ASMR: Augmenting Life Scenario using Large Generative Models for Robotic Action Reflection

Shang-Chi Tsai Seiya Kawano Angel Garcia Contreras Koichiro Yoshino Yun-Nung Chen

Abstract When designing robots to assist in everyday human activities, it is crucial to enhance user requests with visual cues from their surroundings for improved intent understanding. This process is defined as a multimodal classification task. However, gathering a large-scale dataset encompassing both visual and linguistic elements for model training is challenging and time-consuming. To address this issue, our paper introduces a novel framework focusing on data augmentation in robotic assistance scenarios, encompassing both dialogues and related environmental imagery. This approach involves leveraging a sophisticated large language model to simulate potential conversations and environmental contexts, followed by the use of a stable diffusion model to create images depicting these environments. The additionally generated data serves to refine the latest multimodal models, enabling them to more accurately determine appropriate actions in response to user interactions with the limited target data. Our experimental results, based on a dataset collected from real-world scenarios, demonstrate that our methodology significantly enhances the robot's action selection capabilities, achieving the state-of-the-art performance.

Shang-Chi Tsai
National Taiwan University, e-mail: d08922014@ntu.edu.tw

Seiya Kawano
RIKEN, e-mail: seiya.kawano@riken.jp

Angel Garcia Contreras
RIKEN, e-mail: angel.garciacontreras@riken.jp

Koichiro Yoshino
RIKEN, e-mail: koichiro.yoshino@riken.jp

Yun-Nung Chen
National Taiwan University, e-mail: y.v.chen@ieee.org

1 Introduction

As robotic technology advances, robots are increasingly capable of providing a variety of services in different contexts. A key challenge in robotics is understanding and responding to human requests that are not always clear-cut, especially in life-support situations [33]. This requires interpreting both the visual context of the environment and the user's verbal communication, essentially making it a multimodal, multi-class classification problem. Recent advancements have seen the development of large multimodal language models [2, 8, 12, 16, 23], that can process and respond to multiple channels of input data.

Our study utilizes the latest multimodal model, LLaVA [16], as a foundation for predicting responsive actions to human requests. While LLaVA has shown promise in general multimodal interactions, it requires additional data to tailor its responses to specific actions in a human-robot interaction context. Gathering such interaction data is often time-intensive and not easily scalable.

Inspired by the success of the large generative model in the language [40] and vision [36, 39, 38] domains, this paper introduces a framework for automatically enhancing scenario data, specifically in contexts where a robot needs to perform life-support actions in response to human requests. We harness the power of large language models [20, 35, 1, 22], to create plausible dialogue scenarios [13, 11, 4] and describe environmental settings. These narratives are then visualized using advanced diffusion models [15, 24, 30], creating images that represent the robot's perspective during each dialogue.

By using this augmented scenario data, we can train an agent to choose appropriate actions based on everyday user interactions. This training is conducted in a controlled environment, supplemented with a small, real-world dataset collected from human-robot interactions. Our experiments demonstrate that this approach not only generates realistic scenarios but also effectively trains the multimodal language model to respond with appropriate life-support actions based on both verbal requests and environmental cues. The success of this framework highlights its potential to make robotic scenario data more scalable and relevant.

2 Related Work

In this paper, we explore the augmentation of scenario data using large generative models. We provide an overview of the relevant background in large language models (LLMs) and stable diffusion models.

2.1 Large Language Models (LLMs)

Language models have been widely studied for research in language understanding and generation, evolving from statistical to neural network-based models [40]. In recent years, the emergence of pre-trained language models (PLMs) [7, 25, 17], marked a significant advancement. These models, based on Transformer architecture and trained on vast text corpora, have shown remarkable proficiency across various natural language processing tasks. A key finding in this domain is that increasing model size enhances performance. As a result, the term "large language models" (LLMs) has been adopted to describe PLMs of substantial scale [9]. A notable example is ChatGPT [21], which has set new benchmarks in NLP tasks and demonstrates advanced linguistic capabilities in human interactions. The ongoing development and diversification of LLMs across various parameter sizes continue to be a focal point in both academic and industrial research [5, 34, 35, 6].

