PointPrompt: A Multi-modal Prompting Dataset for Segment Anything Model

Jorge Quesada, Mohammad Alotaibi, Mohit Prabhushankar, Ghassan AlRegib

OLIVES Lab, Georgia Institute of Technology
Atlanta, GA 30332, USA
{jpacora3, malotaibi44, mohit.p, alregib}@gatech.edu

Abstract

The capabilities of foundation models, most recently the Segment Anything Model, have gathered a large degree of attention for providing a versatile framework for tackling a wide array of image segmentation tasks. However, the interplay between human prompting strategies and the segmentation performance of these models remains understudied, as does the role played by the domain knowledge that humans (by previous exposure) and models (by pretraining) bring to the prompting process. To bridge this gap, we present the `PointPrompt` *dataset compiled across multiple image modalities as well as multiple prompting annotators per modality. We collected a total of 16 image datasets from the natural, underwater, medical and seismic domain in order to create a comprehensive resource to facilitate the study of prompting behavior and agreement across modalities. Overall, our prompting dataset contains 158880 inclusion points and 52594 exclusion points over a total of 6000 images. Our analysis highlights the following: (i) viability of prompts across heterogeneous data, (ii) that point prompts are a valuable resource in the effort for enhancing the robustness and generalizability of segmentation models across diverse domains, (iii) prompts facilitate an understanding of the dynamics between annotation strategies and neural network outcomes. Information on downloading the dataset, images, and prompting tool is provided on our project website* `https://alregib.ece.gatech.edu/pointprompt/`.

1. Introduction

Efficient and accurate segmentation of different objects in an image is a fundamental task in computer vision, impacting a wide array of domains, from natural image understanding [17, 21] to medical diagnosis [18, 27] and seismic interpretation [13, 30]. In recent years, the rapid collection of growing volumes of image data in various fields coupled with the accelerated progress of deep learning methods led to the development of foundation models. In particular, the Segment Anything Model (SAM) [12] has emerged as a highly flexible segmentation alternative due to its prompting-based flexibility and intuitive use.

However, there are limited studies on principled or efficient prompting strategies, and whether these strategies are transferable across different imaging modalities. Moreover, it has been observed that simply prompting indiscriminately often leads to over-prompting, where an excessive amount of cues leads to poor segmentation performance. While there are works that adapt SAM to a particular domain by finetuning its components on target datasets [11, 20, 28], the base zero-shot transfer capabilities and prompting dynamics of SAM remain poorly understood. This knowledge gap severely hinders the principled transferability of SAM to domains where large amounts of data are not available to finetune it.

To address this gap, we present the `PointPrompt` dataset, a collection of prompting data derived across multiple image modalities and multiple annotators per modality. The images, prompting data and software are available on our project website. Our dataset was collected by showing one of 16 image datasets from the natural, underwater, medical and seismic domain each to a different set of annotators for them to perform point-based prompts on, and collating their responses across 400 images for each dataset. This dataset constitutes a unified resource to study multimodal prompting patterns and advance the development of efficient domain transfer techniques for foundation models.

2. Related Work

Foundation models are distinguished by their extensive training on a large scale of data, enabling generalizability across different domains [5, 31]. Although these models are based on standard ML algorithms, their capabilities are often far more versatile than domain-specific (e.g., fine-tuned) models [19]. These models are often implemented with a prompting interface to users, in which users can guide these models to generate the desired output [31].

