CHAMPAGNE: Learning Real-world Conversation from Large-Scale Web Videos

Seungju Han♠ Jack Hessel♡ Nouha Dziri♡ Yejin Choi♡♢ Youngjae Yu♡♣

♠ Seoul National University ♡ Allen Institute for Artificial Intelligence ♢ University of Washington ♣ Yonsei University

wade3han@snu.ac.kr

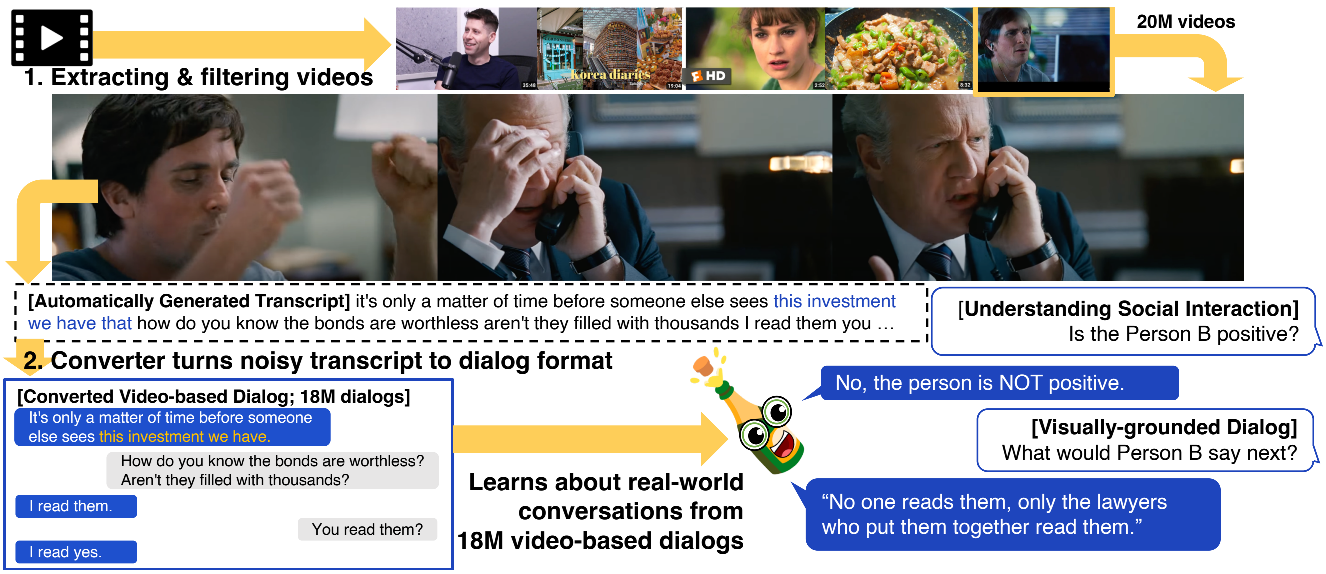

Figure 1: CHAMPAGNE is a generative model of real-world conversational frames trained on YTD-18M, a dataset of 18M video-based dialogues. YTD-18M is derived from public videos and their associated transcripts; a language model automatically grounds the conversation turns to the speakers. CHAMPAGNE attains state-of-the-art results on four competitive vision-language benchmarks after fine-tuning.

Abstract

Visual information is central to conversation: body gestures and physical behaviour, for example, contribute to meaning that transcends words alone. To date, however, most neural conversational models are limited to just text. We introduce CHAMPAGNE, *a generative model of conversations that can account for visual contexts. To train* CHAMPAGNE, *we collect and release* YTD-18M, *a large-scale corpus of 18M video-based dialogues.* YTD-18M *is constructed from web videos: crucial to our data collection pipeline is a pretrained language model that converts error-prone automatic transcripts to a cleaner dialogue format while maintaining meaning.*

Human evaluation reveals that YTD-18M *is more sensible and specific than prior resources (MMDialog [17], 1M dialogues), while maintaining visual-groundedness. Experiments demonstrate that 1)* CHAMPAGNE *learns to conduct conversation from* YTD-18M*; and 2) when fine-tuned, it achieves state-of-the-art results on four vision-language tasks focused on real-world conversations. We release data, models, and code at https://seungjuhan.me/champagne.*

1. Introduction

Conversation often relies on non-verbal cues: visual information like physical expressions, body gesture, or the surrounding environment are used by interlocutors to shape and understand meaning. Figure 1: the two conversation participants appear stressed (in the first image: one person strikes the desk with his fists; in the second, the other person is rubbing his face); but the tension is not apparent simply

Figure 2: Sample dialogues from two different datasets: (a) YTD-18M, and (b) the example from MMDialog presented in their paper [17]. According to the human evaluation results, our YTD-18M dataset has a good balance of both social interactions (left; 62%) and visually-grounded dialogue (right; 38%), in contrast, MMDialog only has 12% social interactions. In addition, MMDialog is derived from social media, resulting in examples that contain emojis or other social media-specific elements, while YTD-18M is derived from videos and thus genuinely capture real-life communications.

from the transcript, which refers mostly to financial topics. Other visual information suggests a broader context as well: that the conversation is taking place over the phone, that one person is older, and that one person is more formally dressed are all potentially important factors for understanding the conversation's meaning, yet none are reflected in the transcript. Indeed, visual perception provides information that can help machines understand the world in ways that text alone cannot [6, 28, 11].

In this paper, we propose 🍾 CHAMPAGNE[1], a generative model of conversations that learns from a large-scale video corpus. CHAMPAGNE takes in video frames, a video title, and a dialogue context as input and returns a dialogue response as output. The model learns from videos about two *conversational frames:* 1) *Social Interaction*, where the conversation is observed from a 3rd-person perspective (*e.g.* movies or interviews; Figure 2, (a)-left); and 2) *Visually-grounded Dialogue*, where the conversation is observed from an embodied, first-person perspective (*e.g.* egocentric videos or chit-chatting through messenger applications; Figure 2, (a)-right)).

