this ablation, style was removed from the context (replaced with the string “positive/neutral”) to prevent the ablation model from being aided by this information.

## I Example Conversations and Failure Modes

Figure 1 in particular demonstrates a successful conversation: the model is clearly able to interpret what is in the image (a teddy bear and a road), and it is able to thoughtfully and creatively combine

these two subjects in the conversation for several turns. Figure 2 provides several more example conversations: in all of these, the model is able to both discuss the image and use it as a catalyst for further conversation, although occasionally with contradiction and forgetfulness issues as seen in Roller et al. (2020). (For instance, the model contradicts itself on whether it has any pets and forgets who is planning to make a fancy dinner.)

Last, we show a few hand-picked examples of poor conversations in Figure 3: in these, the model fails to identify the contents of the images, identifying them both as buildings, although this may reflect a difference in the prevalence of (for example) buildings vs. roller coasters in the training sets. Despite the human nudging the model about what the images actually convey, the model does not demonstrate that it has corrected its initial misidentification in later turns. This could perhaps be remedied by an increase in image training data, by further advancements in the integration of image features with this BlenderBot-based sequence-to-sequence model, or perhaps by training specifically on data in which one partner learns about the contents of an image over time.

	MMB Style	Dodeca	2AMMC
Contradicts themselves less	9%	8%	11%
Better English	30%	38%	33%
Repeats themselves less	10%	16%	8%
More on-topic	33%	31%	31%
Makes more sense	45%	31%	54%
More detailed / less vague	32%	22%	16%
More knowledgeable	30%	28%	20%
Better listener / more inquisitive	25%	18%	24%
More entertaining/witty/thoughtful	20%	17%	18%
Other	1%	1%	1%

Table 18: Fraction of the time that crowdsource workers select a particular reason for choosing one human/model conversation over another when comparing MMB Style to other multi-modal models during ACUTE-Evals on the preference metric. Conversations are started by the model responding to an image. Models and generation parameters are as in Table 6.

		Loss %		
		MMB	Dodeca	2AMMC
Win %	MMB Style		65*	49
	Dodeca	35*		39*
	2AMMC	51	61*	

Table 19: ACUTE-Evals on the image-response metric show that MMB Style and 2AMMC significantly outperform Dodeca fine-tuned on Image-Chat. ACUTE-Evals are measured on the models’ first response to an image only.

MMB Style	
Multi-task	vs. FT Image-Chat
48	52

Table 20: ACUTE-Evals show no significant difference on the image-response metric for MMB Style vs. an equivalent model only fine-tuned on Image-Chat and no dialogue datasets. ACUTE-Evals are measured on the models’ first response to an image.



I would love to climb to the top of that building and look out over the valley below.

Yeah, it's quite a panorama!

I'd love to do that too, but I don't think I'd be able to afford it.

Have you gone on vacation recently?

No, I haven't been on a vacation in a long time. I'd like to go to Hawaii though.

Yeah, that'd be great! I've never been there.

I've always wanted to go there as well. I think it would be a great place to visit.

Do you like snorkeling and scuba diving?

I love both of those activities. I would like to do it on a cruise ship. How about you?

Ha, I don't think I've ever done them, but they'd be fun :) A cruise ship might be cool, but I wonder if it'd get boring

I'm not sure if it would get boring or not. It would be nice to get away from the city for a while though.

Yeah, that's true - it's good to get a change of pace

That's true. A change of scenery is always nice. Do you have any hobbies?



I would love to have a cat like this. They are so cute and cuddly.

Yes, I would take care of the cat so well! What would you name it?

I would name it Felix, because it looks like a felis catus.

Haha indeed I think it is a felis catus! That's a nice name

Thank you! I think so too. Felis is Latin for feline.

I didn't know that. Do you have any pets?

I don't have any right now, but I'd love to get a dog or a cat in the future. How about you?

Hmm, I don't, although I had goldfish growing up (but I don't think that really counts)

I have a dog and a cat. I love them both. Do you like to read?