2.2 Large Diffusion Models

Text-to-image generation has been a significant challenge in the field of computer vision [38]. Early attempts, such as AlignDRAW [18], produced images from text but lacked realism. The introduction of Text-conditional GANs [28] marked a shift towards more sophisticated models capable of generating images from text descriptions, which is the first end-to-end architecture with characters as its input and pixels as its output. However, these GAN-based methods were limited to smaller datasets. The advent of large-scale data utilization in autoregressive models, exemplified by DALL-E [27] and Parti [37], brought improvements but at the cost of high computational demands and sequential error accumulation.

Recently, diffusion models have emerged as the new benchmark in text-to-image generation. These models can be broadly categorized based on their operational domain: pixel space or latent space. Pixel-level approaches, like GLIDE [19] and Imagen [31], generate images directly from high-dimensional data. On the other hand, latent space methods, such as stable diffusion [30] and DALL-E 2 [26], involve compressing images into a lower-dimensional space before applying the diffusion model. This innovation in model design has significantly enhanced the quality and efficiency of text-to-image generation.

3 Proposed Augmentation Framework

In our framework, we approach the challenge of robotic action determination as a multi-class classification problem. The task involves interpreting an ambiguous request $\mathbf{x}$ from a user, coupled with an image depicting the robot's view of the environment. The objective is to accurately predict a suitable action $\mathbf{y} \subseteq \mathbf{Y}$, with $\mathbf{Y}$

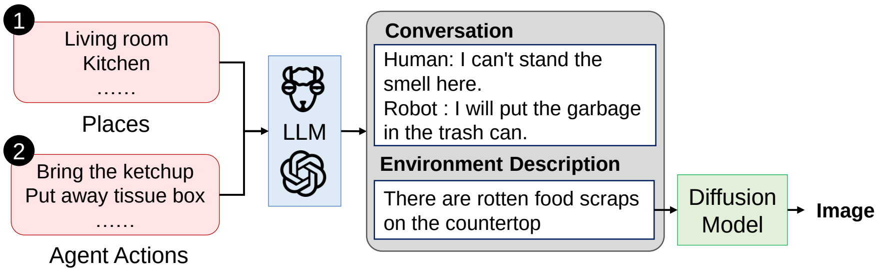


Fig. 1: Illustration of our augmentation method.

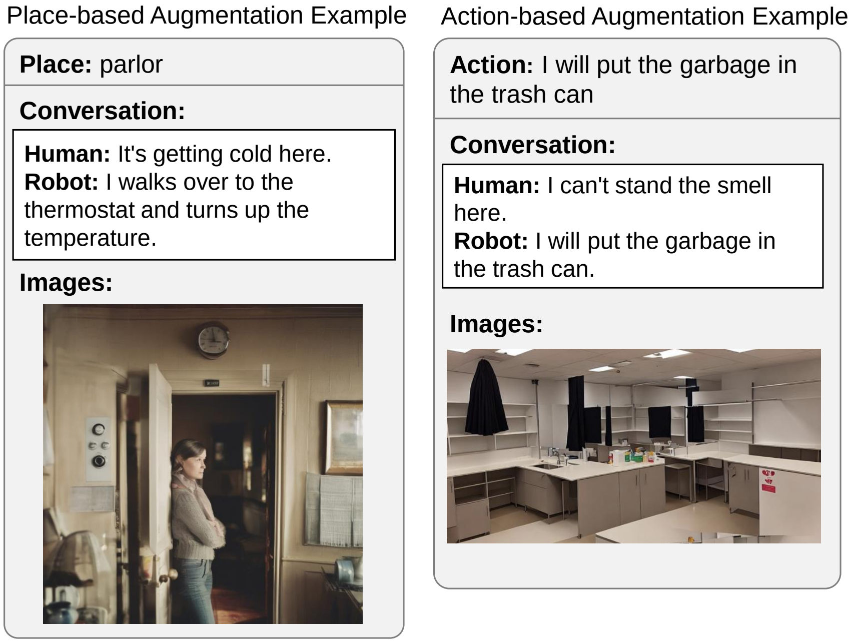
representing the set of all actions available to the robot, to assist the human user effectively.

