Do Vision-Language Models See Dwarf Galaxies the Way We Do?

Dimitrios Tanoglidis
Independent Researcher
tanoglidisdimitrios@gmail.com

Chin Yi Tan
Kavli Institute for Cosmological Physics, University of Chicago
Department of Physics, University of Chicago
NSF-Simons AI Institute for the Sky (SkAI)
chinyi@uchicago.edu

Kate Overdeck
Department of Astronomy & Astrophysics, University of Chicago
NSF-Simons AI Institute for the Sky (SkAI)
koverdeck@uchicago.edu

Alex Drlica-Wagner
Fermi National Accelerator Laboratory
Kavli Institute for Cosmological Physics, University of Chicago
Department of Astronomy & Astrophysics, University of Chicago
NSF-Simons AI Institute for the Sky (SkAI)
kadrlica@fnal.gov

Abstract

With the advent of powerful, general-purpose vision-language models (VLMs), there has been growing interest in their potential to assist astronomical discovery, a field characterized by large volumes of image data. In this work, we evaluate VLMs on the challenging task of identifying ultra-faint dwarf galaxy candidates using multi-panel diagnostic images from survey data. We compare model predictions to human annotations from a large-scale citizen science campaign. We find that zero-shot VLMs closely reproduce aggregate human calibration and perform well on less ambiguous cases. However, there is significant variability at the level of individual examples, and attempts to obtain uncertainty estimates (via self-reported confidence or repeated inference) fail to yield reliable and practically useful measures. Our results highlight both the promise and the current limitations of deploying VLMs for large-scale scientific discovery in realistic settings.

1 Introduction

Large, multimodal, Vision-Language Models (VLMs; e.g., [2]) have demonstrated strong performance across a wide range of scientific benchmarks, multimodal reasoning tasks, and real-world applications(e.g., [7, 11, 13, 22, 23]). As general-purpose foundation models[1] pre-trained on large-scale datasets, they can be rapidly adapted to a variety of downstream tasks (e.g., image classification)

Conference on Physics and AI at Stanford University (PAI 2026).

with minimal task-specific engineering. Importantly, for VLMs, this adaptation can be achieved through natural language instructions, optionally augmented with a small number of example images.

Given their generality and flexibility, there has been growing interest in applying VLMs to astronomical data analysis, where observations are predominantly image-based. Modern astronomical surveys produce vast quantities of imaging data (e.g., [10]), and identifying scientifically interesting objects typically requires inspecting, and rejecting, a large number of contaminants (e.g., artifacts or spurious detections).

While custom deep learning models are widely used, they require substantial development effort, large labeled datasets that are often unavailable, and can inherit biases from their training data (especially when synthetic data are used). Citizen science campaigns (e.g., Zooniverse) provide an alternative, but designing, deploying, and analyzing such campaigns also requires substantial effort.

The application of VLMs to astronomy has been explored in recent work [5, 6, 9, 12, 14, 16, 17, 19, 21, 24]. However, these studies have largely focused on relatively clean, single-image datasets, where the classification task is comparatively straightforward. This leaves open the question of how VLMs perform in more complex, realistic settings that require integrating multiple diagnostic plots.

In this work, we leverage the outputs of a recent citizen science campaign aimed at discovering ultra-faint dwarf galaxies in wide-field survey data. Each candidate is evaluated using multiple diagnostic plots, requiring the integration of complementary visual cues. We ask how well general-purpose VLMs match the behavior of citizen scientists, particularly in ambiguous or challenging cases.

We find that VLMs, without task-specific training, recover aggregate human calibration well, but exhibit substantial per-example variability and systematic mismatches in confidence. These results highlight both the promise and the limitations of using general-purpose VLMs for scientific discovery tasks involving complex, multi-view reasoning.

