
PandaGPT: One Model To Instruction-Follow Them All

Yixuan Su^{♣,*,†}

Tian Lan*

Huayang Li^{◊,*,†}

Jialu Xu

Yan Wang

Deng Cai^{♣,*}

[♣]University of Cambridge

[◊]Nara Institute of Science and Technology

[♣]Tencent AI Lab

<https://panda-gpt.github.io/>

Abstract

We present PandaGPT, an approach to empower large language models with visual and auditory instruction-following capabilities. Our pilot experiments show that PandaGPT can perform complex tasks such as detailed image description generation, writing stories inspired by videos, and answering questions about audios. More interestingly, PandaGPT can take multimodal inputs simultaneously and compose their semantics naturally. For example, PandaGPT can connect how objects look in an image/video and how they sound in an audio. To do so, PandaGPT combines the multimodal encoders from ImageBind and the large language models from Vicuna. Notably, only aligned image-text pairs are required for the training of PandaGPT. Thanks to the strong capability of ImageBind in embedding data from different modalities into the same space, PandaGPT displays emergent, i.e. zero-shot, cross-modal behaviors for data other than image and text (e.g., video, audio, depth, thermal, and IMU). We hope that PandaGPT serves as an initial step toward building AGI that can perceive and understand inputs in different modalities holistically, as we humans do.

1 Introduction

Humans possess remarkable abilities to perceive and understand information from diverse sensory modalities, such as seeing a painting and hearing an audio guide. Analogously, to learn simultaneously, holistically, and directly from many different forms of information holds great promise for enabling machines to have a more comprehensive and better understanding of the world. To this end, there has been an emergent interest in developing artificial intelligence (AI) systems capable of perceiving and understanding information from multiple modalities simultaneously in a manner similar to humans.

However, much of the prior research has focused on tackling individual modalities in isolation. For instance, while significant progress has been made in text-to-image retrieval and generation [18], visually-grounded instruction following [12, 31], and speech understanding and generation [29], these advances have largely been confined to separate combinations of text and other modalities or, at best, a few visual modalities (e.g., image and video). These models are limited in their ability to connect information from different modalities and lack the capacity to perceive and understand

*Major contributors. Contact: ys484@cam.ac.uk and jcykcai@tencent.com.

[†]Work done during internship at Tencent AI Lab.

multimodal inputs holistically, thereby neglecting the inherent richness and complementary nature of multimodal data.

In this paper, we present PandaGPT, the first general-purpose model capable of instruction-following data from six modalities. PandaGPT leverages the power of multimodal encoders from ImageBind [8] and the expressive language models from Vicuna [4], demonstrating impressive and emergent cross-modal capabilities across six modalities: image/video, text, audio, depth, thermal, and inertial measurement units (IMU). Crucially, PandaGPT achieves these capabilities despite being only trained on aligned image-text pairs, thanks to the shared embedding space provided by ImageBind.

This integration of multimodal information enables PandaGPT to perform a wide range of tasks, including generating detailed descriptions of images, composing engaging stories inspired by videos, and providing accurate answers to questions about audio inputs. Most interestingly, the core innovation of PandaGPT lies in its ability to naturally compose the semantics of multimodal inputs, which enables a rich set of compositional multimodal tasks across different modalities. For example, it can seamlessly connect the visual appearance of objects in a photo with their corresponding sounds in an audio clip, producing a cohesive and comprehensive understanding of the scene. These cross-modal capabilities empower the model to go beyond traditional unimodal analysis. We hope PandaGPT serves as an initial step toward building AGI that can perceive and understand inputs in different modalities holistically, as humans do.

2 Related Work

Large Language Models. Large language models (LLMs) pre-trained over massive unlabeled text have dominated the field of natural language processing (NLP) today [3, 5, 19, 20]. With alignment techniques such as supervised instruction tuning [13, 21, 28] and reinforcement learning from human feedback [16, 23], LLMs exhibit surprisingly effective zero- and few-shot generalization abilities to perform almost any NLP tasks. The most successful examples could be OpenAI’s ChatGPT [15] and GPT4 [14], which have made a profound impact on the entire AI research community and beyond. There also have been enormous open-source efforts to replicate the success, such as BLOOM [22], LLaMA [27], Alpaca [26], Vicuna [4], OpenAlpaca [24] among many others.