The primary challenge of training a model to tackle this task is the time-intensive and non-scalable nature of collecting authentic interaction data between humans and robots. To address this, we have developed a framework utilizing large generative models to enrich the dataset with various potential life-support scenarios, encompassing both dialogues and environmental images. Figure 1 illustrates our augmentation pipeline. Our framework comprises two distinct pathways, each tailored to generate robotic scenarios for a specific purpose.

- **Place-based augmentation** focuses on creating dialogues pertinent to a specific location, such as a living room, kitchen, or bedroom, along with a detailed description of the respective environment.
- **Action-based augmentation** focuses on generating dialogues aligned with potential robot actions, like fetching a banana, clearing garbage, or organizing glasses, accompanied by a depiction of the setting where these actions would occur.

3.1 Place-based Augmentation

In the initial phase of our augmentation pipeline, we employ gpt-3.5, a robust large language model, to create various dialogues. These dialogues simulate scenarios where a human presents an ambiguous request in everyday settings, and a robot must respond with an appropriate service action. The process begins by selecting a commonplace setting, such as a bedroom, bathroom, or dining room—areas where robots are likely to offer routine assistance. Next, we prompt gpt-3.5 to generate potential conversations that could occur in these settings, along with descriptions of the surrounding environment. Following this, we use the stable-diffusion-XL model [24] to transform these textual descriptions into visual representations of the respective locations.


Place-based Augmentation Example
Action-based Augmentation Example
Images:
Images:

Place: parlor

Conversation:

Human: It's getting cold here.
Robot: I walks over to the thermostat and turns up the temperature.

Action: I will put the garbage in the trash can

Conversation:

Human: I can't stand the smell here.
Robot: I will put the garbage in the trash can.

Fig. 2: Example of two augmentation methods.

When crafting prompts for gpt-3.5, we do not set constraints on the generated user requests or robot actions. This approach allows the language model to conjure a wide array of scenarios, helping the model to learn and adapt to a diverse range of potential situations. For the image generation via the diffusion model, we emphasize the first-person perspective in the prompt, mirroring what the robot would observe in these environments.

We have identified ten everyday locations, each serving as the basis for generating ten distinct dialogues through the large language model. An illustrative example of this place-based augmentation process is depicted in the left part of Figure 2. The following is the prompt template used in our pipeline.

```
Give me ten conversation examples between two people in
a [location]
Person A made an ambiguous request indirectly without asking
a question to Person B
And Person B responded with a reflected action to A
Each conversation should be one utterance
And describe some related object in the background
```

3.2 Action-based Augmentation

While place-based augmentation focuses on equipping robots with the versatility to navigate various locations, action-based augmentation concentrates on creating scenarios tailored to specific, predefined robot actions [14]. In this second route of our framework, we utilize the same large language model as discussed in Section 3.1 for generating dialogues.

The key difference here lies in the nature of the input constraint. Rather than selecting a location, we choose an action from a robot's predefined action set, such as "I will clean up the table". The gpt-3.5 model is then prompted to formulate potential dialogues where this action is the appropriate response, along with descriptions of the relevant surroundings. This approach allows the model to concentrate on learning and responding to specific, realistic scenarios tied to particular actions.

To generate images that resemble real-world settings, we employ the blip diffusion model [15], known for its ability to create images with a consistent theme or subject. When generating an image from a text description, we incorporate a reference image from our real-world data collection, specifying "room" as the constant subject. This method ensures the generated images closely align with the kind of environments a robot is likely to encounter.

Our framework includes 43 distinct actions, each serving as a basis to prompt the language model to produce ten unique dialogues. An example of this action-based augmentation is showcased in the right part of Figure 2. Below is a prompt template used with the large language model for this purpose.

```
Here is a reflected action from B.
B: [reflected_action]
A is another person talking to B in a room
What ambiguous request may A talk to B indirectly without
asking a question causing B to respond above reflected action.
And describe some related object in the background according
to utterance between A and B
```

After obtaining the scenario data derived from both the place-based and action-based augmentation routes, we construct our comprehensive augmented dataset. This enriched dataset is then utilized to refine the performance of our base multimodal model, specifically designed for predicting robotic action responses. The model, adept at processing both visual and linguistic inputs, is trained to recognize and understand the scenarios presented in our augmented dataset. Upon fine-tuning this base model, we proceed to assess its proficiency in zero-shot accuracy using real-world data. This evaluation helps us measure the model's ability to accurately predict robot actions in previously unseen situations, indicating the effectiveness of our augmentation approach.