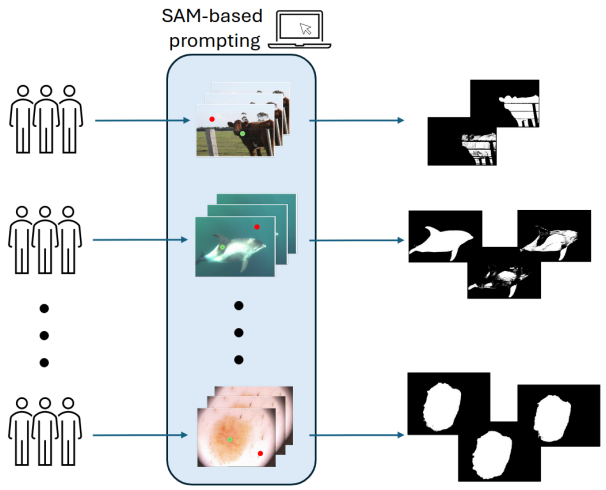

Figure 1. A conceptual summary of our prompting dataset. Multiple sets of annotators (left) are provided with different image datasets across multiple imaging modalities (middle) to prompt using our SAM-based segemntation tool, and produce a sequence of masks (right), each with their own associated prompts and segmentation scores

The revealing of the Transformer architecture [26] has marked a milestone in the development of foundation models, and more specifically Natural language processing (NLP) models. The 345M parameter BERT model [6] stands as a pioneering example of foundation models in the application of NLP. The Generative Pre-trained Transformer (GPT) models [23] stand as another major milestone in the development of foundation models. The most recent GPT-4 model [22] demonstrated state-of-the-art results in different downstream NLP tasks.

Foundation models have been developed for computer vision tasks as well, albeit to a lesser extent compared to their linguistic counterparts. One of the earliest implementations of these models are the text-to-image models (e.g. DALL.E [24], and Midjourney [9]) that convert the textual description of a scene into visuals. In image segmentation, the Segment anything model (SAM) [12] is marked as the first foundation model designated for this task. SAM has been trained on over 1 million images and 1 billion masks to achieve generalizability across different domains. SAM engages with its users through prompts, which can be in various formats, including inclusion and exclusion points, bounding boxes, masks, or even free-form text.

The interaction between foundation models and users via the prompting interface introduces an element of uncertainty, as the precise response of these models to user prompts can be unpredictable. This uncertainty has given rise to the concept of prompt engineering, which aims to

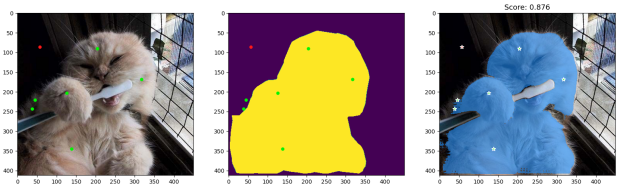

Figure 2. Sample use case of the prompting interface used to collect the data. Left: Original image. Middle: Ground-truth segmentation (acquired from source database). Right: SAM-based segmentation. In this example, a user provides 7 prompts (6 inclusion points in green and 1 exclusion point in red), leading to an IOU score of 0.874.

achieve better results by crafting better prompts. Yet, defining what makes a prompt "better" in the algorithmic perception of a model remains unclear, thus motivating scholars to explore ways of optimizing it. Moreover, the diversity of users engaging with these models—each bringing their own unique perspectives and problem-solving approaches shaped by their individual social experiences—adds another layer of complexity [29].

This uncertainty has motivated scholars to study the impact of the prompts on the model's output [7, 15], and to design prompting tools to analyze how users approach these models through prompts [3, 16]. These studies, however, are limited to the textual prompts. We hypothesize that the use of visual prompts by users will fundamentally differ from textual prompts, which necessitates this line of research.

3. Data Acquisition Methodology

In order to build our prompting dataset, we first curated 16 single-class image datasets across different modalities in order to achieve a high level of heterogeneity in our prompts. Each of these image datasets was collected by randomly sampling 400 images along with their corresponding ground-truth segmentation masks from an existing open-source dataset as detailed below:

- 9 datasets were extracted from the COCO [14] database, each corresponding to a single category: dog, cat, bird, clock, bus, baseball bat, cow, tie, stop sign.
- 2 datasets were extracted from the NDD20 [25] database: dolphins above water and underwater.
- 3 medical imaging datasets: Chest-X [1] (chest tumors), Kvasir-SEG [10] (polyp images) and ISIC [8] (skin lesions).
- 2 seismic imaging datasets extracted from the F3 Facies database [2], corresponding to the salt dome and chalk group categories.