Multi-modal Alignment. Feature alignment among multiple modalities has attracted great interest for its applications such as cross-modal retrieval [2, 6, 7]. Recently, CLIP [18] learns a joint embedding space for image and text. Flamingo [1], BLIP-2 [11], and MAGIC [25] bridge powerful pre-trained vision-only and language-only models and show strong zero-shot abilities. AudioCLIP [9] adds audio into the CLIP framework for audio classification. ImageBind [8] learn a joint embedding across six different modalities (image/video, text, audio, depth, thermal, and IMU data) using image-paired data only. More recently, there has been a surge of interest to combine multi-modal alignment and large language models for multi-modal instruction following. LLaVa [12], Mini-GPT4 [31], and Video-LLaMA [30] enable visually-grounded instruction following. DetGPT [17] proposes reasoning-based object detection. SpeechGPT [29] adds speech understanding and generation abilities to LLMs. However, these advances have largely been confined to separate combinations of text and other modalities (e.g., image/video or audio).

3 Method

PandaGPT combines the multi-modal encoders from ImageBind and the large language models from Vicuna, achieving impressive capabilities in vision- and audio-grounded instruction following tasks. To align the feature space of multimodal encoders from ImageBind and large language models from Vicuna³, we train PandaGPT using 160k image-language instruction-following data released by [12] and [31]. Each training instance consists of an image \mathcal{I} and a multi-turn conversation data $(\mathbf{x}_1, \mathbf{y}_1, \dots, \mathbf{x}_n, \mathbf{y}_n)$, where \mathbf{x}_i and \mathbf{y}_i are the human’s instruction and the system’s response at the i -th turn. To reduce the number of trainable parameters, we only train (i) a linear projection matrix f to connect the representation produced by ImageBind to Vicuna; and (ii) additional LoRA [10] weights on the Vicuna’s attention modules.⁴ Figure 1 illustrates the architecture of PandaGPT.

³We use the version-0 of Vicuna-13B as our base language model.

⁴The total number of trainable parameters is around 0.4% of the parameters of Vicuna.

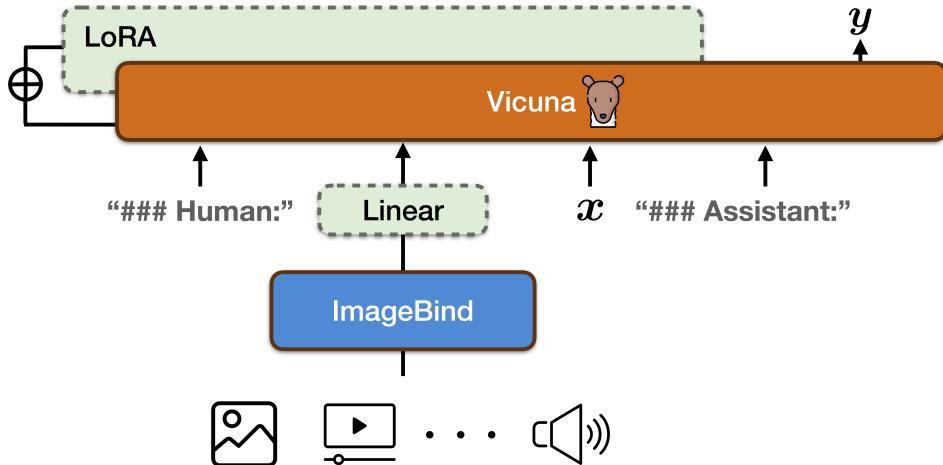


Figure 1: Illustration of PandaGPT. During training, we only train the linear projection matrix and the additional LoRA weights (as indicated with dashed boxes) while keeping the parameters of ImageBind and Vicuna frozen.

The training objective of PandaGPT is defined as

$$\mathcal{L}(\theta_f, \theta_l) = \prod_{i=1}^n p_\theta(\mathbf{y}_i | \mathbf{x}_{<i}, \mathbf{y}_{<i-1}, f(h_i)), \quad (1)$$

where θ_f and θ_l correspond to the learnable parameters of the linear projection matrix and LoRA weights. The h_i is the image representation produced by ImageBind and $\theta = \{\theta_f, \theta_l, \theta_1, \theta_2\}$, where θ_1 and θ_2 are frozen parameters of ImageBind and Vicuna. Note that the loss is only computed from the part of system responses during training. We train PandaGPT on the image-language instruction-following dataset for two epochs using a learning rate of 5e-4 with linear decay. The maximum sequence length for Vicuna-13B is set to 400 based on our computation resources ($8 \times$ A100 40G GPUs). The training takes around 7 hours to complete.

