TIGER: A Unified Generative Model Framework for Multimodal Dialogue Response Generation

Fanheng Kong, Peidong Wang, Shi Feng†, Daling Wang, Yifei Zhang
Northeastern University, China
{kongfanheng, pdongwang}@stumail.neu.edu.cn
{fengshi, wangdaling, zhangyifei}@cse.neu.edu.cn

Abstract

Responding with multimodal content has been recognized as one of the essential functionalities of intelligent conversational agents. However, existing research on multimodal dialogues primarily focuses on two topics: (1) textual response generation that ground the conversation on a given image; and (2) visual response selection based on the dialogue context. In light of the aforementioned gap, we propose mul**TI**modal **GE**nerator for dialogue **R**esponse (**TIGER**), a unified generative model framework for multimodal dialogue response generation. Through extensive experiments, TIGER has demonstrated new state-of-the-art results, providing users with an enhanced conversational experience. A multimodal dialogue system based on TIGER is available at https://github.com/friedrichor/TIGER. A video demonstrating the system is available at https://www.youtube.com/watch?v=Kd0CMwDs8Rk.

Keywords: Multimodal, Dialogue Generation, Low-resource, Demonstration

1. Introduction

Open-domain conversational agents (Zhang et al., 2020; Adiwardana et al., 2020; Roller et al., 2021) have demonstrated outstanding performance on text-only dialogue generation, offering interesting and engaging experiences for users. However, relying solely on a single text modality falls short of fully simulating the rich visual perception of the real physical world (Liang et al., 2021; Lin et al., 2023).

Multimodal dialogue (Shuster et al., 2021; Sun et al., 2022), which refers to the conversation with various modal contents (e.g., text, image, audio), is an emerging research direction in the field of natural language processing. Multimodalities enhance the interactivity and expressiveness of the conversation (Kusal et al., 2022), making dialogue systems highly applicable in a wide range of scenarios.

Communication based on images is more engaging than text-only conversations (Hu et al., 2014). Although OFA (Wang et al., 2022), BLIP (Li et al., 2022, 2023) have demonstrated the ability to receive multimodal information and generate appropriate text responses, limited research has been conducted on generating multimodal responses in conversational scenarios. Sun et al. (2022) formulate a new problem: Multimodal Dialogue Response Generation (MDRG), where models should have the ability to generate responses in multiple modalities, and present Divter, which consists of two Transformer-based (Vaswani et al., 2017) components: a textual dialogue response generator, and a text-to-image translator. We found that: (1) the prediction of response modal has not received adequate attention, despite its significant impact on the overall conversational experience; and (2) an inherent limitation of Divter is its inability to guarantee consistently high-quality and contextually appropriate images generated from the dialogue context, thereby presenting an ongoing challenge.

In this paper, we delve into open-domain multimodal dialogue response generation in a low-resource setting. We incorporate text-to-image translation into text-only dialogue, which allows multimodal dialog response generation to be achieved with small-scale multimodal dialogue data.

Our work makes contributions as follows:

- We propose mul**TI**modal **GE**nerator for dialogue **R**esponse (**TIGER**), a unified generative model framework designed for multimodal dialogue response generation. Notably, this framework is capable of handling conversations involving any combination of modalities.
- We implement a system for multimodal dialogue response generation, incorporating both text and images, based on TIGER.
- Extensive experiments show that TIGER achieves new state-of-the-art results on both automatic and human evaluations, which validate the effectiveness of our system in providing a superior multimodal conversational experience.

2. Related Work

2.1. Open-domain Dialogue

Text-only Dialogue Open-domain dialogue generation is a popular research topic in artificial in-

†Corresponding author.