From all of the above described datasets, the seismic datasets are the only ones that contain 200 images rather than 400, due to structural constraints in the presence of

the desired categories across the volumetric seismic data. We gather prompting data by presenting each of these 16 datasets to separate groups of 3 to 4 annotators, each of which generates individual prompts for their corresponding dataset, leading to multiple prompting annotations per category.

In order for the annotators to be able to interact with each of these datasets, we developed a SAM-based prompting tool that allowed annotators to add prompts in the form of inclusion (inside the region of interest) and exclusion (outside the region of interest) points through clicks in an image (the tool is accessible through our project website). Figure 2 shows the standard interface of the tool.

This prompting tool updated the SAM-generated segmentation in a live fashion (after at least 1 inclusion and 1 exclusion points were provided), allowing the annotators to visualize the result of each consecutive prompt and adjust their strategies accordingly. Annotators were generally instructed to try to maximize their Intersection Over Union (IOU) score, calculated by comparing the SAM-generated mask with the ground-truth mask, which was displayed on top of the segmentation result after each prompt. Once they were satisfied with their results, users could close the interface and move on to the next image in the set.

Given this live score update scheme, we saved each set of prompts, their respective IOU scores, and the corresponding SAM-generated masks at each timestep, which allows for a detailed study of prompt progression and strategy through time. We notice that oftentimes, adding more points may hinder rather than improve segmentation performance. We account for this by closing a round of prompts if the IOU score did not improve after 5 consecutive prompts and restarting another prompting round for the corresponding image from scratch (for a maximum of two rounds per image). Since the data in both rounds (when available) is saved, this allows for a study both on updated prompting strategies as well as on the difference in prompting difficulty or uncertainty that different modalities might entail.

4. Dataset and Discussion

We showcase in Figure 3 some of the different image modalities in our dataset, along with two sample prompt-based segmentations for each image. The first row shows the segmentations performed by two different annotators, and it shows how an excessive amount of prompts can be highly detrimental to the resulting segmentation mask. The second row shows consecutive prompts performed by the same annotator, which differ only in a single point (the second one has an additional green point on the tail of the dolphin) and it shows a surprising fact: very similar prompts can lead to significantly different masks. The third row also shows two dissimilar sets of prompts performed by different annotators, which generate qualitatively similar masks.

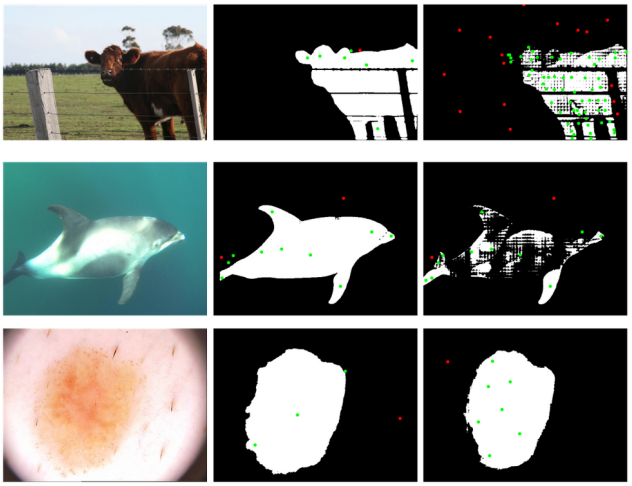

Figure 3. Image examples of the different prompting modalities and segmentations in our dataset. The first column corresponds to an image from a given modality (natural, underwater and medical) and the two other columns corresponds to different prompts of that image and their corresponding segmentation. First row: dissimilar prompts, dissimilar mask. Second row: similar prompts, dissimilar mask. Third row: dissimilar prompts, similar mask.

These examples illustrate that prompt and mask similarity are not always correlated and there are inherent properties of different types of image datasets that also affect the mask generation process.