2 Methods

Data: We use images and annotations from a recently completed Zooniverse[1] campaign aimed at discovering ultra-faint Milky Way satellite galaxies in the wide-field DELVE survey, which covers $\sim 20{,}000, \mathrm{deg}^2$ of the southern sky [8]. The observed abundance of dwarf galaxies orbiting the Milky Way provides strong constraints on dark matter models, making their discovery both scientifically important and observationally challenging [4, 15, 20]. In practice, the search is dominated by large numbers of spurious candidates, necessitating either custom deep learning pipelines or large-scale citizen science efforts to reduce the candidate pool.

The campaign presented 46,334 subjects to more than 1,650 human volunteers. Each subject received multiple (25) independent votes, resulting in over 1.1 million classifications. The subjects include 43,677 candidates from the data, 2,600 simulated galaxies, and 57 known dwarf galaxies. Each subject consists of a set of diagnostic plots (Fig. 1), accompanied by instructions to distinguish genuine dwarf galaxy candidates from artifacts ("junk"). Classification requires integrating information across multiple panels, which makes this a more realistic and challenging setting than standard single-image benchmarks.

Models and Prompting: In this work, we use the multimodal GPT-5-mini model[18] for all experiments. This model provides a practical balance between performance, speed, and cost – an important consideration in realistic scenarios, where a VLM-based approach must be both accurate and scalable to large candidate sets. We leverage the OpenAI Batch API, which enables asynchronous processing of large classification workloads at reduced cost.

We focus on a zero-shot prompting setup, using text-only instructions that describe how to distinguish genuine dwarf galaxies from artifacts. These instructions closely mirror those provided to human annotators in the citizen science campaign. While VLMs can also incorporate visual examples (few-shot prompting), we defer this direction to future work in order to isolate the capabilities of instruction-only guidance.

All prompts and code used in this work are provided in the accompanying GitHub repository[2].

[1] `https://www.zooniverse.org/`

[2] `https://github.com/dtanoglidis/Zooniverse-Dwarf-Multimodal`

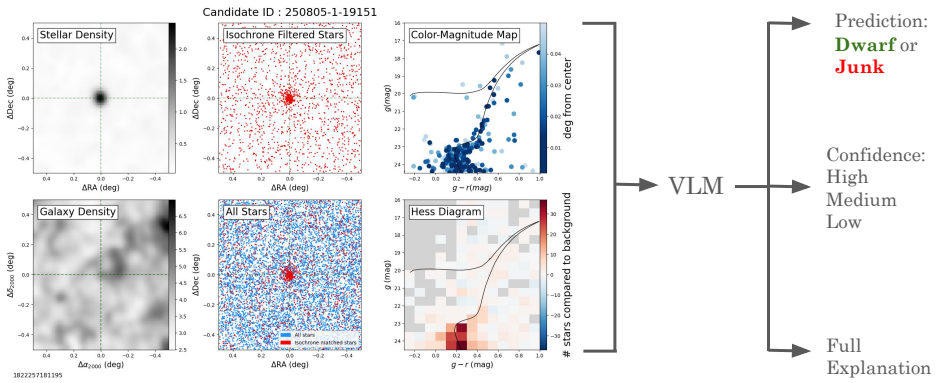
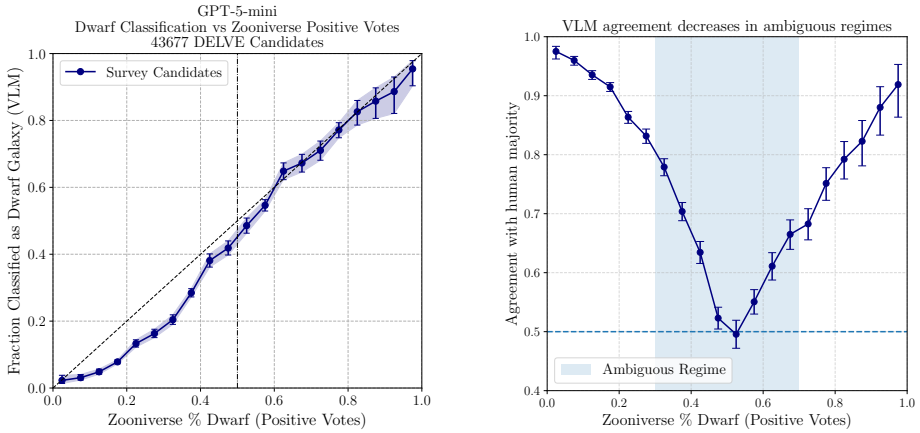


Figure 1: Example of the diagnostic panel shown to both Zooniverse volunteers and the VLM. This instance corresponds to the known dwarf galaxy Horologium I, for which all panels display a clear positive signal. The VLM outputs a binary prediction (Dwarf/Junk), an associated confidence level, and a free-form explanation.