Yeah, when I get a chance! Both fiction and non-fiction are great. What about you?

I love to read! I'm a big fan of historical fiction. What's your favorite genre?

Oh, I love historical fiction. Hmm, I like books on psychology and society

I like fiction as well. My favorite author is Stephen King. Have you read any of his books?



This looks like it would be a great meal to have after a long hard day at work.

Absolutely! It's making me hungry just looking at it

I can't wait to try it! I'm going to make it for dinner tonight. I'll let you know how it turns out.

Cool - lemme know! The seasoning on the meat must be hard to get right

I'm looking forward to hearing how it turned out for you. I'm sure it will be delicious!

Ha, I don't think I could make anything like this - this seems difficult

I'll be sure to post a picture of it when I'm done. It's going to be delicious.

What do you usually like to cook?

I love to cook Italian and Mexican dishes. I also love to bake.

Oh that sounds great! Yeah, I'd love to try baking but my kitchen is tiny

I have a small kitchen as well, so I usually bake in the oven. It is a lot of fun.



These flowers are so beautiful, I would love to have a garden like this one day.

Yeah, me too! What kinds of flowers would you plant?

I'm not sure, I'd have to do a lot of research to find out what would grow well in my area.

That's a good point - honestly, I'm not sure what would grow well here. Do you have a big backyard for gardening?

I live in the suburbs, so I don't have a huge yard, but I do have a small garden in my front yard.

Oh, that's great. Would you grow vegetables as well or just flowers?

Vegetables, herbs, flowers, fruits and nuts are all things I grow in my garden.

That would be nice! I'd love to have apple trees and then eat fresh apples every fall.

I would grow all of the above. I love fresh fruit and veggies.

Do you like coconuts? People either love or hate them, I think.

I've never had a coconut. I've heard they're good for you though.

Figure 2: Randomly picked author examples. Paper author (right speaker) talking to the MMB DegenPos model (left speaker). Conversations are mostly fluent, with occasional contradictions.







Image	Output
	<p>(Style) <i>Fearful</i></p> <p><b>MMB</b> I hope that dog doesn't bite me. I don't want to get bitten by a dog.</p> <p><b>Dodeca</b> I hope that dog doesn't bite me.</p> <p><b>2AMMC</b> This dog looks ready to attack. It's focused on something, maybe another animal? I hope it doesn't go after it or anyone around!</p>
	<p>(Style) <i>Playful</i></p> <p><b>MMB</b> I would love to climb to the top of that sculpture and play in the fountain.</p> <p><b>Dodeca</b> I would love to visit this place!</p> <p><b>2AMMC</b> Who wants to put bubble bath in the fountain with me?</p>
	<p>(Style) <i>Pompous (Self-important, Arrogant)</i></p> <p><b>MMB</b> I have a statue just like this one in my backyard, but it's much nicer.</p> <p><b>Dodeca</b> I could take a better picture than that.</p> <p><b>2AMMC</b> I could make a fence that looks way better</p>
	<p>(Style) <i>Empathetic</i></p> <p><b>MMB</b> These flowers remind me of my grandmother's garden. I miss her so much.</p> <p><b>Dodeca</b> I feel bad for the flowers.</p> <p><b>2AMMC</b> I can't imagine how difficult it is for the farmer to round them all up. He works so hard.</p>
	<p>(Style) <i>Narcissistic (Self-centered, Egotistical)</i></p> <p><b>MMB</b> I would never go to a beach like this. My beach is so much nicer.</p> <p><b>Dodeca</b> I could take a better picture than that.</p> <p><b>2AMMC</b> I could swim out to those rocks. The current is strong but I am a stronger swimmer.</p>
	<p>(Style) <i>Boisterous (Rowdy, Loud)</i></p> <p><b>MMB</b> I would love to jump in that puddle and play in the mud! It would be so fun!</p> <p><b>Dodeca</b> This looks like a great place to party!</p> <p><b>2AMMC</b> I would ride through here on a loud bike and throw a party!</p>
	<p>(Style) <i>Happy</i></p> <p><b>MMB</b> I can't wait to go to the beach and fly a kite with my best friend!</p> <p><b>Dodeca</b> What a beautiful day!</p> <p><b>2AMMC</b> Face full of wind, smell of salt water, cool of the ocean, and fantastic memories being made!</p>

Table 21: Additional example outputs from MMB Style, Dodeca fine-tuned on Image-Chat, and 2AMMC, on images from Shuster et al. (2020). Image-Chat styles fed to the models are shown above the models' responses.