Figure 4 depicts the average amount of total prompting timestamps per image (blue), and the average amount of prompts until achieving the optimal score per image (orange) for each of the image categories in the dataset. Note that "optimal score" refers to the mask that attains the highest IOU with respect to the ground truth. We can see that for every category, on average, people prompt approximately twice as much as they need to, or by a range of approximately 2 to 9 unnecessary prompts. This further corroborates the issue showcased in the first two rows of Figure 1, in which people arrive at poor segmentation masks by prompting excessively.

In Figure 5, we provide statistics on the number of inclusion (blue) and exclusion (orange) prompting points at the optimal segmentation (highest IOU score) timestamp for each category. As intuition would dictate, we can see that the number of inclusion points significantly dominates the amount of exclusion points. An interesting pattern however, is that in many cases the total amount of prompts (taken by summing the number of inclusion and exclusion points) does not closely resemble the average amount of optimal prompting timpestamps (orange bars in Figure 4), implying that many prompts are being performed solely on the first timestamp.

We analyze this pattern more closely in Figure 6, in

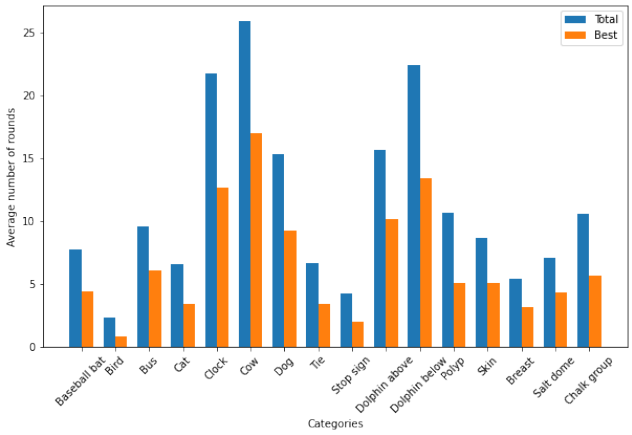

Figure 4. Average amount of total (blue) and optimal (orange) prompting timestamps for each image category.

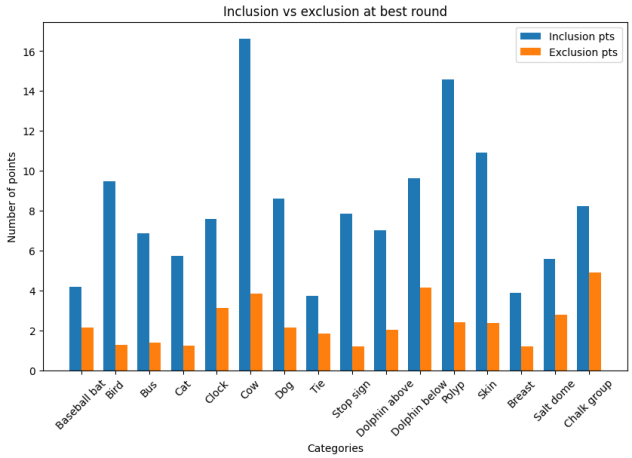

Figure 5. Average number of inclusion (blue) and exclusion (orange) points for each image category at the optimal segmentation performance.

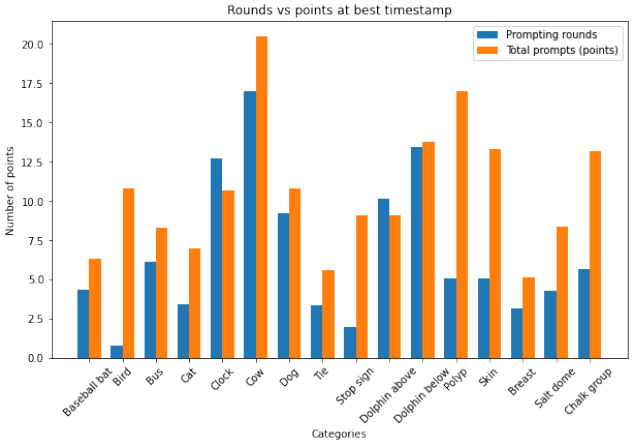

Figure 6. Total number of prompting timestamps (orange) compared against total number of prompts (blue).