3 Results



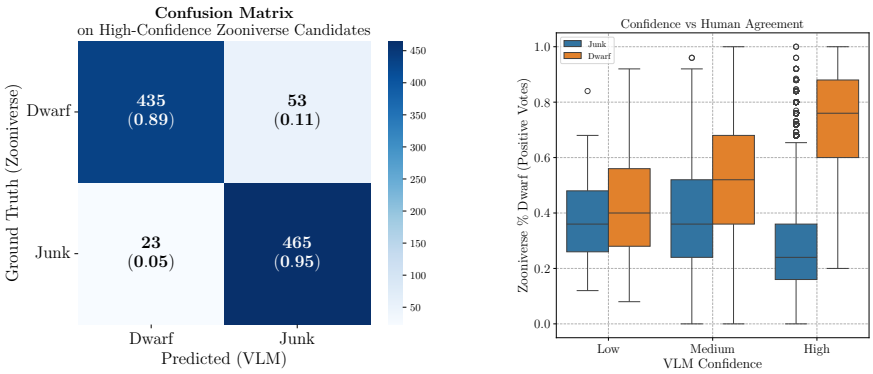
(a) (b)

Figure 2: (a) VLM dwarf classification fraction vs. human positive vote fraction, showing strong agreement at the population level. (b) Agreement between VLM and human labels decreases in ambiguous regimes.

Using GPT-5-mini with a zero-shot instruction prompt, we classify the 43,677 DELVE survey candidates, yielding 12,599 labeled as Dwarf. We compare these predictions to the fraction of positive ("is Dwarf") votes from Zooniverse volunteers (Fig. 2). Binning candidates by human positive fraction (Fig. 2a), the fraction classified as Dwarf by the VLM closely matches the human signal, indicating good aggregate calibration. In Fig. 2b we binarize the human labels (positive fraction > 0.5 as Dwarf, < 0.5 as Junk) and compute agreement with the VLM across bins. Agreement is strongest in low-ambiguity regimes (very low/high vote fractions). Furthermore, the model correctly identifies 55/57 known dwarfs, with the two failures corresponding to challenging cases, even for humans.

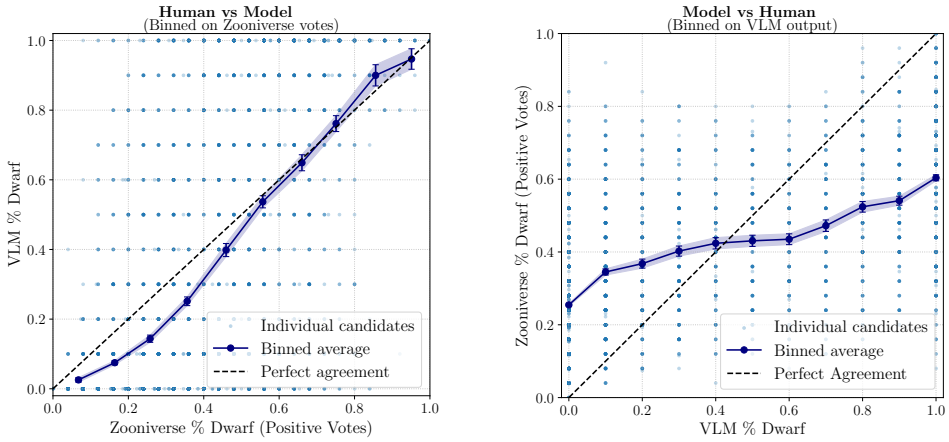
To further quantify performance, we construct a high-confidence subset using human votes (> 0.85 for dwarf, < 0.15 for junk). On this subset, the VLM achieves 95% precision and 89% recall (Fig. 3a), consistent with prior work on clear, less-ambiguous astronomical classification problems.