It is worth noting that the current version of PandaGPT is only trained with aligned image-text data. However, by leveraging the binding property across six modalities (image/video, text, audio, depth, thermal, and IMU) inherited from the frozen ImageBind encoders, PandaGPT demonstrates emergent, i.e. zero-shot, cross-modal capabilities across all of the modalities.

4 Capabilities of PandaGPT

Compared to existing multimodal instruction-following models trained individually for one particular modality, PandaGPT can understand and combine the information in different forms together, including image/video, text, audio, depth (3D), thermal (infrared radiation), and inertial measurement units (IMU) readings. We find that the capabilities of PandaGPT (see concrete examples in Section 6) include but are not limited to:

- **image/video-grounded question answering:** see examples of Figure 2, 3, and 4.
- **image/video-inspired creative writing:** see examples of Figure 5.
- **visual and auditory reasoning:** see examples of Figure 6, 7, and 8.
- **multimodal arithmetic:** PandaGPT is also capable of working with input composed across modalities. By arithmetically adding information from different modalities as input, PandaGPT can produce results that reflect concepts from different parts. See Figure 9 and 10 for examples of image and audio arithmetic, and see Figure 11 and 12 for examples of video and audio arithmetic.

5 Limitations

Despite the amazing ability in handling multiple modalities and their combinations. There are multiple ways to further improve PandaGPT.

1. The training of PandaGPT can be enriched by using other alignment data, for instance, other modalities paired with text (e.g., audio-text pairs).
2. We only use one embedding vector for the content in other modalities than text, more research into fine-grained feature extraction such as cross-modal attention mechanisms could be beneficial to the performance.
3. PandaGPT currently only allows multimodal information to be used as input, future possibilities include generating richer multimedia content (e.g., creating images and response in audio).
4. New benchmarks to evaluate the composition ability of multimodal inputs is demanded.
5. PandaGPT can also exhibit several common deficiencies of existing language models, including hallucination, toxicity, and stereotypes.

Lastly, we would like to note that PandaGPT is a research prototype and cannot be readily used for real-world applications.

6 Examples



Introducing this image to me in detail.



The image features a beautiful golden brown dog standing on a sandy beach near the ocean. The dog appears to be a German Shepherd, and it is enjoying the beach environment. The dog is positioned in the center of the image, with the ocean and the beach visible in the background. The dog's fur and the beach sand contrast, creating a striking scene.

Figure 2: Example showing PandaGPT’s capability in image-grounded question answering.



Who's this guy?



The guy in the image is Elon Musk, a well-known entrepreneur, inventor, and CEO of multiple companies, including SpaceX and Tesla. He is often associated with innovative projects and technological advancements.

Can you list some other companies he's involved with?



In addition to SpaceX and Tesla, Elon Musk is also involved with several other companies, including Neuralink, The Boring Company, and OpenAI.

Figure 3: Example showing PandaGPT's capability in image-grounded question answering.



A video of a Super-hero Movie.

Who are the people in this video?



The people in this video are a group of superheroes, including characters like Iron Man, Captain America, and the Hulk.

I wanna know if Spider-Man showed up in the video.



Yes, Spider-Man is present in the video along with the other superheroes.

Figure 4: Example showing PandaGPT's capability in video-grounded question answering.



A video tutorial on how
to make a delicious plate
of spaghetti.



What's the dish in the video?



The dish in the video is a stir-fried noodle dish, which appears to be a soup or a stir-fry with noodles.

How can I make this dish in the video?



To make this dish, follow these steps:

1. Cook the noodles according to the package instructions.
2. In a large pot or wok, heat some oil over medium heat.
3. Add the chopped onions and stir-fry until they are translucent.
4. Add the chopped carrots and stir-fry for a few more minutes.
5. Add the broccoli and stir-fry until it is tender.
6. Add the noodles to the pot, stirring to coat them with the vegetables and oil.
7. Add the sauce to the pot, stirring to combine with the ingredients.
8. Cook the mixture over medium heat for a few minutes, stirring occasionally, until the sauce thickens.
9. Serve the dish hot, garnished with chopped green onions or other desired toppings.

Note: The specific ingredients and sauce used in the video may vary, so adjust the recipe according to your preferences or the ingredients available.

Figure 5: Example showing PandaGPT's capability in image/video-inspired creative writing.



A video of a SpaceX
Starship rocket
launch.

