

# How do people talk about images? A study on open-domain conversations with images.

**Yi-Pei Chen**

The University of Tokyo  
ypc@g.ecc.u-tokyo.ac.jp

**Nobuyuki Shimizu**

Yahoo Japan  
nobushim@yahoo-corp.jp

**Takashi Miyazaki**

Yahoo Japan  
takmiyaz@yahoo-corp.jp

**Hideki Nakayama**

The University of Tokyo  
nakayama@ci.i.u-tokyo.ac.jp

## Abstract

This paper explores how humans conduct conversations with images by investigating an open-domain image conversation dataset, ImageChat. We examined the conversations with images from the perspectives of *image relevancy* and *image information*. We found that utterances/conversations are not always related to the given image, and conversation topics diverge within three turns about half of the time. Besides image objects, more comprehensive non-object image information is also indispensable. After inspecting the causes, we suggested that understanding the overall scenario of image and connecting objects based on their high-level attributes might be very helpful to generate more engaging open-domain conversations when an image is presented. We proposed enriching the image information with image caption and object tags based on our analysis. With our proposed *image<sup>+</sup>* features, we improved automatic metrics including BLEU and Bert Score, and increased the diversity and image-relevancy of generated responses to the strong SOTA baseline. The result verifies that our analysis provides valuable insights and could facilitate future research on open-domain conversations with images.

## 1 Introduction

A picture is worth a thousand words. Human communication often involves both text and images. Understanding the image content and chatting about it is crucial for a chatbot to interact with people. Current multimodal dialogue systems often equip with an object detector, and adapt similar architecture as text-based dialogue systems, except fusing text and image modalities through concatenation (Shuster et al., 2020b,c) or an attention mechanism (Ju et al., 2019).

To investigate whether an additional object detector is enough, and to understand what factors direct the conversation content when an image exists, we conducted a deep analysis of the ImageChat

dataset (Shuster et al., 2020a). We aimed to answer the following questions: (1) How is a conversation with image different from an open-domain conversation? Is the image necessary or supplemental? How related is the image to the conversation topic? (2) Does the topic of the three-turns conversation always be on the image? How does the transfer happen if the conversation topic transits from the image to others? Can we predict the shift from the image? (3) What types of image information are used in the conversation? More specifically, we want to know how helpful image objects are in the conversation since baseline models usually use an object detection model as the image encoder.

We addressed the questions by sampling and analyzing ImageChat dataset from the aspects of *image relevancy* and *image information*, which are independent but intertwined. The former labels whether the given image is relevant to the conversation theme, and the latter marks the type of image information in utterances. The annotation results show that about 31% of utterances are not on the image-related theme, i.e., the utterances do not describe or could be generated without the image. In terms of the conversation, people transit conversation topics 54% of the time within three utterances, and surprisingly 7% of conversations entirely consist of non-image-related utterances. In these conversations where the image is optional, the topic often derived from attributes of one of image objects. We also discovered that 45% of utterances contain image objects, 23.7% have non-object image information such as the description of events in the image, and 31.3% do not have any image information at all. This result hints that including comprehensive descriptions of the image beyond image objects could benefit the generation of utterances with image information.

Based on our analysis, we propose to enhance the generation of image-dependent response by augmenting image features from image caption




Image	Style	Utterance	Related
	<i>Cowardly:</i>	Never had this food before and not sure if I’m ready to try it today.	✓
	<i>Grateful:</i>	I am always up to trying new things. It looks like a lot of effort went into this food and I plan to enjoy every bite.	✓
	<i>Cowardly:</i>	I don’t know, it looks like it might be too much.	✓
	<i>Extraordinary:</i>	What an unusual place! The colors of the train really bounce off the grey backdrop of the city.	✓
	<i>Narcissistic:</i>	Well, of course this is a fantastic picture, since it was MY magnificent photographic skills that produced it!	✓
	<i>Extraordinary:</i>	I had no idea you have such talent!	✗
	<i>Spontaneous:</i>	That’s it, I going to Vegas tomorrow. Who’s coming with me?	✗
	<i>Morbid:</i>	Someone died in that Vegas spot.	✗
	<i>Spontaneous:</i>	Lets go on a vegas trip this weekend!	✗

Table 1: Examples of conversation themes are related and unrelated to the given image.

and object tags, and using the text information explicitly rather than fusing image captioning and object detection models to the text-based conversation model. Our model with enhanced image features outperforms the strong SOTA model Multimodal BlenderBot (MMB) (Shuster et al., 2020c) on BLEU and BertScore. In addition, we also generate more image-related and more diverse responses than MMB.