To support training CHAMPAGNE, we collect and release a large-scale dataset, 🎥 YTD-18M, which is the largest publicly available dataset for real-world conversation learning. YTD-18M is constructed from 20M YouTube videos: we use a language model to convert the noisy transcripts automatically generated by YouTube into well-formatted dialogues associated with video frames. Human evaluation shows that YTD-18M covers both social interaction and visually-grounded dialogue frames in balance, and surpasses the prior resource in terms of quality and scale (see §3.2).

After training, we demonstrate that CHAMPAGNE models generate high-quality next-turn utterances that account for visual contexts. Then, we conduct fine-tuning experiments, finding that: 1) CHAMPAGNE exhibits strong performance on open-domain text-only conversation benchmarks (§4.1); 2) CHAMPAGNE outperforms existing SOTA models on two *social interaction* understanding benchmarks: CMU-MOSEI [62] and Visual Comet [38] (§4.2); 3) CHAMPAGNE outperforms SOTA models on two *visually-grounded dialogue* benchmarks: Visual Dialog [12] and Image Chat [50] (§4.3); and 4) ablations confirm the importance of various components of YTD-18M (§4.4), *e.g.* video frames.

In summary, our main contributions are:

1. 🎥 YTD-18M, a large-scale dataset that contains 18M video-based dialogue that covers real-world conversational frames derived from 20M web videos.

2. 🍾 CHAMPAGNE, a generative model that learns about real-world conversations from YTD-18M without any manual annotation.

3. Experiments and ablations that demonstrate learning from a large-scale video-based dialogue dataset improves model performance on various tasks related to conversation.

We publicly release code, the 🎥 YTD-18M dataset, and 🍾 CHAMPAGNE model checkpoints to facilitate future research on understanding real-world conversations from a visually-grounded perspective.

2. Related Work

Although recent language modeling advancements have enabled machines to engage in conversation and comprehend dialogue similar to humans [65, 47, 9, 3, 10, 16, 20], these efforts have largely been confined to textual contexts. In our work, we consider three real-world conversational frames: social interaction, visually-grounded dialogue, and

[1] **C**onvers**A**tional **M**ultimodal **P**rompted **G**e**NE**rator.

open-domain text conversation. Social interaction involves conversations that incorporate visual cues and are observed from an external viewpoint (*i.e.*, third-person perspective). Examples of these conversations include speaker sentiment analysis in videos as proposed by CMU-MOSEI [62], and human-focused commonsense-based captioning as introduced by Visual Comet [38]. Visually-grounded dialogue, on the other hand, refers to conversations that involve visual contexts from the perspective of the agent, *i.e.*, from the first-person point of view. Visual Dialog [12], which involves answering questions about an image in a conversational setting, and Image Chat [50], a chit-chat grounded to the image, are examples of visually-grounded dialogue. Finally, open-domain text conversation refers to conversations that rely solely on textual contexts, and there have been multiple endeavors to teach machines to engage in natural communication on a wide range of topics [53, 13].

Text-only dialogue models have access to abundant training resources [25, 55, 16, 52, 5, 15], allowing them to generate natural responses based on conversational context. However, despite their striking conversational abilities they lack the capability to comprehend visual contexts. To address this shortcoming, several studies have proposed multimodal dialogue models [51, 54]. Few large-scale datasets such as MMDialog [17] are available for conversations that are grounded on visual contexts, yet the scale of the dataset is quite smaller than those of text-only dialogue datasets, making it difficult to train such models at scale.

Large Vision-Language Models. Recent studies in vision-language research have shown that scaling the size of the model in conjunction with the dataset size can significantly enhance model performance across a range of tasks [40, 43]. Several frameworks have been introduced, including Unified-IO [33] and OFA [58], which propose a unified sequence-to-sequence model for modeling vision-language tasks. These frameworks demonstrate that a single model trained on a broad range of tasks can perform well on various tasks. Additionally, BLIP [30] presents a technique for pretraining models on noisy datasets for both vision-language understanding and generation tasks. The Flamingo model [4], trained on a large amount of noisy web data, demonstrates that it can adapt to tasks with few examples. While these models provide a general vision-language model, our research aims to enhance conversation capability by focusing on specifically learning about conversations.

3. Approach

3.1. Dataset Collection: YTD-18M

In this section, we describe our pipeline to collect YTD-18M, which is outlined in Figure 1.

Extracting and Filtering Videos. The data collection process starts with downloading public YouTube videos. Each video is associated with a user-provided title, metadata, and a transcript generated by the YouTube's internal Automatic Speech Recognition system. We then filter videos by applying several steps using the associated metadata, following the strategy from [64, 63]. In particular, we use Python *cld3* library, which uses a neural network for language identification, to filter out videos whose transcripts have a probability of being English less than 80%. We also discard videos that do not contain visual variation (*e.g.*, a video of podcast with a static thumbnail) or those lacking objects in the thumbnails according to an image classification model [48]. These steps resulted in the extraction of 20M videos.

We further process these 20M videos to build video segments. First, we build a list of segments by iterating through each video with a 60-second sliding window. Second, we filter out segments which have transcripts containing less than 30 words: short transcripts are unlikely to contain multiple conversation turns. Finally, we filter out transcripts that contain unsafe topics to ensure that the model does not learn harmful language. We utilize Rewire API [1] to detect toxic content in the transcripts. At the end of the filtering stage, we end up with 18M video segments.