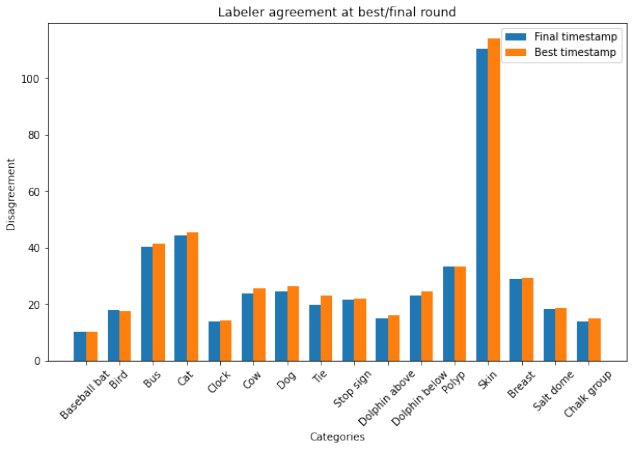

Figure 7. Average disagreement across different datasets for last (blue) and best (orange) timestamp.

which we directly compare the total number of prompting timestamps (orange bars) against the total number of prompts (blue bars). A large difference between these two bars (as in the case of the bird, stop sign and polyp categories) implies that annotators perform more than 10 prompts on average for these images before even seeing the first round of 'feedback' in the form of a result generated by SAM. In contrast, cases where these two bars are very similar (like the dolphins or clock categories) imply that annotators start off with a single pair of inclusion/exclusion points and continue prompting based on the visual result returned by SAM. This patterns implies that prompting heavily early on before seeing the result stems either from an established degree of confidence in the domain one is prompting on (and therefore an expectation of a good segmentation on that first attempt), or from a preconceived notion of complexity in the image one is looking at, which is preemptively accounted for by performing an initially large number of prompts. These questions could be further explored by analyzing the relationship between the structures at the dataset and image levels and the prompting patterns elicited by the data.

In Figure 7 we compare the level of disagreement between the prompts performed by the annotators within each dataset. The disagreement is calculated for each possible pair of annotators within a dataset using the Chamfer distance [4] between two sets of points (each corresponding to the prompt locations) $P_1 = \{x_i \in \mathbb{R}^2\}_{i=1}^{n}$ and

$P2 = \{x_j \in \mathbb{R}^2\}_{j=1}^{m}$ as:

$$d(P_1, P_2) = \frac{1}{2n}\sum_{i=1}^{n} ||x_i - \text{NN}(x_i, P_2)||_2 + \frac{1}{2m}\sum_{j=1}^{m} ||x_j - \text{NN}(x_j, P_1)||_2 \quad (1)$$

where $\text{NN}(x, P)$ denotes the nearest neighbor of x in the point set P. For a given pair of annotators, we calculate the disagreement by taking the average of this metric across all prompted images. The final disagreement for each dataset is then computed by averaging the disagreement between all possible pairwise combinations of annotators in that dataset. For this metric we consider only the inclusion points, given that the exclusion points are much more arbitrary (since they correspond to the conceptual background) and lead to a disagreement almost an order of magnitude higher. From the plot, it is interesting to observe that the skin lesion dataset is the most controversial one by a large margin, even though it has on average far fewer points than (for instance) the cow dataset, which although has on average more prompts (see Figure 5), is drastically more agreed on. We also want to highlight that both the final (blue) and best (orange) disagreements are extremely close for all datasets, which when contrasted with the data on Figure 4 implies that even when adding excessive prompts, annotators do it in very similar and consistent manners. This seems to suggest overprompting is highly homogeneous (it does not increase disagreement once an optimal segmentation has been achieved), which provides further evidence on the value of studying the structure of overprompting patterns in order to develop strategies to alleviate or prevent it.