4 Experiments

To assess the impact of our augmentation data, we conducted experiments using the Do-I-Demand dataset [32], a collection of real interactive records between humans and robots. This dataset comprises 400 samples and serves as a benchmark for evaluating our method. We apply two base models with differing parameter sizes to test the efficacy of our proposed augmentation approach.

4.1 Experimental Setup

The evaluation dataset features two primary text elements: the human's ambiguous request and a description of the environment, inferred from an image. We develop two input settings based on these elements:

- **Utterance**: Here, only the human's request is used as input, with the output being one of the 43 predefined actions.
- **Utter + Description**: This setting combines the human's request with the environmental description as input, aiming to predict one of the 43 predefined actions as output.

For our experiments, we select LLaVA, a large multimodal model renowned for its multimodal chat capabilities, as our base model. LLaVA integrates a vision encoder with a large language model (LLM) to facilitate general visual and linguistic understanding. We chose its two versions, 13B and 7B parameters, for subsequent fine-tuning.

We fine-tune the base models using our augmentation dataset for five epochs, keeping the hyperparameters largely consistent with those used in the original LLaVA model. The training input comprised the image and the ambiguous human request, with the goal of maximizing the likelihood of the model predicting the correct response action. This process was carried out on 4 A6000 GPUs, each with 40GB of memory, utilizing the LoRA technique [10] for efficient training.

In evaluating the fine-tuned models, we focused on measuring zero-shot accuracy on the evaluation dataset. To match LLaVA's responses with specific actions, we employ a sentence encoder to process both the model's response and each potential action. We calculated the cosine similarity between each pair, selecting the action with the highest similarity as the final prediction. For this purpose, we experiment with two encoders: the Sentence-BERT model (SBERT) [29] and the GPT-3 model [3], both of which have shown excellent performance in various NLP benchmarks.

Model	**Utterance-Only**		**Description + Utterance**	
	SBERT	**GPT3**	**SBERT**	**GPT3**
LLaVA-13B	20.3	24.5	28.3	34.8
+ place-based augmentation	29.0†	34.3†	33.3†	39.0†
+ action-based augmentation	31.5†	31.5†	45.5†	45.5†
+ both	**36.3**†	**35.3**†	**48.5**†	**47.8**†
LLaVA-7B	19.5	22.5	27.8	36.3
+ place-based augmentation	30.3†	**36.1**†	36.0†	42.3†
+ action-based augmentation	32.3†	32.5†	41.8†	41.5†
+ both	**34.0**†	33.8†	**48.8**†	**47.5**†

Table 1: Results on the DO-I-DEMAND (%). † indicates the significant improvement achieved by the augmented data. The best score for each base predictor is marked in bold.

4.2 Results

The effectiveness of our augmentation methods on the Do-I-Demand dataset is summarized in Table 1. We evaluate the accuracy of each method by comparing the exact match rates across all labels. The baseline results, achieved using the original multi-modal model LLaVA in two distinct sizes, are presented in the first row. The data clearly indicates that both our place-based and action-based augmentation methods significantly enhance the performance of the base models. However, it is noteworthy that action-based augmentation generally outperforms place-based augmentation. This is likely because action-based augmentation is specifically tailored to align with the action categories in the evaluation dataset, whereas place-based augmentation aims to broadly improve the model's versatility in various scenarios.

Interestingly, we observe the highest performance when combining both place-based and action-based augmentations, except in one instance: the LLaVA-7B model with a GPT-3 encoder under the utterance-only setting. The top accuracy is recorded at 36.3% for LLaVA-13B with the SBERT encoder in the utterance setting, and 48.8% for LLaVA-7B with SBERT in the utterance plus description setting. These results reinforce the value of environmental descriptions in enhancing action prediction accuracy.