For VLM classifiers to be useful in ambiguous and hard cases, it is important to obtain a meaningful and reliable score together with the binary label. Such a score enables targeted follow-up, focusing on the highest-confidence candidates—analogous to selecting objects with high human agreement. In



(a) (b)

Figure 3: (a) Confusion matrix on high-confidence human-labeled candidates, showing high precision and recall. (b) VLM confidence correlates with human votes, but high-confidence predictions are sparse.



(a) (b)

Figure 4: (a) Binned on human votes: model frequency tracks human consensus. (b) Binned on model output: human agreement does not map cleanly to model frequency.

Fig. (3b), the model's self-reported confidence correlates with human vote fraction. However, only a small fraction (135 of those classified as Dwarf) of the examples are assigned High confidence, limiting their practical utility.

As an alternative, we estimate uncertainty via repeated inference. We select a random subset of 4,000 candidates (for speed and cost) and run the VLM 10 times per example, defining a model positive fraction (fraction of runs classified as Dwarf). In Fig. 4, we compare this quantity to human votes. In Fig. 4a, we bin by human positive fraction and compute the mean VLM dwarf frequency in each bin. The solid points show these means, with error bars given by the standard error of the mean ($\sigma/\sqrt{n}$), reflecting the uncertainty on the average within each bin. In this view, the model frequency increases with human consensus, indicating good average calibration. In contrast, Fig. 4b bins by the VLM output (model positive fraction) and computes the corresponding mean human vote fraction. Here, while the mean human vote fraction increases with VLM "Dwarf" classification frequency, the mapping is not well calibrated: each VLM bin contains a broad spread of human votes (low % and high %), reducing its usefulness as a probabilistic estimate.

Overall, while VLMs reproduce aggregate human trends and perform well on clear cases, their outputs remain poorly calibrated at the level of individual, ambiguous examples.

4 Discussion and Conclusions

In this work, we compare the performance of VLMs in discovering ultra-faint dwarf galaxies in survey data against human annotations from a large-scale citizen science campaign. We find that VLMs perform strongly on clear, low-ambiguity cases and reproduce aggregate human behavior, suggesting they can serve as a scalable first-pass filter. However, reliable uncertainty estimation remains a key limitation. Both self-reported confidence and repeated inference fail to provide well-calibrated, useful scores, limiting the direct use of VLM outputs for candidate prioritization and follow-up.

In future work we plan to explore alternative proprietary and open-source VLMs (or even results based on an ensemble of VLM, see for example [6]), few-shot [3] visual prompting, and improved calibration methods. We also plan to investigate human–AI collaborative workflows in which VLM predictions are used to prioritize candidate review and guide scientific annotation efforts. Overall, while VLMs show clear promise, robust uncertainty and per-example reliability remain open challenges for deployment in scientific workflows.

References

[1] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models. *arXiv e-prints*, art. arXiv:2108.07258, August 2021. doi: 10.48550/arXiv.2108.07258.

[2] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C. Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, Mark Ibrahim, Melissa Hall, Yunyang Xiong, Jonathan Lebensold, Candace Ross, Srihari Jayakumar, Chuan Guo, Diane Bouchacourt, Haider Al-Tahan, Karthik Padthe, Vasu Sharma, Hu Xu, Xiaoqing Ellen Tan, Megan Richards, Samuel Lavoie, Pietro Astolfi, Reyhane Askari Hemmat, Jun Chen, Kushal Tirumala, Rim Assouel, Mazda Moayeri, Arjang Talattof, Kamalika Chaudhuri, Zechun Liu, Xilun Chen, Quentin Garrido, Karen Ullrich, Aishwarya Agrawal, Kate Saenko, Asli Celikyilmaz, and Vikas Chandra. An Introduction to Vision-Language Modeling. *arXiv e-prints*, art. arXiv:2405.17247, May 2024. doi: 10.48550/arXiv.2405.17247.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. *arXiv e-prints*, art. arXiv:2005.14165, May 2020. doi: 10.48550/arXiv.2005.14165.