LREC-COLING 2024, pages 16135–16141

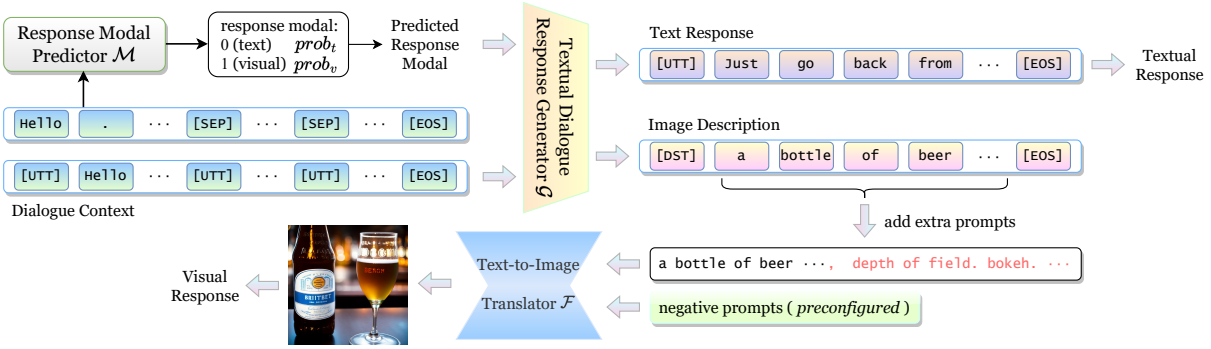

Figure 1: The overview of TIGER framework. Given the dialogue context, response modal predictor $\mathcal{M}$ determines the timing to respond with images. If the predicted response modal is text, textual dialogue response generator $\mathcal{G}$ generates the text response. Conversely, $\mathcal{G}$ produces an image description, and text-to-image translator $\mathcal{F}$ leverages this description to generate an image as the visual response.

telligence, due to its successful applications in social chatbots (e.g., Microsoft XiaoIce (Shum et al., 2018)) and virtual assistants (e.g., Amazon Alexa (Ram et al., 2018)). Recently, pre-trained open-domain dialogue generation models, such as DialoGPT (Zhang et al., 2020), Menna (Adiwardana et al., 2020), and BlenderBot (Roller et al., 2021; Komeili et al., 2022; Shuster et al., 2022), have shown excellent conversation performance.

Multimodal Dialogue Existing works (Gao et al., 2020; Zang et al., 2021; Liang et al., 2021; Lu et al., 2023) have studied dialogue systems with the ability to generate multimodal responses, but they focus on retrieval-based methods, which are constrained by the richness of the dataset, making it challenging for them to perform effectively in new and unfamiliar scenarios. Sun et al. (2022) present Divter, a conversational agent powered by large-scale visual world experiences, with pending improvements in response modal prediction and image generation. Large multimodal models, such as GILL (Koh et al., 2024), SEED (Ge et al., 2023), and Emu (Sun et al., 2023b,a), can perceive and generate images, but the explorations on dialogue are scarce. MiniGPT-5 first explores the performance of large multimodal models on dialogue. Our research centers on multimodal conversations in a low-resource setting and emphasizes multimodal response generation and its implementation through generative methods.

2.2. Text-to-Image Generation

Recently, there has been a surge of research in text-to-image generation, resulting in numerous works that have demonstrated impressive performance. OpenAI's DALL-E (Ramesh et al., 2021) takes the text-to-image field to a completely new level. For the first time, it shows a great zero-shot generalization of text-to-image models. Nichol et al. (2022) combine Diffusion Models (Sohl-Dickstein et al., 2015; Ho et al., 2020) with text-based guidance which allows GLIDE to achieve better diversity and fidelity. DALL-E 2 (unCLIP) (Ramesh et al., 2022), Imagen (Saharia et al., 2022), Latent Diffusion Models (LDMs) (Rombach et al., 2022) are several of the best performing text-to-image models so far. Stable Diffusion (Rombach et al., 2022) is implemented based on LDMs and has become the most popular text-to-image model. Beyond the basic text-to-image generation, our work explores this in multimodal dialogue scenarios. Ensuring the contextual relevance of generated images in conversational scenarios is important and complex.