Converting Noisy Transcripts into Dialogues. Although each video is equipped with its own transcript, they are usually not suitable to train a video-based dialogue model directly. For example, transcripts do not provide any delimitation between the interlocutors. A naive approach to address this problem is to use speaker diarization systems to determine "who spoke when" [56]. However, these systems suffer from low accuracy [39], thus perform poorly when turning a sequence of words to a well-structured dialogue.

Instead, we train a *converter* model to transform noisy transcripts into structured dialogues. Motivated by the in-context learning capabilities of GPT-3 models [7], we prompt GPT-3 with few-shot examples and ask the model to generate well-formatted dialogues given noisy transcripts. However, using GPT-3 to process millions of samples is computationally expensive. Therefore, we train a smaller model using denoised transcripts sampled from GPT-3. That is, we collect 40K input-output pairs from GPT-3 where the input is a noisy transcript and the output is the converted dialogue. Our *converter* is a Unified-IO [33] Base (241M params) fine-tuned on the generated pairs. We exclude video segments that have more than 150 words as the model cannot handle excessively long inputs. The pairs are divided into train and validation sets at a ratio of 99:1. The *converter* achieves a high accuracy of 90.1% on the validation set when using teacher-forcing to predict tokens, suggesting high-quality dialogue generation given noisy transcripts (*e.g.,* in Figure 1, Step 2).

Converting Videos into Video-based Dialogues. To ensure that each utterance in the obtained dialogue matches

Dataset	YTD-18M	MMDialog
Number of Dialog	**18M**	1M
Is the interlocutors of dialog *visible*?		
Visible	**61.6%***	11.5%
If **NOT** visible, then		
how *related* the dialog is to the image?	**2.589**	2.580
how *grounded* the dialog is to the image?	2.429	**2.463**
How *sensible* is the dialog?	**2.495**	2.479
How *specific* is the dialog?	**2.650***	2.464
Is the data containing *explicit* content?		
Sexually Explicit	**0.5%***	1.6%
Hatespeech	**0.3%***	2.5%
Others	**0.3%***	0.9%

Table 1: Human evaluation results on YTD-18M and MMDialog about the visual contexts and the quality of dialogues. YTD-18M covers more cases of social interactions, and dialogues exhibit better sensibleness and specificity compared to MMDialog. * denotes statistically significance after independent two-sample t-test ($p < 0.5$). Full breakdown of numbers are in Appendix B.

correctly with the corresponding video frames, we employ Dynamic Time Warping [36] to align the dialogues with the original noisy transcripts. After the alignment, we use the timing information in the original noisy transcripts to estimate the start time of each dialogue turn, which in turn helps extract the corresponding video frame and minimize alignment errors caused by the conversion process. Figure 2 shows some examples from YTD-18M. More details about collecting dataset can be found in Appendix A.

3.2. Dataset Analysis

To better understand our dataset, we conduct a human evaluation study comparing YTD-18M with MMDialog [17] — the largest dataset for visually-grounded dialogue. MMDialog is a dataset sourced from people's interaction in social media. We randomly sample 500 examples from each dataset and ask three workers to assess each example for several factors. For more information regarding the human evaluation, please refer to the Appendix B.

Dataset quality. To compare the dialogue quality, we ask workers to assess examples using two criteria: sensibleness and specificity [3]. To be sensible, a dialogue should be reasonable, logical, and not confusing. Specific dialogue is one that is not dull or generic. Each human annotators rate the dialogue in specific aspect on a 3-point Likert scale, *e.g.* for sensibleness "1" being "Not Sensible" and "3" being "Sensible". Evaluation results are shown in Table 1. On average, YTD-18M received higher scores than MMDialog across the axes of sensibleness and specificity. We suspect that MMDialog, which is derived from social media, may lack a natural conversation flow due to the non-consecutive nature of social media interactions.

Social Interaction. Given the prevalence of third-person point of view in web videos, we postulate that a substantial proportion of such videos feature social interactions. To

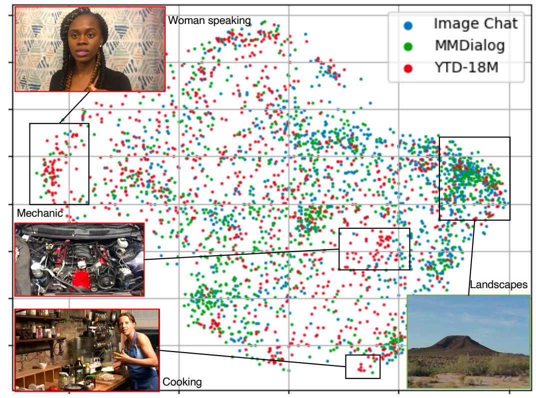

Figure 3: Visual feature distributions of YTD-18M and other visually-grounded dialogue datasets. Our YTD-18M includes a wide range of visual contexts, with a particular emphasis on frames in which a person is speaking, in contrast to the other datasets (shown in the upper left cluster).

investigate this claim, we enlisted the aid of workers to determine whether the conversation's interlocutors were visible in the image frames and, if so, whether their body language was present. Our findings, presented in Table 1, indicate that our video-based dialogue dataset has a considerably higher proportion of visible interlocutors (61.6%) than MMDialog (11.5%). This discrepancy can be explained by the fact that it is uncommon for interlocutors in social media interactions to reveal their identities through images. Moreover, when interlocutors are visible, workers accurately identified facial expressions in 83.6% of cases, and tagged body posture in 64.7% of them. These results suggest that our dataset presents a valuable resource for exploring body language in communication.

Visual Grounding. Real-life conversations do not always have a direct relationship or grounding to images. Therefore, a higher degree of relevance between dialogues and images does not necessarily imply a higher quality dataset, although it offers more visual grounding opportunities for models to learn from. We ask workers to assess the degree of grounding between conversations and images on a 3-point Likert scale when interlocutors were not visible in the images. According to Table 1, YTD-18M exhibits grounding scores that are comparable to those of MMDialog, implying that models can acquire visual grounding knowledge from YTD-18M.