5. Conclusions and future work

In this paper, we introduced a comprehensive image prompting dataset based on SAM, comprising multiple imaging modalities and several human annotators per modality, in order to provide a resource to improve the understanding of the complex interplay between prompting strategies and segmentation results, and the role domain knowledge (in humans and models) plays into this dynamic. Our dataset provides insights into prompting behaviour across multiple imaging categories, and presents opportunities for improving zero-shot transfer capabilities of foundation models in domains where large amounts of labeled data are not necessarily available for finetuning.

We are actively working on expanding this dataset, and plan to acquire prompting data in other modalities, as well as re-prompt on the already gathered modalities after finetuning SAM to each specific category, in order to better study how prompting strategies evolve as models receive more exposure to specific domains. The dataset, images and software can be accessed at https://alregib.ece.gatech.edu/pointprompt/.

References

[1] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020. 2

[2] Yazeed Alaudah, Patrycja Michałowicz, Motaz Alfarraj, and Ghassan AlRegib. A machine-learning benchmark for facies classification. *Interpretation*, 7(3):SE175–SE187, 2019. 2

[3] Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan Fries, Maged S. Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-Jian Jiang, and Alexander M. Rush. Promptsource: An integrated development environment and repository for natural language prompts, 2022. 2

[4] P.J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992. 4

[5] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 1

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 2

[7] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models, 2020. 2

[8] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016. 2

[9] David Holz, 2022. 2

[10] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, pages 451–462. Springer, 2020. 2

[11] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1, 2

[13] Kiran Kokilepersaud, Mohit Prabhushankar, and Ghassan AlRegib. Volumetric supervised contrastive learning for seismic semantic segmentation. In *Second International Meeting for Applied Geoscience & Energy*, pages 1699–1703. Society of Exploration Geophysicists and American Association of Petroleum . . . , 2022. 1

[14] 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. 2

[15] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, 2021. 2

[16] Vivian Liu and Lydia B. Chilton. Design guidelines for prompt engineering text-to-image generative models, 2023. 2

[17] Xiaolong Liu, Zhidong Deng, and Yuhan Yang. Recent progress in semantic image segmentation. *Artificial Intelligence Review*, 52:1089–1106, 2019. 1

[18] Xiangbin Liu, Liping Song, Shuai Liu, and Yudong Zhang. A review of deep-learning-based medical image segmentation methods. *Sustainability*, 13(3):1224, 2021. 1

[19] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024. 1

[20] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 1

[21] Hossein Mobahi, Shankar R Rao, Allen Y Yang, Shankar S Sastry, and Yi Ma. Segmentation of natural images by texture and boundary compression. *International journal of computer vision*, 95(1):86–98, 2011. 1

[22] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. 2

[23] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018. 2

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

Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. 2

[25] Cameron Trotter, Georgia Atkinson, Matt Sharpe, Kirsten Richardson, A Stephen McGough, Nick Wright, Ben Burville, and Per Berggren. Ndd20: A large-scale few-shot dolphin dataset for coarse and fine-grained categorisation. *arXiv preprint arXiv:2005.13359*, 2020. 2

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. 2

[27] Risheng Wang, Tao Lei, Ruixia Cui, Bingtao Zhang, Hongying Meng, and Asoke K Nandi. Medical image segmentation using deep learning: A survey. *IET Image Processing*, 16(5): 1243–1267, 2022. 1

[28] Junde Wu, Rao Fu, Huihui Fang, Yuanpei Liu, Zhaowei Wang, Yanwu Xu, Yueming Jin, and Tal Arbel. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023. 1

[29] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. Why johnny can't prompt: How non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2023. Association for Computing Machinery. 2

[30] Tao Zhao and Xiaoli Chen. Enrich the interpretation of seismic image segmentation by estimating epistemic uncertainty. In *SEG Technical Program Expanded Abstracts 2020*, pages 1444–1448. Society of Exploration Geophysicists, 2020. 1

[31] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1