3. Approach

In this section, we introduce our research strategy and the architecture of TIGER.

3.1. Low-resource Setting

End-to-end models often require a high number of training instances. Due to the limitation of less available multimodal dialogue instances, we conduct the study in a low-resource setting. Figure 2 shows our research strategy. Given a dialogue context U, the target is to generate a textual response r^t or a visual response r^v. The process from U to r^t is text dialogue generation, while it is extremely challenging and less effective to generate a visual response r^v from the dialogue context U. Fortunately, there exists large-scale text dialogue data $\mathcal{D}_C$ (e.g., Reddit comments) and image-text pair data $\mathcal{D}_P$ (e.g., LAION-5B (Schuhmann et al., 2022)). A feasible approach is to introduce text-to-image into text-only dialogue. In this way, our target is to learn a generative multimodal dialogue model $P(R \mid U; \theta)$ with $\mathcal{D} = \{\mathcal{D}_C, \mathcal{D}_P, \widetilde{\mathcal{D}}_S\}$, where $\widetilde{\mathcal{D}}_S$ is a small-scale multimodal dialogue dataset, R is multimodal response, and θ is the parameters of the model.

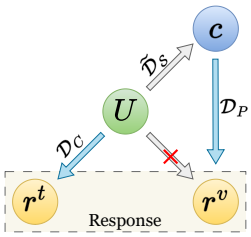

Figure 2: Our research strategy, inspired by Sun et al. (2022). Blue arrows indicate that large-scale datasets exist for pre-training, and gray arrows indicate that only very few training instances are available. "×" indicates bad performance.

3.2. TIGER

Figure 1 presents the overview of the framework. TIGER consists of three components: (1) response modal predictor $\mathcal{M}$; (2) textual dialogue response generator $\mathcal{G}$; (3) text-to-image translator $\mathcal{F}$.

3.2.1. Response Modal Predictor

The timing to respond with images is particularly important for the experience of the conversation. However, existing work (Sun et al., 2022) lacks attention toward the response modal prediction. We formulate the problem as a binary classification task. Given a dialogue context U, the target is to predict the modality $m \in \{t, v\}$ of the next response r_j, where t is textual modality and v is visual modality. Formulaically, response modal prediction is defined as a binary classification task:

$$\forall j \in [1, h], \mathcal{M}(U, R_{<j}) \in \{0, 1\}, \quad (1)$$

where $\mathcal{M}(\cdot, \cdot)$ is the response modal predictor that takes as input the dialogue context U and the previous speaker's responses $R_{<j}$ and provides the modality of the next response r_j. Specifically, the predictor should output 0 when r_j is a textual utterance and 1 when r_j is a visual image.

In this paper, we design a predictor with a T5 encoder (Raffel et al., 2020) as the backbone, followed by a MLP layer, whose output dimension is 2. We concatenate all the previous turns by $[\mathrm{SEP}]$ as the input, and the images in the context are replaced by their semantically equivalent textual descriptions.

3.2.2. Textual Dialogue Response Generator

The textual dialogue response generator $\mathcal{G}$ is a standard decoder-only causal transformer model $P(R_\mathcal{G}|U, m; \theta_\mathcal{G})$. Given a dialogue context $U = \{u_1, u_2, \ldots, u_K\}$, the target is to generate a textual output $R_\mathcal{G} \in \{r^t, c\}$. Text response $r^t = \{[\mathrm{UTT}], w_1, \ldots, w_T\}$ and image description $c = \{[\mathrm{DST}], w_1, \ldots, w_L\}$ with w_i the i-th word. $[\mathrm{UTT}]$ and $[\mathrm{DST}]$ as special identifiers guide the model to distinguish whether an utterance is a text dialogue or a textual image description. The causal language modeling loss is defined as:

$$\mathcal{L}_{\text{text}} = -\sum_{i=1}^{k} \log p(w_i|U, w_1, \ldots, w_{i-1}; \theta_\mathcal{G}) \quad (2)$$