Distribution of Visual Contexts. In order to gain a better understanding of the visual representations encoded in YTD-18M from videos, we examine the distributions of visual features across three distinct datasets: Image Chat, MMDialog, and YTD-18M, that are grounded in dialogue. Similar to [32], we use CLIP ViT-L14 [40] trained on LAION-2B [49].[2] We utilize this model to extract embeddings, and the embeddings are then projected into a 2D fea-

[2] [32] conducted a similar study, but using ResNet50 ImageNet features.

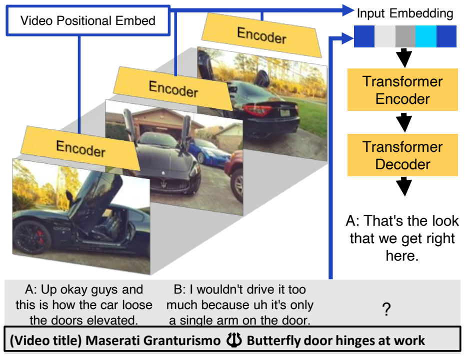

Figure 4: Training 🍾CHAMPAGNE on 📽YTD-18M. The model takes a video title, a dialogue context, and image frames as an input and learns to predict the response.

ture space by way of UMAP [35]. In order to conduct our analysis, we randomly sample 1K images from each of the aforementioned datasets.

Figure 3 illustrates the results, indicating that the datasets share some similarities but differ in certain aspects. Notably, YTD-18M has a distinct distribution pattern from Image Chat and MMDialog, with a higher proportion of images featuring a person speaking, which is consistent with the previous discovery that YTD-18M emphasizes social interactions. Additionally, YTD-18M encompasses a broad range of diverse and specific topics, such as cooking or mechanics. Further information regarding the analysis of visual contexts can be found in the Appendix C.

Content Safety. We ask workers to identify whether the dialogues or images contain any potentially unsafe content, like sexually explicit material or hatespeech. As indicated in Table 1, our YTD-18M has fewer occurrences of sexually explicit content and hate speech when compared to MMDialog. We suspect that this discrepancy is due to the inclusion of safety filtering step in the data collection process for YTD-18M.

Video Title as an Additional Feature. We further ask the workers to rate the relevance of the video title to the dialogue on a 3-point Likert scale, ranging from 1 (Not Related) to 3 (Related). The results show that 65% of the videos have a title that is relevant to the dialogues, and 21% have titles that were somewhat related. This led us to utilize the video titles as prompts when training our model. Our qualitative findings (§4.1) suggest that CHAMPAGNE can be conditioned effectively by such prompts.

3.3. Model: 🍾CHAMPAGNE

CHAMPAGNE is a multimodal conversational agent that takes image frames, a prompt, and a dialogue context as an input and generates a response. The overview of training process is described in Figure 4. The collected YouTube titles for each video serve as prompts and are denoted as P. Finally, the vision-based dialogue is denoted as $D = (P, I_1, T_1, ..., I_n, T_n)$, where I_i and T_i denote an image frame and the dialogue turn, respectively.

Figure 5: Validation perplexities during the training process of CHAMPAGNE models on YTD-18M.

Architecture. The architecture used in CHAMPAGNE is based on the Unified-IO model proposed by [33]. This model is designed for vision-and-language tasks and operates as a sequence-to-sequence model. Although it can handle image and text inputs together, it is unable to handle multiple images. To address this limitation, we introduce *video position embeddings* that can be learned and incorporated into CHAMPAGNE. Specifically, CHAMPAGNE converts each frame of the video into a sequence of patch encodings using a visual encoder. It then adds video position embeddings to the patch encodings to capture temporal information of the video frames. The patch encodings from multiple image frames are averaged through mean pooling and fed into a Transformer encoder. Our experiments use three image frames per dialogue to train CHAMPAGNE.

Training. We initialize CHAMPAGNE training with the pretrained weight from Unified-IO model that is pretrained on collection of C4 [41], Wikipedia, ImageNet21K [45], and YFCC15M [40] with a denoising objective. Unified-IO model is trained in two stages, pre-training and the multitask stage, and the weights from the pre-training stage, referred to as Unified-IO$_{PT}$[3], are used to initialize the CHAMPAGNE model. We train the model using a next token prediction objective, which aims to maximize the likelihood of the target response T_k when taking multiple images $I_{i\leq k}$, dialogue context $T_{i<k}$, and the video title P as an input, *i.e.*, $p_\theta(T_k|I_1, T_1, ..., I_{k-1}, T_{k-1}, T_k, P)$.

We present three versions of the model, BASE (241M), LARGE (776M), and XL (2.9B). We train models for 3 epochs on YTD-18M, and training CHAMPAGNE-XL takes approximately 3 days on TPU v3-256 on Google Cloud Virtual Machines with T5X framework [46]. More details about hyperparameters are in Appendix D.

4. Experiments

Figure 5 shows the training curves for CHAMPAGNE models that were trained on YTD-18M. We evaluate the

[3]We run a pilot study and find that the model initialized from Unified-IO$_{PT}$ weights perform better on downstream tasks compared to the model initialized from Unified-IO.

Figure 6: CHAMPAGNE-XL predicting next utterance when the visual and dialogue contexts are given from the unseen video.