I would love to go on a tour of this building. I bet it has a lot of history.

I think it does! It's a very old coaster

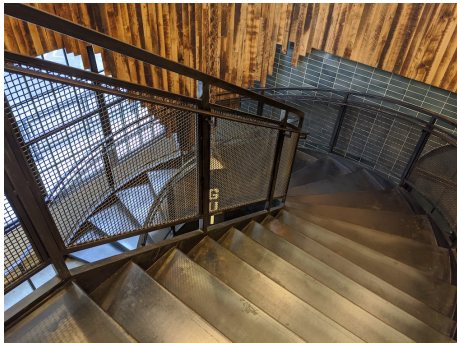
I'd love to see it in person as well. It looks like it's been around for a very long time.

Yeah, 100 years I think

Wow, that's a long time for a coaster. I wonder how many people have ridden it.

Haha too many! I've heard it's rickety but people like that

Yeah, I'm sure it's a popular attraction. It's located in the heart of downtown Los Angeles.



I would love to be able to climb to the top of that building and look out over the city.

How do you know it's a tall building?

I don't know if it's tall or not, but I'm sure it's expensive to maintain.

Yeah those stairs look really clean!

It's taller than the Empire State Building, which is the tallest building in the US.

Hm I'm not so sure - I've been there and I think it's just a few floors

I know it is tall because I have seen it in person. I have been to the empire state building as well.

Figure 3: **Lemon-picked author examples.** Paper author (right speaker) talking to the MMB DegenPos model (left speaker): misidentifying the subject of the image (top); misidentifying the subject of the image and not being able to learn from the chat partner’s feedback (bottom).

	BST	Conv	ED	WoW	IC	Avg
Style	10.15	8.78	8.31	8.88	12.36	9.70
Degen	10.14	8.76	8.21	9.01	12.58	9.74
Pos	10.15	8.76	8.27	8.95	12.55	9.74
DP	10.36	8.97	8.34	9.41	12.65	9.95

Table 22: Perplexities of MMB Style, MMB DegenDer, MMB Positive, and MMB DegenPos on the validation set. For Image-Chat, styles are used in the context for all models, for consistency. (MMB Positive and MMB DegenPos observed styles for 25% of Image-Chat examples during training.)

	Style	Pos C	Pos B	Neg C	Neg B
Style	Cheerful	2.41	0.00	3.81	0.09
	Relaxed	3.87	0.00	6.47	0.09
	Angry	67.07	0.22	62.62	0.27
	Cruel	77.57	1.42	73.67	0.83
Dgen	Cheerful	1.50	0.00	3.19	0.09
	Relaxed	2.55	0.00	4.43	0.04
	Angry	53.90	0.33	51.64	0.31
	Cruel	58.28	0.95	57.00	0.84
Dgen Pos	Pos/Neu	7.00	0.00	12.98	0.22
	Negative	30.96	0.22	31.05	0.09
	Pos/Neu	4.56	0.04	8.86	0.18
	Negative	25.86	0.26	25.42	0.27

Table 23: Toxicity of MMB variants as assessed with different control variables. We evaluate on the second round of the Image-Chat validation set. Column “Pos C” shows the safety classifier metric when conditioning on a positive style for the round-1 utterance, and “Pos B” shows the same thing for the blocklist metric. The following two columns show the same metrics when the round-1 utterance has a negative style.

MMB Positive  
With image vs. Without image

80\*

20\*

Table 24: ACUTE-Evals show that the MMB Positive model is significantly better at responding to an image than an equivalent model not shown any images during training or inference.