We train the textual dialogue response generator with two stages: (1) Pre-training: training it with large-scale text dialogue data $\mathcal{D}_C$ to make it have basic text dialogue generation capabilities; (2) Fine-tuning: training on a small amount of multimodal dialogue data $\widetilde{\mathcal{D}}_S$, where, for each image in $\widetilde{\mathcal{D}}_S$, we use its textual description instead to ensure that the data used in this subsection is pure text. Given a dialogue context, the textual dialogue response generator will generate descriptions besides text responses by this approach.

Inference The predicted modality m is used to guide the text generation, forcing the first generated token to be $[\mathrm{UTT}]$ if $m = t$ and $[\mathrm{DST}]$ if $m = v$.

3.2.3. Text-to-Image Translator

If the predicted response modality $m = v$, the text-to-image translator $P(r^v|c; \theta_\mathcal{F})$ generates a high-quality, high-resolution image r^v as a visual response, on conditional of the context-sensitive image description c. We adopt the diffusion model as the text-to-image translator $\mathcal{F}$.

The training procedure also has two stages: (1) Pre-training: the text-to-image translator $\mathcal{F}$ is trained on large-scale image-text pair data $\mathcal{D}_P$ to have a basic and general image generation capability; (2) Fine-tuning: training on a few image-text pairs from multimodal dialogue data $\widetilde{\mathcal{D}}_S$ to make $\mathcal{F}$ generate images with a style similar to $\widetilde{\mathcal{D}}_S$. Different from Divter (Sun et al., 2022), our model is much lighter on data requirements in the fine-tuning stage and provides better performance. We can use image captioning models (e.g., BLIP-2 (Li et al., 2023)) to generate textual descriptions of images from $\widetilde{\mathcal{D}}_S$, which can avoid situations where the image descriptions are not provided in $\widetilde{\mathcal{D}}_S$ or the image descriptions have low quality. Furthermore, achieving excellent performance in the fine-tuning stage requires very limited data, typically around 100 or even fewer instances.

Inference The design of the prompt is important for the performance of our text-to-image translator $\mathcal{F}$. BestPrompt (Pavlichenko and Ustalov, 2023) illustrates that adding additional information to the descriptions improves the quality of the generated images. The image descriptions c generated by the textual dialogue response generator are brief texts that express the main image semantics, while richer prompts will provide better visual responses.

Models	Modal	Text Response Generation			Image Description Generation			Image Generation	
	F1	BLEU-1	BLEU-2	ROUGE-L	BLEU-1	BLEU-2	ROUGE-L	IS ↑	FID ↓
Divter†	56.2	**6.52**	1.66	5.69	15.08	11.42	15.81	15.8±0.6	**29.16**
TIGER	**61.9**	6.02	**1.72**	**8.42**	**40.95**	**25.64**	**37.15**	**22.3±0.9**	42.30

Table 1: Automatic evaluation results of TIGER and baseline Divter on PhotoChat test set. † means the results are reported by Sun et al. (2022). Bolded numbers indicate statistically significant improvements (t-test with p-value < 0.01).

We add extra prompts to refine c, which leads to higher quality and stability of the generated image. Negative prompts are also adopted to improve the image quality.

3.3. Arbitrary Modal Compatibility

We utilize large-scale text dialogue data $\mathcal{D}_C$ and image-text pairs $\mathcal{D}_P$ to assist small-scale multimodal dialogue data $\widetilde{\mathcal{D}}_S$. The pre-training stages of the textual dialogue response generator and the text-to-image translator are independent, allowing for a more lightweight fusion between modalities. This means that large-scale multimodal dialogue data is not necessary.