Dataset	BST		ConvAI2		ED		WOW		WOI		IC		IC First Turn	
Metric (on Test Set)	PPL	Dist-3	PPL	Dist-3	PPL	Dist-3	PPL	Dist-3	PPL	Dist-3	PPL	Dist-3	PPL	Dist-3
Without Fine-tuning														
CHAMPAGNE-XL	34.1	0.417	34.2	0.371	32.8	0.544	31.2	0.705	34.8	0.641	50.5	0.213	69.9	0.135
Reddit 2.7B [47]	20.0	0.390	24.2	0.286	14.0	0.418	24.3	0.458	23.5	0.473	40.9	0.152	66.3	0.011
Fine-tuned														
Multimodal Blender 2.7B [51]	**14.6**	0.418	13.8	0.300	**11.4**	0.306	**15.6**	0.527	19.8	0.522	24.0	0.200	25.9	0.193
Unified-IO$_{PT}$ BASE	20.4	0.280	16.5	0.199	19.8	0.352	22.6	0.480	26.0	0.449	29.9	0.106	29.6	0.113
Unified-IO$_{PT}$ LARGE	17.7	0.284	15.0	0.176	15.8	0.304	18.6	0.484	21.3	0.436	26.8	0.108	26.7	0.096
Unified-IO$_{PT}$ XL	15.7	0.346	12.9	0.250	14.9	0.393	16.5	0.580	18.9	0.515	20.5	0.182	22.2	0.163
CHAMPAGNE-BASE	19.1	0.395	15.2	0.361	19.0	0.421	21.3	0.651	24.1	0.584	27.2	0.174	27.4	0.220
CHAMPAGNE-LARGE	16.9	0.390	13.6	0.353	18.6	0.454	18.5	0.665	21.7	0.590	22.9	0.205	23.6	0.256
CHAMPAGNE-XL	15.1	**0.434**	**12.3**	**0.390**	16.2	**0.499**	16.1	**0.690**	**18.5**	**0.624**	**19.3**	**0.247**	**21.2**	**0.278**

Table 2: Automatic evaluation results on open-domain text conversation benchmarks and Image Chat (IC). For fair comparison, all perplexities (PPL, ↓) are normalized to be in the space of GPT-2 tokenizer [60]. Dist-3 (↑) is calculated over corpus-level (Inter Dist-3). *IC First Turn* indicates the situation in which the model generates the response given image but no textual context from IC, which is to highlight model's ability to ground on visual contexts. Note that Multimodal Blender 2.7B is a fine-tuned version of Reddit 2.7B.

sensibleness and specificity of the response generated by CHAMPAGNE models by computing perplexities [24] on the validation set of YTD-18M. Perplexity (PPL) has been shown to be a reliable indicator of these qualities [3]. The training curves in Figure 5 demonstrate that the models are capable of learning to generate appropriate next-turn utterances given visual and dialogue contexts, and that they generalize well to the validation set when trained on YTD-18M. A qualitative example of this successful generalization can be seen in Figure 6.

We next conduct experiments on a range of benchmarks that assess the model's performance on real-world conversation tasks. Additional information about the benchmarks and the evaluation metrics used can be found in Appendix E.

4.1. Open-domain Text Conversation

To investigate whether CHAMPAGNE models are capable of learning chit-chat skills from YTD-18M, we conduct automatic evaluations on open-domain text conversation benchmarks. These benchmarks assess the model's ability to produce appropriate responses given a textual dialogue context in a conversational scenario. Specifically, we evaluate the model's performance on five text-only benchmarks: BST [53], ConvAI2 [13], ED [44], WOW [14], and WOI [27]. We employ two metrics to evaluate the models: PPL and Dist-3 [29]. Dist-3 is a metric that has been shown to reflect the diversity and interestingness of the response [19]. Results are shown in Table 2 (left).

Prior to fine-tuning, both Unified-IO$_{PT}$ XL and Unified-IO XL exhibited perplexities over 100, which suggests that these models generate unintelligible sentences. In contrast, CHAMPAGNE-XL achieved significantly lower perplexities on all benchmarks, indicating that the model has effectively learned chit-chatting capabilities from YTD-18M. Next, we fine-tune both Unified-IO$_{PT}$ and CHAMPAGNE on a mixture of dialogue benchmarks, with the multi-task training weight set to BST: ConvAI2: ED: WOW: WOI: IC $= 1:3:3:3:3:3$, following the approach described in [47]. After fine-tuning, CHAMPAGNE achieves significantly higher Dist-3 scores and lower PPL values than the same-sized Unified-IO$_{PT}$, indicating that CHAMPAGNE has successfully learned about textual conversations from YTD-18M. Our experimental results further suggest that larger models tend to perform better overall, with CHAMPAGNE-XL reporting Dist-3 scores that surpassed those of other baselines on all benchmarks, demonstrating its ability to generate diverse and interesting responses. However, in terms of PPL, CHAMPAGNE-XL shows weaker performance on BST, ED, and WOW compared to the Multimodal Blender 2.7B. We conjecture that this may be due to the influence of the large text-based dialogue dataset (the Pushshift dataset [5]), which contains 3B Reddit comments. Indeed, the Reddit 2.7B model, the text-only model which is trained on the Pushshift dataset, performs better

Metric	Acc. (%) (↑)	F1 (↑)
UniMSE [22]	85.8	0.858
MAG-BERT [42]	84.7	0.845
Unified-IO$_{PT}$ BASE	70.1	0.780
Unified-IO$_{PT}$ LARGE	72.9	0.812
Unified-IO$_{PT}$ XL	81.5	0.869
CHAMPAGNE-BASE	82.3	0.879
CHAMPAGNE-LARGE	83.9	0.892
CHAMPAGNE-XL	**86.1**	**0.899**

Table 3: Evaluation results on test split of CMU-MOSEI sentiment analysis task.