4.3 Effectiveness with Diverse Prompts

For place-based augmentation, our original prompt is "make an ambiguous request indirectly without asking a question". To explore variations, we test two alternative prompts: "make an ambiguous request without asking a question" and simply "make an ambiguous request." After merging data generated from all three prompts, we observe a general decline in accuracy, as shown in Table 2. This suggests that our

Model	Utterance-Only		Description + Utterance	
	SBERT	GPT3	SBERT	GPT3
LLaVA-13B (original prompt)	29.0	34.3	33.3	39.0
+ diverse prompts	27.3	32.0	35.0	43.5
LLaVA-7B (original prompt)	30.3	36.1	36.0	42.3
+ diverse prompts	23.3	25.5	34.0	42.0

Table 2: Results of the original place-based and diverse prompts on the DO-I-DEMAND (%).

Model	Utterance-Only		Description + Utterance	
	SBERT	GPT3	SBERT	GPT3
LLaVA-13B (both augmentation)	36.3	35.3	48.5	47.8
w/o blip diffusion	30.5	31.8	46.0	45.5
LLaVA-7B (both augmentation)	34.0	33.8	48.8	47.5
w/o blip diffusion	32.8	33.3	44.8	46.0

Table 3: Results of the framework with and without blip diffusion on the DO-I-DEMAND (%).

original prompt is sufficiently detailed, leading to the generation of high-quality dialogues for model training.

4.4 Ablation of Blip Diffusion

In our action-based augmentation, the use of the blip diffusion model for generating environmental images is crucial. We experiment by substituting blip diffusion with stable-diffusion-XL, as used in place-based augmentation. Table 3 reveals a consistent decrease in accuracy across all scenarios, including a notable 6% drop in the LLaVA-13B utterance setting. This highlights the significant role of the blip diffusion model in our augmentation strategy.

4.5 Effectiveness on Low-Performing Labels

Our analysis reveals that a significant portion, over one-third, of the actions predicted by the original multi-modal model perform zero accuracy. To delve into how our proposed augmentation methods impact these lower-performing labels, we group the labels into four categories based on their accuracy levels. Each group represents a quartile of performance, with bucket 1 consisting of the ten labels with the lowest accuracy, all at zero initially.

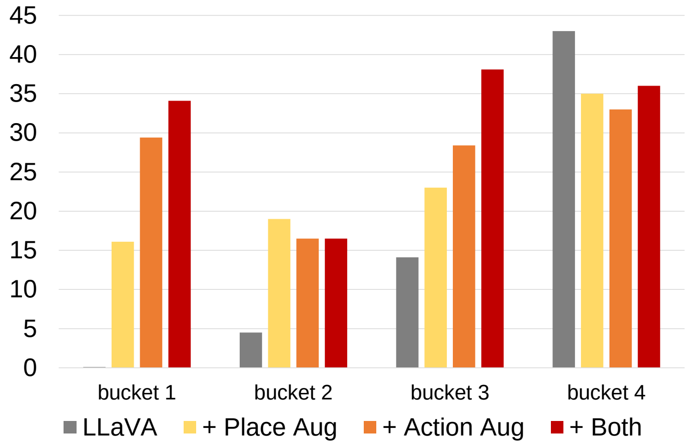


Fig. 3: performance of different buckets in utterance setting.

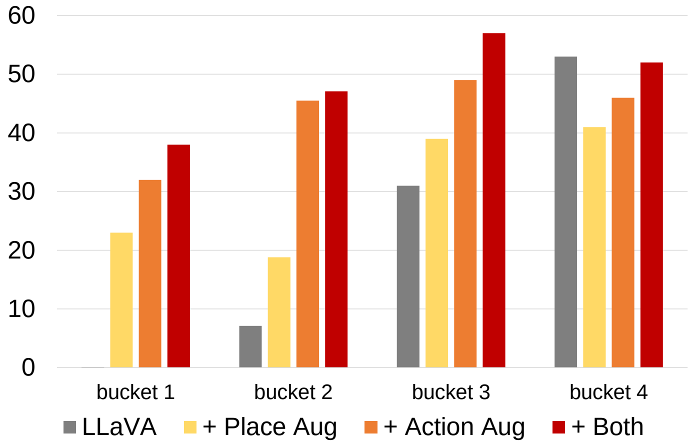


Fig. 4: performance of different buckets in utterance+description setting.