Introducing this video to me



The image is a time-lapse video of a rocket launch, captured from a bird's eye view. The rocket can be seen ascending into the sky, leaving a trail of smoke behind it. The video captures the excitement and beauty of the launch, showcasing the rocket's journey from the ground to its eventual cruising altitude.

How are people feeling in the video?



The people in the video are feeling excited and thrilled as they watch the rocket launch. The rocket's launch is a significant event, and the spectators are likely filled with awe and anticipation as they witness this incredible feat of engineering and science.

So, did the rocket launch go well?



Based on the time-lapse video, it is not possible to determine the exact outcome of the rocket launch. However, the fact that the rocket is visible in the sky suggests that it has successfully lifted off and is on its way into space. The excitement and anticipation among the spectators indicate that the launch has been a success, at least from their perspective.

Figure 6: Example showing PandaGPT's capability in visual reasoning.



A dog barking audio.

Tell me what you hear about this audio.



In the audio, a dog is barking loudly, possibly due to being excited or startled. The dog's barking is heard in the background, with the sound gradually getting louder as the audio progresses. The dog's barking is the main focus of the audio, and it is clear that the dog is the source of the noise.

Can you guess what might be the reason why dogs in the audio start barking?



While it is impossible to know the exact reason for the dogs barking in the audio, there are several possibilities that could explain the sudden increase in barking. Some possible reasons include:

1. Playfulness: Dogs often bark when they are playing or engaging in playful activities with their owners or other dogs.
 2. Excitement: Dogs may bark when they are excited or curious about something new, such as a visitor, a toy, or a treat.
 3. Fear or anxiety: Dogs may bark when they are scared or anxious, such as when they hear strange noises or see unfamiliar people or animals.
 4. Separation anxiety: Dogs may bark when they are left alone or feel anxious about being separated from their owners.
 5. Communication: Dogs may bark to communicate their needs, such as wanting attention, food, or to play.
- These are just a few examples of the many reasons why dogs might bark. It is important to consider the context of the situation and the dog's behavior to determine the exact cause of the barking in this particular audio.

Figure 7: Example showing PandaGPT's capability in auditory reasoning.



The audio of a gunshot.

Describe what you hear from this audio.



The audio is a sound of a gunshot, with the distinctive cracking sound of the bullet being fired and the muffled echoes from the surrounding environment.

This is so scary, how can we avoid encountering such a thing?



To avoid encountering dangerous situations like gunshots, it is essential to prioritize safety and awareness. Here are some suggestions:

1. Stay informed: Keep yourself updated on local news, crime reports, and any potential threats in your area.
 2. Be aware of your surroundings: Pay attention to your environment and any unusual or suspicious activities.
 3. Avoid potentially dangerous situations: Avoid walking alone in isolated or high-crime areas, especially at night.
 4. Learn self-defense techniques: Knowing how to protect yourself in case of an emergency can be a valuable skill.
 5. Report any suspicious activity: If you witness or experience something unusual, immediately report it to the authorities.
 6. Install security systems: Installing security systems like alarms, cameras, and locks can help protect you and your property.
 7. Maintain a strong support system: Stay connected with friends, family, and neighbors to ensure that you have a network of people who can help you in case of an emergency.
- By following these guidelines, you can minimize the risk of encountering the gunshot.

Figure 8: Example showing PandaGPT's capability in auditory reasoning.



A recording of thunder and pouring rain.

Write a story to a child about what you see and hear.



As I look out the window, I see a beautiful scene of a person walking on the road in the distance. The person is walking under a cloudy sky, which is dark and stormy. The sound of the wind is blowing, and I can hear the distant sound of thunder rumbling in the background. The person is walking with a backpack, and it seems like they are on a journey or adventure. The stormy weather might make the journey a bit more challenging, but the person seems determined to continue their journey despite the conditions.

Figure 9: Example showing PandaGPT's capability in multimodal arithmetic (Image and Audio).



A recording of calm ocean waves.

Write a story to a child about what you see and hear.



In the image, there is a woman standing on a rock near the water, looking out at the ocean. The ocean is calm and reflects the sky, creating a serene atmosphere. The woman is enjoying the view and listening to the soothing sound of the waves crashing against the shore. As she stands there, she notices a bird flying by, adding a touch of life and movement to the scene. The woman feels peaceful and content, taking in the beauty of the ocean and the natural world around her.

What color is the woman's hair?



The woman has blonde hair.