Metric	CIDEr-D (↑)		BLEU-4 (↑)	
Split	Val	Test	Val	Test
Unified-IO XL	0.212	-	0.073	-
SOTA on Public Leaderboard [2]	-	0.184	-	0.041
Unified-IO$_{PT}$ BASE	0.344	-	0.095	-
Unified-IO$_{PT}$ LARGE	0.356	-	0.098	-
Unified-IO$_{PT}$ XL	0.384	-	0.105	-
CHAMPAGNE-BASE	0.351	-	0.097	-
CHAMPAGNE-LARGE	0.369	0.322	0.102	0.080
CHAMPAGNE-XL	**0.399**	**0.354**	**0.109**	**0.092**

Table 4: Evaluation results on Visual Comet.

than CHAMPAGNE-XL without fine-tuning, suggesting that the Pushshift dataset has a distribution that is more similar to those of these benchmarks than YTD-18M. We expect that further improving the model's performance on these benchmarks may require training CHAMPAGNE-XL on the Pushshift dataset in addition to fine-tuning it on these benchmarks.

4.2. Understanding Social Interactions

We evaluate the performance of three variants of CHAMPAGNE and Unified-IO$_{PT}$ on two benchmarks that measure social interactions: the sentiment analysis task in CMU-MOSEI [62] and Visual Comet [38]. Table 3 and Table 4 show the results on CMU-MOSEI and Visual Comet, respectively. Results show that fine-tuning CHAMPAGNE-XL on these benchmarks leads to the highest accuracy and F1 score among the baselines, including the state-of-the-art model (UniMSE) on CMU-MOSEI. Additionally, our model outperforms the state-of-the-art result on Visual Comet, both on the validation and test sets. Furthermore, the models fine-tuned from CHAMPAGNE consistently demonstrate significant improvements compared to the models fine-tuned from same-sized Unified-IO$_{PT}$ on both CMU-MOSEI and Visual Comet. This indicates that our model effectively learns meaningful representations about social interactions from YTD-18M.

4.3. Visually-grounded Dialogues

We run experiments on two visually-grounded dialogue benchmarks: Visual Dialog [12] and Image Chat [50]. Table 5 presents the results of models on Visual Dialog in both zero-shot and fine-tuned settings. Similarly, Table 2 (right) shows the results of models fine-tuned on combination of dialogue datasets on Image Chat and Image Chat First Turn. Fine-tuning CHAMPAGNE on those benchmarks outperforms fine-tuning the same sized Unified-IO$_{PT}$ model on both tasks. This suggests that the model effectively learns from YTD-18M, and larger models tend to perform better. For Visual Dialog zero-shot setting, CHAMPAGNE-XL outperforms ESPER and FROMAGe in terms of NDCG [23]. In the fine-tuned setting, CHAMPAGNE-XL achieves the highest NDCG compared to Flamingo-80B and the state-of-the-art model (AlignVD). However, in the zero-shot setting, Flamingo-80B performs significantly better than CHAMPAGNE-XL, which might be due to Flamingo's larger parameters and training data. Based on our scaling observations, we suspect that training Flamingo-80B or a similar sized model on YTD-18M could lead to further performance improvements. For Image Chat and Image Chat First Turn, CHAMPAGNE-XL shows the lowest PPL and highest Dist-3 scores among the baselines.

Metric	NDCG (×100) (↑)
Zero-shot	
Flamingo-80B	52.0
ESPER [61]	22.3
FROMAGe [26]	16.5
CHAMPAGNE-XL	25.5
Fine-tuned	
Flamingo-80B	61.8
AlignVD [8]	67.2
Unified-IO$_{PT}$ BASE	58.9
Unified-IO$_{PT}$ LARGE	60.3
Unified-IO$_{PT}$ XL	65.6
CHAMPAGNE-BASE	60.0
CHAMPAGNE-LARGE	62.5
CHAMPAGNE-XL	**68.2**

Table 5: Evaluation results on Visual Dialog valid set in finetuned and zero-shot settings. CHAMPAGNE-XL shows the best zero-shot performance except for Flamingo-80B, which has 30× more parameters, and CHAMPAGNE-XL achieves the state-of-the-art result on finetuned setting. For fair comparison, we report baselines that do not use additional dense annotations to finetune the model.

We additionally carry out a human evaluation to compare our best model with the baselines trained on Image Chat. We randomly choose 100 examples from both Image Chat and Image Chat First Turn. For each example, we asked three workers to rate the dialogue responses generated by the model on four aspects: (1) sensibleness, (2) specificity, (3) grounding to the image, and (4) relevance to the given style. The workers were presented with positive/negative options for each aspect. For instance, for the grounding aspect, they were asked to choose between "Yes, the response is grounded to the image" or "No, the response is not grounded to the image." Our evaluation results, presented

(a) A conversation with the person
User: Hi, what are you doing now?
Bot: I'm working on my computer. I'm a programmer.
User: Where are you working now?
Bot: I'm working at a restaurant. I'm in the kitchen.
User: Are you alone?
Bot: No, I'm with my girlfriend. We've been together for a while.

(b) A conversation between two people about the picture
User: Hi, what is the man doing in the picture?
Bot: He's typing. He's typing on his laptop.
User: Where is he now?
Bot: He's in his office. He's working on a project.
User: Is he alone?
Bot: No, he's with his wife. They have three kids.

Figure 7: Examples of conversation based on unseen images from COCO [31] and prompts (colored in blue) between person and CHAMPAGNE-XL finetuned on mixture of dialogue benchmarks. The utterances with the prefix *Bot:* are generated by the model. (a) represents a conversational frame of social interaction (bot pretends to be the person in the picture), while (b) represents a conversational frame of visually-grounded dialogue (bot talks about the person in the picture).