Our framework is also suitable for multimodal dialogues that integrate arbitrary modalities (e.g., text and video, text and audio). To achieve this, one simply needs the predicted response modal $m \in \{\text{text}, \text{target modal}\}$ of the response modal predictor and replace the text-to-image translator with a Text-to-<Target Modal> translator.

4. Experiment

We conduct extensive experiments on each module to evaluate the performance of our model.

4.1. Dataset

Existing data on multimodal dialogue remains scarce, and we conduct comprehensive experiments on the PhotoChat dataset (Zang et al., 2021), which consists of 12,286 multi-turn multimodal conversations, with each conversation containing several text dialogues along with an image and its description. The dataset has been split into 10,286 train, 1,000 dev, and 1,000 test instances. We leverage BLIP-2 $\text{OPT}_{6.7\text{B}}$ (Li et al., 2023) to generate more detailed and accurate image descriptions to replace those provided by PhotoChat, which consists of multiple discrete object labels that lack information about the scene, action, etc.

Additionally, we constructed a small-scale high-quality image-text pair dataset. Specifically, we filter 120 high-quality images from the PhotoChat training set and further enhance the clarity and resolution by Gigapixel[1]. The description of each image is also generated by BLIP-2.

Models	F1	Precision	Recall
ALBERT-base*	52.2	44.8	62.7
BERT-base*	53.2	56.1	50.6
T5-base*	58.1	58.2	57.9
T5-3b*	58.9	54.1	64.6
Divter†	56.2	-	-
T5-base Encoder	**61.9**	57.8	**66.6**
T5-large Encoder	60.0	**61.5**	58.5

Table 2: Performance of response modal prediction on PhotoChat test set. * reported by Zang et al. (2021) and † reported by Sun et al. (2022).

4.2. Implementation Details

For the response modal predictor, focal loss (Lin et al., 2017) is the loss function, considering the extremely uneven distribution of labels in PhotoChat. For the textual dialogue response generator, we fine-tune DialoGPT (Zhang et al., 2020) that has been pre-trained on Reddit comment chains, and the version "DialoGPT-medium" is adopted. For the text-to-image translator, we use Stable Diffusion v2-1 as pre-trained model initialization and only fine-tune UNet, freezing the VAE and text encoder. In inference, various extra prompts are added for different categories of descriptions, and negative prompts are applied to improve image quality. More details and prompt templates can be found in our open-source repository. Our experiments are conducted on 4 NVIDIA A6000 48G GPUs.

4.3. Evaluation

We conducted a fair comparison with Divter (Sun et al., 2022), the sole small-scale model designed for MDRG as far as we know. In fairness, we exclude comparisons with MiniGPT-5, whose training data scale and model scale are much larger than our framework.

To comprehensively evaluate the performance of our model, we conduct both automatic and human evaluations. Following Sun et al. (2022), we focus

[1] https://www.topazlabs.com/gigapixel-ai

	Win (%)	Tie (%)	Lose (%)
Fidelity	71.5	24.5	4.0
Clarity	98.5	1.5	0.0
Consistency	33.0	46.5	20.5

Table 3: Human evaluation results.

on four aspects: (1) Response Modal Prediction; (2) Text Response Generation; (3) Image Description Generation; (4) Image Generation.

4.3.1. Automatic Evaluation

As shown in Table 1, TIGER demonstrates state-of-the-art performance on MDRG from most of the automatic evaluation metrics, which means that: (i) TIGER can accurately judge the timing of response with images. More detailed results are shown in Table 2; (ii) TIGER achieves comparable performance on text response generation with Divter; (iii) the generated image descriptions are more detailed and contextualized, and the generated images have better clarity and diversity. We fine-tune our model on limited data leading to a higher FID. Considering the lack of overfitting judgments and human alignment in FID and IS, we will provide a complement through human evaluation.