## 2 Analysis of Conversations on Image

### 2.1 ImageChat Dataset

We analyzed the ImageChat dataset (Shuster et al., 2020a), which is so far the only dialogue dataset that focuses on *open-domain conversations on images*, to the best of our knowledge. Each conversation is paired with one image from YFCC 100M (Thomee et al., 2016) and consists of three turn utterances from two speakers with assigned speaking styles. There are total 215 style types, such as sympathetic or optimistic. The images are highly diverse ones across multiple domains. We obtained the object tags by Scene Graph Benchmark (Han et al., 2021) implementation of Faster R-CNN (Ren et al., 2016), which is also the image encoder used in the baseline model MMB. We also generated the caption of each image using the SOTA language-vision pretrained model VinVL (Zhang et al., 2021).

### 2.2 Annotation

We randomly sampled 300 utterances (100 conversations) from the validation set and annotated each utterance for its image relevancy and what image information it contains.

#### 2.2.1 Image Relevance to Dialogue Theme

We first asked whether the *conversation theme* is always related to the image, and if not, how often is each utterance directly related to the image. We defined *image relevancy* as a binary classification of whether the given image is necessary for generating each utterance. If one could generate the utterance without the given image, the utterance is labeled as unrelated. Examples of image-related and unrelated utterances are shown in Table 1.

#### 2.2.2 Image Information in the Dialogue

Based on our observation of the data, we categorized each utterance into one of the 8 classes, indicating the type of image information mentioned in the utterance. Classes start with **O** mean image objects are mentioned in the utterance; classes start with **R** mean there are non-object image related information mentioned in the utterance; and **NI** class means there is no image information in the utterance at all. See Table 2 for the details and examples of each category.

Class	Explanation	Utterance (U)   Object Tags (T)
O	Words in utterance <b>exactly match</b> object tags	U: I guess this is an interesting <b>building</b> . T: ['cloud', 'window', 'sky', ' <b>building</b> ']
OS	<b>Synonyms</b> of object tags in the utterance, including hyponym/hypernym pairs, e.g. "seagull" in U and "bird" in T.	U: I'd like to party with that <b>guy</b> ! T: ['watch', ' <b>man</b> ', 'phone', 'guitar', ...]
OP	<b>Pronoun</b> is used to refer to image objects.	U: Would <b>she</b> shut up already? T: ['book', 'jacket', 'tree', ' <b>woman</b> ', ...]
OF	Words in the utterance refer to image objects but have no overlap with object tags, probably due to <b>false object detection</b> results.	U: The <b>aluminum</b> art was different. T: ['rock', 'ground', 'foil']
R	Words in the utterance referring to <b>non-object image information</b> , e.g., the scene of the image.	U: It's obviously a <b>festival</b> . T: ['sunglasses', 'hat', 'balloon', ...]
RI	The utterance is about the <b>image itself</b> , not the content of the image.	U: A <b>screenshot</b> by definition does not die. T: ['man', 'hat', 'photo', 'glass']
RP	<b>Pronoun</b> is used to refer to image-related information in the utterance.	U: <b>It's</b> beautiful! I would love to visit. T: ['leaf', 'flower', 'branch', 'tree']
NI	<b>No image-related</b> information mentioned in the utterance.	U: yeah sure does. T: ['sunglasses', 'hat', 'man', 'light', ...]

Table 2: Classes of image information in the utterance.

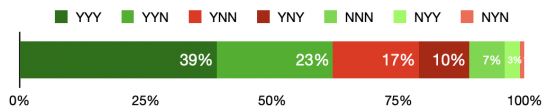


Figure 1: Different combination of image-related utterances in 3-turns dialogues. Y: image-related utterance; N: non-related utterance. Green hue indicates the dialogue is more image-dependent, and the red family suggests the opposite.

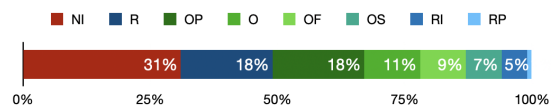


Figure 2: Classes of image information in utterances. Green hue refers to image objects, blue hue refers to non-object image information, and red means there is no image information at all.