Dataset	Image Chat					Image Chat First Turn				
Metric	Sensible	Specific	Grounding	Style	Avg.	Sensible	Specific	Grounding	Style	Avg.
Multimodal Blender 2.7B	81.3%	66.7%	**98.7%**	69.0%	78.9%	**93.3%**	70.0%	85.0%	82.3%	82.6%
Unified-IO$_{PT}$ XL	**84.7%**	88.3%	98.0%	84.3%	88.8%	92.0%	83.0%	87.7%	82.0%	86.3%
🍾 CHAMPAGNE-XL	82.7%	**93.7%**^{UM}	**98.7%**	**90.0%**^{UM}	**91.3%**^{UM}	92.0%	**93.0%**^{UM}	**90.1%**^{M}	**88.3%**^{UM}	**90.8%**^{UM}

Table 6: Human evaluation results on Image Chat and Image Chat First Turn. *Grounding* denotes how well the generated response is grounded to the image, *Style* denotes how well the generated response reflects the given style, and *Avg.* denotes the average percent of results. We run independent two-sample t-tests, and we denote as ^{M} and ^{U} when CHAMPAGNE-XL is statistically significant ($p < 0.05$) over Multimodal Blender 2.7B and Unified-IO$_{PT}$ XL, respectively.

Dataset	CMU-MOSEI	IC First Turn	
Metric	Acc. (%)	PPL (↓)	Dist-3 (↑)
🍾 CHAMPAGNE-LARGE trained on:			
YTD-18M	**83.9**	**21.6**	**0.260**
YTD-2M	83.2	28.3	0.190
YTD-2M w/ single video frame	78.9	29.3	0.066
YTD-2M w/o video frames	75.6	29.7	0.073
YTD-2M w/o transcript to dialog	74.9	29.0	0.143

Table 7: Ablation study on the test set of CMU-MOSEI and Image Chat First Turn to validate the effect of YTD-18M.

in Table 6, indicate that our model, CHAMPAGNE-XL, outperforms the baselines in terms of specificity, grounding, and style. As for sensibleness, our model shows comparable scores to the best results. More details on the human evaluation can be found in Appendix B. Lastly, Figure 7 presents some sample conversations between a human and our model. The figure demonstrates that the model can engage in conversations based on images and can be prompted to respond accordingly.

4.4. Ablations

To investigate the impact of components in YTD-18M, we conduct ablation studies on CMU-MOSEI and Image Chat First Turn. Our evaluation of models on Image Chat First Turn only involves fine-tuning on Image Chat and excludes any text-only dialogue datasets.

Number of Examples. The performance drop in both tasks is evident when reducing the number of examples in YTD-18M from 18M to 2M, as seen in Table 7. These findings are consistent with previous works [21, 7] and highlight the benefits of learning from a larger dataset. Notably, the significance of the performance drop is greater in the visually-grounded dialogue task of Image Chat, underscoring the crucial role of the number of examples in such tasks.

Video Frames. To validate whether the models learn from visual contexts, we conduct two experiments. The first experiment involves training the model without video frames, while the second experiment utilizes only a single video frame for training. Table 7 presents the results of these experiments, which show a decrease in performance for both tasks. These findings support the idea that YTD-18M provides an opportunity for the model to learn visual grounding. Notably, the accuracy drop is more significant for CMU-MOSEI, where training with a single frame results in an accuracy drop from 83.2% to 78.9%, and training without video frames results in an even greater accuracy drop to 75.6%. This suggests that relying on visual cues is essential for understanding sentiment in social interaction. Furthermore, in Image Chat, there is a substantial drop in Dist-3, indicating that models trained without visual contexts may generate generic responses that are irrelevant to the images.

Dialogue Format. By comparing the model trained on the noisy transcript to the model trained on dialogue format, we aim to demonstrate how the latter can improve the machine's ability to learn representations of conversation. In Table 7, we observe a decrease in performance for both tasks. Notably, on the CMU-MOSEI dataset, the accuracy drops from 83.2% to 74.9% without the dialogue conversion process. We speculate that the conversion process helps the dataset recover crucial speaker information that is missing from the noisy transcript. This information could prove beneficial for learning representations of social interaction.

5. Conclusion

We introduce 🎥 YTD-18M, a large-scale video-based dialogue dataset and 🍾 CHAMPAGNE, a model that learns about the real-world conversation from YTD-18M. Our experiments show that CHAMPAGNE exhibits strong performance in various real-world conversation tasks, indicating that video-based dialogues can help models to learn about real-world conversations. We will release our data and model checkpoints for research purposes. In future work, we plan to explore the potential of utilizing auditory signals from videos to further enhance our understanding of real-world conversations.

6. Ethical Considerations

In this paper, we introduce a large-scale dataset derived from publicly available YouTube videos. With an emphasis on teaching machines about real-world conversation, our dataset includes frames that present the interlocutors. It might capture their facial expressions and body cues, however, may give rise to user privacy concerns. To mitigate this issue, we release only the video IDs instead of raw videos, following the prior works [64, 63]. It allows users to delete their video from YouTube, thereby excluding them from the dataset. To strengthen user privacy even further, future directions may include works such as anonymizing faces and speech from the videos, and deidentifying personal information from the transcripts.

References

[1] Rewire api. `https://rewire.online/`.

[2] Visual comet public leaderboard. `https://leaderboard.allenai.org/visualcomet/submissions/public`.

[3] Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. Towards a human-like open-domain chatbot. *ArXiv*, abs/2001.09977, 2020.

[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198, 2022.

[5] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839, 2020.

[6] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online, Nov. 2020. Association for Computational Linguistics.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[8] Feilong Chen, Duzhen Zhang, Xiuyi Chen, Jing Shi, Shuang Xu, and Bo Xu. Unsupervised and pseudo-supervised vision-language alignment in visual dialog. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4142–4153, 2022.

[9] Jiaao Chen and Diyi Yang. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118, 2020.

[10] Jiaao Chen and Diyi Yang. Simple conversational data augmentation for semi-supervised abstractive dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6605–6616, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.

[11] Yuchen Cui, Qiping Zhang, Brad Knox, Alessandro Allievi, Peter Stone, and Scott Niekum. The empathic framework for task learning from implicit human feedback. In *Conference on Robot Learning*, pages 604–626. PMLR, 2021.