4.3.2. Human Evaluation

We randomly sample 200 descriptions for image generation, containing four categories: people, animals, food, and products, and each category has the same number. We assign five workers to compare all instances of images generated by both models in terms of three aspects: (1) fidelity: whether the image is realistic, i.e., whether it corresponds to what we see in our daily lives; (2) clarity; and (3) consistency: the consistency of the visual content with the textual description. The workers are graduate students in the field of Natural Language Processing (NLP), who are not involved in the development of the model and work without knowing which model the output sources from. As shown in Table 3, TIGER outperforms Divter on image generation. Specifically, TIGER owns better fidelity and clarity, which is proportional to the results of the IS metrics. For the consistency between descriptions and generated images, TIGER has superior text comprehension, which facilitates generating contextualized images and reduces the possible bias of the text-to-image process. Despite not dominating the FID metrics, these human evaluations clearly illustrate that the images generated by TIGER exhibit superior overall quality and are more favorably perceived by humans.

5. Conclusion

We introduce TIGER, a unified generative model framework for multimodal dialogue response generation. We incorporate text-to-image into text-only dialogues, enabling the original dialogue model to generate multimodal responses without the need for extensive multimodal dialogue data. Through sufficient experiments, we demonstrate that TIGER outperforms other models on various automatic evaluation metrics and is also preferred by humans, offering a more satisfying conversational experience.

6. Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62272092, No. 62172086) and the Fundamental Research Funds for the Central Universities of China (No. N2116008).

7. Bibliographical 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*.

Shen Gao, Xiuying Chen, Chang Liu, Li Liu, Dongyan Zhao, and Rui Yan. 2020. Learning to respond with stickers: A framework of unifying multi-modality in multi-turn dialog. In *Proceedings of the Web Conference 2020*, pages 1138–1148.

Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. 2023. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 6840–6851.

Yuheng Hu, Lydia Manikonda, and Subbarao Kambhampati. 2014. What we instagram: A first analysis of instagram photo content and user types. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 595–598.

Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. 2024. Generating images with multimodal

language models. *Advances in Neural Information Processing Systems*, 36.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478.

Sheetal Kusal, Shruti Patil, Jyoti Choudrie, Ketan Kotecha, Sashikala Mishra, and Ajith Abraham. 2022. Ai-based conversational agents: A scoping review from technologies to future directions. *IEEE Access*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

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

Zujie Liang, Huang Hu, Can Xu, Chongyang Tao, Xiubo Geng, Yining Chen, Fan Liang, and Daxin Jiang. 2021. Maria: A visual experience powered conversational agent. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5596–5611.

Hongpeng Lin, Ludan Ruan, Wenke Xia, Peiyu Liu, Jingyuan Wen, Yixin Xu, Di Hu, Ruihua Song, Wayne Xin Zhao, Qin Jin, et al. 2023. Tiktalk: A multi-modal dialogue dataset for real-world chitchat. *arXiv preprint arXiv:2301.05880*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Hua Lu, Zhen Guo, Chanjuan Li, Yunyi Yang, Huang He, and Siqi Bao. 2023. Towards building an open-domain dialogue system incorporated with internet memes. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. 2022. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR.

Nikita Pavlichenko and Dmitry Ustalov. 2023. Best prompts for text-to-image models and how to find them. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2067–2071.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. 2018. Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. 2021. 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.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models.

Advances in Neural Information Processing Systems, 35:25278–25294.

Heung-Yeung Shum, Xiao-dong He, and Di Li. 2018. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19:10–26.

Kurt Shuster, Eric Michael Smith, Da Ju, and Jason Weston. 2021. Multi-modal open-domain dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4863–4883.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR.

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

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, et al. 2023a. Generative multimodal models are in-context learners. *arXiv preprint arXiv:2312.13286*.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023b. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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

Xiaoxue Zang, Lijuan Liu, Maria Wang, Yang Song, Hao Zhang, and Jindong Chen. 2021. Photochat: A human-human dialogue dataset with photo sharing behavior for joint image-text modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6142–6152.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.