## 2.3 Analysis Result and Finding

### 2.3.1 Image Relevancy

We found that conversation themes of ImageChat dialogues are not always about the image. In fact, the conversation often goes back and forth between image-related to non-related topics even within only three conversation turns. Figure 1 illustrates such a phenomenon with dialogues of different combinations of the image-relevance utterances. While an image-related utterance is labeled as 'Y' and non-image-related utterance is labeled as 'N', 'YYY' means all three turns in a dialogue are image-related utterances and 'YYN' means the conversation diverse from image-related topics to other domain not related to the given image.

Further investigating the combination of image-related and non-related utterances in a dialogue, we could roughly classify them into two schemas: (1) One speaker responds to the other, and if one extends out of the image-related topic, the following conversation is diverse, and vice versa. 'YNN', 'YYN', 'NYY' are in this category. The transition between 'Y' and 'N' may result from the mention of an object related to objects in the image but not related to the image itself. In this case, the related object often links to the image object with some high-level attributes, such as the object's category, shape, or material. Alternatively, the 'N' utterance might be a general non-informative response or an invented non-image-related scenario. (2) Some dialogues seem unnatural because one of the speakers

continues their previous (self-)expression and does not respond to the other’s utterance. ‘YNY’ and ‘NYN’ usually belong to this schema. Note that there is no combination of ‘NNY,’ showing that it is less likely to talk about the image after chatting on off-image topics.

We found that about 7% of dialogues are non-image-related (‘NNN’), although most utterances are still image-related (Y: 69% vs. N:31%). Investigating the reason, we noticed that many of the non-image-related dialogues are stimulated by attributes of one of objects in the image. For example, a conversation about fighting in a ring is given an image with a ring-shaped object. This observation suggests that capturing attributes of objects and linking objects to much broader scenarios are essential directions to generate natural utterances.

### 2.3.2 Image Information

Figure 2 shows the distribution of image information classes. The green hue represents the utterances with image objects (Ox, 45.0%). Among them, a great portion of utterances have objects referred by a pronoun (OP, 17.7%), 11.3% of utterances have the exact match of image objects (O), 9.3% contain objects not in the tag set (OF), and the rest of 6.7% have objects mentioned in synonyms (OS). While many objects are indicated by pronouns, linking the objects and their attributes to mentions in the utterance becomes a vital task for utterance generation.

On the other hand, the blue hue refers to the utterances with non-object image information (Rx, 23.7%), which usually describes the event, action, or scenario in the image. Thus, knowing the scene beyond the given objects is also important.

The rest of 31.3% of utterances represents in red are the class NI without any image information. These utterances are usually on the off-image theme and the only hint to reconstruct such utterances is from their conversational context.

## 3 Augmenting Image Information

Our analysis suggests the importance of the non-object image information, which is often the scene in the image. Therefore, we augmented the image feature by image caption to capture the scenario. We also found that explicitly using texts of objects tags and captions works better than fusing the latent vectors from captioning and object detection models. Given object tags, we replace the single full-image feature in the baseline model with several

image region features to facilitate the extraction of image object information.

## 3.1 Experiments

### 3.1.1 Settings

We ran our experiments on the ImageChat dataset (Shuster et al., 2020a) which is described in Sec 2.1. All our experiments are conducted using the ParlAI (Miller et al., 2017) framework. We compared with the SOTA multimodal dialogue system: Multimodal Blenderbot (MMB) (Shuster et al., 2020c).

We obtain image tags from Scene Graph Benchmark (Han et al., 2021) and the image caption from pretrained VinVL model (Zhang et al., 2021). The image feature dimension is set to 2054, with additional 6-dim image information such as weight and height to the 2048-dim FasterRCNN feature in the original model. Each image is paired with 1 to 10 unique tags, an image caption with maximum 12 tokens, and at most 32 image object features. All models are finetuned from the Reddit pretrained model, following the instruction from MMB<sup>1</sup>.

Following previous works, we reported the number of perplexity (PPL), Rouge-L, BLEU-4, and F1 score. As existing research has reported that these numbers are not highly correlated with human evaluation (Liu et al., 2016; Li et al., 2016), we also reported Bert Score (rescale) (Zhang\* et al., 2020), which reflects the semantics similarity instead of the token-wised matching. To show how relevant the generated response is to the image, we ran the image-text retrieval task using VinVL (Zhang et al., 2021). We also reported the number of average length, unique vocabularies, and Distinct-1 (Li et al., 2015) to show the diversity of utterances.