In Figure 3, we plot the mean performance of the LLaVA-1.5-13B model on the Do-I-Demand utterance set, categorized by label performance ranking. The graph clearly shows that our augmentation methods significantly improve the accuracy of labels in bucket 1. There is also a noticeable increase in accuracy across the other buckets.

Similarly, Figure 4 illustrates the mean performance on the Do-I-Demand utterance plus description set, again broken down by label performance ranking. This figure further confirms the positive impact of augmentation, especially in buckets 1 and 2, compared to the relatively lesser gains in buckets 3 and 4.

These findings underscore that each augmentation method we propose not only boosts overall performance but also effectively redistributes the performance across different labels, enhancing the model's ability to predict a wide range of actions with improved accuracy.

5 Conclusion

In this paper, we introduce a novel pipeline designed to enhance the collection of robotic life-support scenario data, traditionally a time-consuming process. Our approach leverages a large language model to simulate dialogues between humans and robots, and a large diffusion model to create corresponding images of the environments. We design two distinct types for dialogue generation: place-based augmentation, which focuses on scenarios occurring in specific places, and action-based augmentation, which centers around specific actions the robot might perform. Both approaches have proven effective in generating realistic and relevant data, significantly aiding in the training of the LLaVA model. This model is fine-tuned to predict suitable actions based on ambiguous user requests and environmental imagery. The experiments conducted on real-life collected data demonstrate that the augmented data not only significantly enhances the model's accuracy with all types of actions, especially low-performing ones, but also contributes to making robotic scenario data more scalable and adaptable. This advancement underscores the potential of our augmentation methods in enriching the training datasets for robotic action prediction models, thereby broadening their applicability in real-world scenarios.

References

1. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N.S., Chen, A.S., Creel, K.A., Davis, J., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L.E., Goel, K., Goodman, N.D., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D.E., Hong, J., Hsu, K., Huang, J., Icard, T.F., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P.W., Krass, M.S., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X.L., Li, X., Ma, T., Malik, A., Manning, C.D., Mirchandani, S.P., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J.C., Nilforoshan, H., Nyarko, J.F., Ogut, G., Orr, L., Papadimitriou, I., Park, J.S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y.H., Ruiz, C., Ryan, J., R'e, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K.P., Tamkin, A., Taori, R., Thomas, A.W., Tramèr, F., Wang, R.E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S.M., Yasunaga, M., You, J., Zaharia, M.A., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., Liang, P.: On the opportunities and risks of foundation models. ArXiv (2021). URL https://crfm.stanford.edu/assets/report.pdf

2. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., Florence, P., Fu, C., Arenas, M.G., Gopalakrishnan, K., Han, K., Hausman, K., Herzog, A., Hsu, J., Ichter, B., Irpan, A., Joshi, N., Julian, R., Kalashnikov, D., Kuang, Y., Leal, I., Lee, L., Lee, T.W.E., Levine, S., Lu, Y., Michalewski, H., Mordatch, I., Pertsch, K., Rao, K., Reymann, K., Ryoo, M., Salazar, G., Sanketi, P., Sermanet, P., Singh, J., Singh, A., Soricut, R., Tran, H., Vanhoucke, V., Vuong, Q., Wahid, A., Welker, S., Wohlhart, P., Wu, J., Xia, F., Xiao, T., Xu, P., Xu, S., Yu, T., Zitkovich, B.: Rt-2: Vision-language-action models transfer web knowledge to robotic control (2023)
3. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners (2020)
4. Chang, W.Y., Chen, Y.N.: SalesBot 2.0: A human-like intent-guided chit-chat dataset. arXiv preprint arXiv:2308.14266 (2023)
5. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A.M., Pillai, T.S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., Fiedel, N.: Palm: Scaling language modeling with pathways (2022)
6. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S.S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E.H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q.V., Wei, J.: Scaling instruction-finetuned language models (2022)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)
8. Driess, D., Xia, F., Sajjadi, M.S.M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., Florence, P.: Palm-e: An embodied multimodal language model (2023)
9. Floridi, L., Chiriatti, M.: GPT-3: Its nature, scope, limits, and consequences. Minds and Machines **30**, 681–694 (2020)
10. Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2021)
11. Huang, C.W., Hsu, C.Y., Hsu, T.Y., Li, C.A., Chen, Y.N.: CONVERSER: Few-shot conversational dense retrieval with synthetic data generation. In: Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 381–387 (2023)
12. Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O.K., Patra, B., Liu, Q., Aggarwal, K., Chi, Z., Bjorck, J., Chaudhary, V., Som, S., Song, X., Wei, F.: Language is not all you need: Aligning perception with language models (2023)
13. Kim, H., Hessel, J., Jiang, L., West, P., Lu, X., Yu, Y., Zhou, P., Bras, R., Alikhani, M., Kim, G., Sap, M., Choi, Y.: SODA: Million-scale dialogue distillation with social commonsense contextualization. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 12,930–12,949 (2023)
14. Lai, C.M., Hsu, M.H., Huang, C.W., Chen, Y.N.: Controllable user dialogue act augmentation for dialogue state tracking. In: Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 53–61 (2022)

15. Li, D., Li, J., Hoi, S.C.H.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing (2023)
16. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023)
17. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach (2019)
18. Mansimov, E., Parisotto, E., Ba, J.L., Salakhutdinov, R.: Generating images from captions with attention (2016)
19. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models (2022)
20. OpenAI: Gpt-4 technical report (2023)
21. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback (2022)
22. Peng, B., Li, C., He, P., Galley, M., Gao, J.: Instruction tuning with gpt-4 (2023)
23. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world (2023)
24. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis (2023)
25. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer (2023)
26. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents (2022)
27. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation (2021)
28. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis (2016)
29. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks (2019)
30. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2022)
31. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding (2022)
32. Tanaka, S., Yamasaki, K., Yuguchi, A., Kawano, S., Nakamura, S., Yoshino, K.: Do as i demand, not as i say: A dataset for developing a reflective life-support robot. IEEE Access **12**, 11,774–11,784 (2024). DOI 10.1109/ACCESS.2024.3350174
33. Tanaka, S., Yoshino, K., Sudoh, K., Nakamura, S.: Arta: Collection and classification of ambiguous requests and thoughtful actions (2021)
34. Thoppilan, R., Freitas, D.D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H.S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., Chen, D., Xu, Y., Chen, Z., Roberts, A., Bosma, M., Zhao, V., Zhou, Y., Chang, C.C., Krivokon, I., Rusch, W., Pickett, M., Srinivasan, P., Man, L., Meier-Hellstern, K., Morris, M.R., Doshi, T., Santos, R.D., Duke, T., Soraker, J., Zevenbergen, B., Prabhakaran, V., Diaz, M., Hutchinson, B., Olson, K., Molina, A., Hoffman-John, E., Lee, J., Aroyo, L., Rajakumar, R., Butryna, A., Lamm, M., Kuzmina, V., Fenton, J., Cohen, A., Bernstein, R., Kurzweil, R., Aguera-Arcas, B., Cui, C., Croak, M., Chi, E., Le, Q.: Lamda: Language models for dialog applications (2022)
35. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta,

R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models (2023)
36. Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., Yang, M.H.: Diffusion models: A comprehensive survey of methods and applications (2023)
37. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., Hutchinson, B., Han, W., Parekh, Z., Li, X., Zhang, H., Baldridge, J., Wu, Y.: Scaling autoregressive models for content-rich text-to-image generation (2022)
38. Zhang, C., Zhang, C., Zhang, M., Kweon, I.S.: Text-to-image diffusion models in generative ai: A survey (2023)
39. Zhang, T., Wang, Z., Huang, J., Tasnim, M.M., Shi, W.: A survey of diffusion based image generation models: Issues and their solutions (2023)
40. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.Y., Wen, J.R.: A survey of large language models (2023)