Figure 10: Example showing PandaGPT's capability in multimodal arithmetic (Image and Audio).



A video about a couple taking a stroll together.

A recording of thunder and pouring rain.

Now, I am closing my eyes. Please describe what you can see and hear for me.



As you close your eyes, you can still see the image of the two people walking down the street at night, carrying umbrellas to protect themselves from the rain. The streetlights illuminate the scene, casting a warm glow on the people and the surrounding area. You can hear the sound of the rain falling gently on the umbrellas, creating a soothing and calming atmosphere. The sound of footsteps on the pavement complements the overall ambiance of the scene.

Figure 11: Example showing PandaGPT's capability in multimodal arithmetic (Video and Audio).

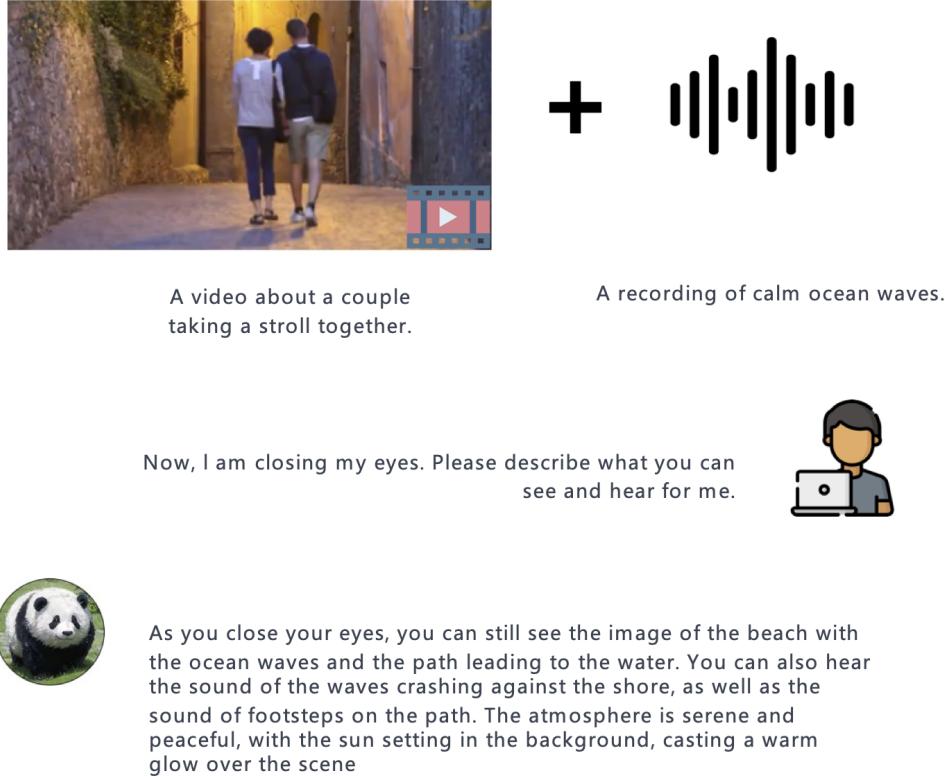


Figure 12: Example showing PandaGPT's capability in multimodal arithmetic (Video and Audio).

References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.
- [2] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramanujam, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. 2020. Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems*, 33:25–37.
- [3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- [4] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.
- [5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- [6] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.
- [7] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc’Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26.
- [8] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. *arXiv preprint arXiv:2305.05665*.
- [9] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE.
- [10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- [11] 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*.
- [12] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.
- [13] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Natural instructions: Benchmarking generalization to new tasks from natural language instructions. *arXiv preprint arXiv:2104.08773*, pages 839–849.
- [14] OpenAI. 2022. Gpt-4 technical report.
- [15] OpenAI. 2022. Introducing chatgpt.
- [16] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

- [17] Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Lingpeng Kong, and Tong Zhang. 2023. Detgpt: Detect what you need via reasoning.
- [18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- [19] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.
- [20] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- [21] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.
- [22] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Lucioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.
- [23] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.
- [24] Yixuan Su, Tian Lan, and Deng Cai. 2023. Openalpaca: A fully open-source instruction-following model based on openllama. <https://github.com/yxuansu/OpenAlpaca>.
- [25] Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. 2022. Language models can see: plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*.
- [26] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.
- [27] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- [28] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.
- [29] Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities.
- [30] Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-finetuned visual language model for video understanding.
- [31] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.