## 3.2 Results and Analysis

Table 3 demonstrates that our enhanced image features improve the strong baseline without training on many additional datasets. This result implies that *image*<sup>+</sup> provides much more useful information that neither additional text-only dialogue datasets (BST+) nor image captioning pretraining is needed. Besides, the result also suggests that a pipeline approach of explicitly adding image caption to the input is better than end-to-end training on the additional image captioning task.

<sup>1</sup>[https://github.com/facebookresearch/ParlAI/blob/main/parlai/zoo/multimodal\\_blenderbot/README.md](https://github.com/facebookresearch/ParlAI/blob/main/parlai/zoo/multimodal_blenderbot/README.md)



Model	Datasets	PPL	Rouge	BLEU	F1	Bert Score		
						P	R	F1
MMB	R,I,C,B	13.60	12.40	0.386	12.94	33.81	25.21	29.49
MMB	R,I,C	15.00	11.35	0.278	11.81	31.73	23.52	27.61
MMB	R,I	12.89	13.04	0.419	13.52	32.58	24.23	28.39
MMB + <i>image</i> <sup>+</sup>	R,I,C,B	<b>12.63</b>	<b>13.36</b>	0.447	13.75	34.76	26.36	30.54
MMB + <i>image</i> <sup>+</sup>	R,I	12.76	13.29	<b>0.461</b>	<b>13.82</b>	<b>35.36</b>	<b>26.38</b>	<b>30.85</b>

Table 3: We compare models pretrained on Reddit (R) (Baumgartner et al., 2020) and finetuned on different datasets such as COCO Captioning (C) (Chen et al., 2015), text-only dialogue datasets BST+(B) (Smith et al., 2020; Dinan et al., 2019a,b; Rashkin et al., 2019), and ImageChat (I). *image*<sup>+</sup> refers to our proposed enhanced image features (image caption and object tags).

Model	Image-to-Text		Text-to-Image		Length	Vocabs	Distinct-1
	R@1	R@10	R@1	R@10			
Gold	0.02	0.14	0.03	0.32	<b>9.90</b>	<b>9,431</b>	<b>0.064</b>
MMB	<b>0.04</b>	0.16	0.03	0.29	7.87	3,436	0.029
MMB + <i>image</i> <sup>+</sup>	<b>0.04</b>	<b>0.26</b>	<b>0.04</b>	<b>0.35</b>	8.04	3,865	0.032

Table 4: We evaluate how much the utterance is related to the image by image-text retrieval task. We also show the average length, vocabulary size, and diversity of utterances in the validation set. Gold refers to the reference utterances by human.

Feature	PPL	R	B	BS
Tags	13.9	12.26	0.325	30.29
Caption	13.8	12.33	0.373	30.20
Both	12.8	13.29	0.461	30.85

Table 5: Ablation results of MMB + *image*<sup>+</sup> trained on Reddit and ImageChat datasets. PPL: perplexity, R: Rouge, B: BLEU, BS: Bert Score

We also found that the Reddit pretraining is essential for dialogue generation. Without pretraining, the perplexity would boost to about 34, and all other metrics get much worse based on our empirical results. In fact, the perplexity is already around 26 at the very beginning of the training when finetuning on the Reddit pretrained model.

Our ablation experiment (Table 5) shows that the model with the caption feature has better Rouge and BLEU scores compared with the model only with tags, but the Bert Score is about the same. The result suggests that both tags and the caption can generate semantically equivalent utterances.

As shown in Table 4, we demonstrated our model’s superiority in generating more diverse and image-relevant responses. We got the best retrieval result in both image-to-text and text-to-image retrieval, which even outperforms the human refer-

ence, showing that our generated responses are the most relevant to the given image. We also generated longer sentences with more diverse vocabularies than the MMB baseline. We provided some example outputs from MMB and our MMB + *image*<sup>+</sup> in the Appendix.