[12] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017.

[13] Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander H. Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur D. Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander I. Rudnicky, Jason Williams, Joelle Pineau, Mikhail S. Burtsev, and Jason Weston. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition*, pages 187–208. Springer, 2020.

[14] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*, 2018.

[15] Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar R Zaiane. Augmenting neural response generation with context-aware topical attention. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 18–31, 2019.

[16] Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M Ponti, and Siva Reddy. FaithDial: A Faithful Benchmark for Information-Seeking Dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490, 12 2022.

[17] Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. Mmdialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. *arXiv preprint arXiv:2211.05719*, 2022.

[18] Jessica Ficler and Yoav Goldberg. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104, 2017.

[19] Seungju Han, Beomsu Kim, and Buru Chang. Measuring and improving semantic diversity of dialogue generation. *ArXiv*, abs/2210.05725, 2022.

[20] Seungju Han, Beomsu Kim, Seokjun Seo, Enkhbayar Erdenee, and Buru Chang. Understanding and improving the exemplar-based generation for open-domain conversation. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 218–230, 2022.

[21] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and L. Sifre. Training compute-optimal large language models. *ArXiv*, abs/2203.15556, 2022.

[22] Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. *arXiv preprint arXiv:2211.11256*, 2022.

[23] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[24] Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63, 1977.

[25] Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. Soda: Million-scale dialogue distillation with social commonsense contextualization. *ArXiv*, abs/2212.10465, 2022.

[26] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal generation. *arXiv preprint arXiv:2301.13823*, 2023.

[27] Mojtaba Komeili, Kurt Shuster, and Jason Weston. Internet-augmented dialogue generation. *arXiv preprint arXiv:2107.07566*, 2021.

[28] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. 2022.

[29] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, 2016.

[30] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[32] Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, 2021.

[33] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022.

[34] Leland McInnes and John Healy. Accelerated hierarchical density based clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 33–42. IEEE, 2017.

[35] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[36] Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

[37] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[38] Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visualcomet: Reasoning about the dynamic context of a still image. In *European Conference on Computer Vision*, pages 508–524. Springer, 2020.

[39] Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J Han, Shinji Watanabe, and Shrikanth Narayanan. A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, 72:101317, 2022.

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[41] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), jan 2020.

[42] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369, Online, July 2020. Association for Computational Linguistics.

[43] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray,

Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[44] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*, 2018.

[45] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.

[46] Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, Curtis Hawthorne, Aitor Lewkowycz, Alexandru Salcianu, Marc van Zee, Jacob Austin, Sebastian Goodman, Livio Baldini Soares, Haitang Hu, Sasha Tsvyashchenko, Aakanksha Chowdhery, Jasmijn Bastings, Jannis Bulian, Xavier García, Jianmo Ni, Andrew Chen, Kathleen Kenealy, J. Clark, Stephan Lee, Daniel H Garrette, James Lee-Thorp, Colin Raffel, Noam M. Shazeer, Marvin Ritter, Maarten Bosma, Alexandre Passos, Jeremy B. Maitin-Shepard, Noah Fiedel, Mark Omernick, Brennan Saeta, Ryan Sepassi, Alexander Spiridonov, Joshua Newlan, and Andrea Gesmundo. Scaling up models and data with t5x and seqio. *ArXiv*, abs/2203.17189, 2022.

[47] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online, Apr. 2021. Association for Computational Linguistics.

[48] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[49] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[50] Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. Image chat: Engaging grounded conversations. *arXiv preprint arXiv:1811.00945*, 2018.

[51] Kurt Shuster, Eric Michael Smith, Da Ju, and Jason Weston. Multi-modal open-domain dialogue. *arXiv preprint arXiv:2010.01082*, 2020.

[52] Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, W.K.F. Ngan, Spencer Poff, Naman Goyal, Arthur D. Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *ArXiv*, abs/2208.03188, 2022.

[53] Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. Can you put it all together: Evaluating conversational agents' ability to blend skills. *arXiv preprint arXiv:2004.08449*, 2020.

[54] Qingfeng Sun, Yujing Wang, Can Xu, Kai Zheng, Yaming Yang, Huang Hu, Fei Xu, Jessica Zhang, Xiubo Geng, and Daxin Jiang. Multimodal dialogue response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2854–2866, 2022.

[55] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam M. Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, Yaguang Li, Hongrae Lee, Huaixiu Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, I. A. Krivokon, Willard James Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Hartz Søraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Díaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravindran Rajakumar, Alena Butryna, Matthew Lamm, V. O. Kuzmina, Joseph Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Huai hsin Chi, and Quoc Le. Lamda: Language models for dialog applications. *ArXiv*, abs/2201.08239, 2022.

[56] SE Tranter, K Yu, DA Reynolds, G Evermann, DY Kim, and PC Woodland. An investigation into the interactions between speaker diarisation systems and automatic speech transcription. Technical report, Tech. Rep. CUED/F-INFENG/TR-464, Cambridge University Engineering Department, 2003.

[57] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[58] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022.

[59] Shuhe Wang, Yuxian Meng, Xiaoya Li, Xiaofei Sun, Rongbin Ouyang, and Jiwei Li. Openvidial 2.0: A larger-scale, open-domain dialogue generation dataset with visual contexts. *arXiv preprint arXiv:2109.12761*, 2021.

[60] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics.

[61] Youngjae Yu, Jiwan Chung, Heeseung Yun, Jack Hessel,

Jae Sung Park, Ximing Lu, Prithviraj Ammanabrolu, Rowan Zellers, Ronan Le Bras, Gunhee Kim, and Yejin Choi. Multimodal knowledge alignment with reinforcement learning. *ArXiv*, abs/2205.12630, 2022.

[62] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018.

[63] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387, 2022.

[64] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, 2021.

[65] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*, 2019.