## 4 Conclusions

In this paper, we analyzed the factors that influence open-domain conversations with images, from aspects of (a) image relevancy to the conversation theme and (b) image information in the conversation. According to our observations, open-domain conversations with images often branch off from one topic to another even within only three turns. The relation between the attributes of objects is the key to linking utterances with different themes. From the dynamics of image-relevancy, it is also interesting to notice that some conversation is a solo play where people just express themselves without responding. Moreover, a comprehensive view of the whole image and the understanding of the image scene are also critical image information in utterances, besides image objects. Therefore, we proposed incorporating image captions that could capture the overall image semantics beyond objects and may provide some hints to the links to other objects. We also found empirically that explicitly

using texts of caption and object tags work better than incorporating captioning and object detection models in latent space. With our enhanced image features *image*<sup>+</sup>, we outperformed MMB on BLEU, F1, and Bert Score, and generated more image-related and diverse conversation responses, confirming the effectiveness of our findings. We believe that our in-depth analysis and proposed findings would benefit the future research on the open-domain conversations with images.

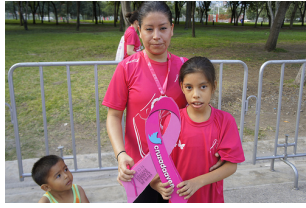
## Acknowledgements

We thank Hong Chen, Minh-Duc Vo, Noriki Nishida, and the anonymous reviewers for their valuable comments on this work.

## References

- 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.
- 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*.
- Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2019a. The second conversational intelligence challenge (convai2). *arXiv preprint arXiv:1902.00098*.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and J. Weston. 2019b. Wizard of wikipedia: Knowledge-powered conversational agents. *ArXiv*, abs/1811.01241.
- Xiaotian Han, Jianwei Yang, Houdong Hu, Lei Zhang, Jianfeng Gao, and Pengchuan Zhang. 2021. [Image scene graph generation \(sgg\) benchmark](#).
- Da Ju, Kurt Shuster, Y-Lan Boureau, and J. Weston. 2019. All-in-one image-grounded conversational agents. *ArXiv*, abs/1912.12394.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.
- Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*.
- Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.
- A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. 2017. Parlai: A dialog research software platform. *arXiv preprint arXiv:1705.06476*.
- 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. 2016. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149.
- Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2020a. [Image-chat: Engaging grounded conversations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2414–2429, Online. Association for Computational Linguistics.
- Kurt Shuster, Da Ju, Stephen Roller, Emily Dinan, Y-Lan Boureau, and J. Weston. 2020b. The dialogue dodecathlon: Open-domain knowledge and image grounded conversational agents. *ArXiv*, abs/1911.03768.
- Kurt Shuster, Eric Michael Smith, Da Ju, and Jason Weston. 2020c. Multi-modal open-domain dialogue. *arXiv preprint arXiv:2010.01082*.
- Eric Michael 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. *arXiv preprint arXiv:2004.08449*.
- Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.
- Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Making visual representations matter in vision-language models. *CVPR 2021*.
- Tianyi Zhang\*, Varsha Kishore\*, Felix Wu\*, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). In *International Conference on Learning Representations*.

## A Example Outputs



[1st Turn]	
Curious: What did they have to do to help support cancer? Fundraisers always intrigue me.	Irrational: Is she doing a backflip twist?
MMB: <u>What is that kid doing?</u>	MMB: <u>I don't think she's going to catch the ball</u>
Ours: <b>I wonder what they are doing.</b>	Ours: <b>I bet they are going to fall and hurt themselves.</b>
[2nd Turn]	
Patriotic: Maybe they donated. All Americans should donate to show that we are a proud and caring country.	Obnoxious: I don't think so. I think she is passing gas.
MMB: <u>They are fighting for our country.</u>	MMB: <u>She is doing a back flip.</u>
Ours: <b>They are supporting their country by raising money for cancer research.</b>	Ours: <b>I hope she falls and breaks her neck.</b>
[3rd Turn]	
Curious: They don't look american, I wonder if this is somewhere else?	Irrational: She needs to find a restroom quick!
MMB: <u>I wonder what they did to help.</u>	MMB: <u>She is definitely not passing gas.</u>
Ours: <b>I wonder what they did to help.</b>	Ours: <b>I think she is trying to fly.</b>

Table 6: Example conversations. The 1st row in each turn is the ground truth speaking style and utterance, 2nd and 3rd rows are utterances generated by MMB (underlined) and our MMB + *image*<sup>+</sup> (bold), given the speaking style, image, and ground truth utterance(s) in previous turn(s).