[4] James S. Bullock and Michael Boylan-Kolchin. Small-Scale Challenges to the ΛCDM Paradigm. , 55(1): 343–387, August 2017. doi: 10.1146/annurev-astro-091916-055313.

[5] Claire Chen, Jiabao Sean Xiao, Shuze Daniel Liu, Facundo Perez Paolino, Luke Handley, Theophile Jegou du Laz, Ricky Nilsson, Alice Zou, Matthew Graham, and Ashish Mahabal. AstroAlertBench: Evaluating the Accuracy, Reasoning, and Honesty of Multimodal LLMs in Astronomical Classification. *arXiv e-prints*, art. arXiv:2605.05573, May 2026. doi: 10.48550/arXiv.2605.05573.

[6] Marco Chiaberge, Elias Stengel-Eskin, Massimo Stiavelli, and Colin Norman. Vision-Language Model Ensembles Achieve Human-Expert Accuracy for Galaxy Merger Classification. *arXiv e-prints*, art. arXiv:2606.00415, May 2026. doi: 10.48550/arXiv.2606.00415.

[7] Avi Cooper, Keizo Kato, Chia-Hsien Shih, Hiroaki Yamane, Kasper Vinken, Kentaro Takemoto, Taro Sunagawa, Hao-Wei Yeh, Jin Yamanaka, Ian Mason, and Xavier Boix. Rethinking VLMs and LLMs for Image Classification. *arXiv e-prints*, art. arXiv:2410.14690, October 2024. doi: 10.48550/arXiv.2410.14690.

[8] A. Drlica-Wagner, J. L. Carlin, D. L. Nidever, P. S. Ferguson, N. Kuropatkin, M. Adamów, W. Cerny, Y. Choi, J. H. Esteves, C. E. Martínez-Vázquez, S. Mau, A. E. Miller, B. Mutlu-Pakdil, E. H. Neilsen, K. A. G. Olsen, A. B. Pace, A. H. Riley, J. D. Sakowska, D. J. Sand, L. Santana-Silva, E. J. Tollerud, D. L. Tucker, A. K. Vivas, E. Zaborowski, A. Zenteno, T. M. C. Abbott, S. Allam, K. Bechtol, C. P. M. Bell, E. F. Bell, P. Bilaji, C. R. Bom, J. A. Carballo-Bello, D. Crnojević, M.-R. L. Cioni, A. Diaz-Ocampo, T. J. L. de Boer, D. Erkal, R. A. Gruendl, D. Hernandez-Lang, A. K. Hughes, D. J. James, L. C. Johnson, T. S. Li, Y.-Y. Mao, D. Martínez-Delgado, P. Massana, M. McNanna, R. Morgan, E. O. Nadler, N. E. D. Noël, A. Palmese, A. H. G. Peter, E. S. Rykoff, J. Sánchez, N. Shipp, J. D. Simon, A. Smercina, M. Soares-Santos, G. S. Stringfellow, K. Tavangar, R. P. van der Marel, A. R. Walker, R. H. Wechsler, J. F. Wu, B. Yanny, M. Fitzpatrick, L. Huang, A. Jacques, R. Nikutta, A. Scott, and Astro Data Lab. The DECam Local Volume Exploration Survey: Overview and First Data Release. , 256(1):2, September 2021. doi: 10.3847/1538-4365/ac079d.

[9] Mariia Drozdova, Erica Lastufka, Vitaliy Kinakh, Taras Holotyak, Daniel Schaerer, and Slava Voloshynovskiy. Radio Astronomy in the Era of Vision-Language Models: Prompt Sensitivity and Adaptation. *arXiv e-prints*, art. arXiv:2509.02615, August 2025. doi: 10.48550/arXiv.2509.02615.

[10] Željko Ivezić, Steven M. Kahn, J. Anthony Tyson, Bob Abel, Emily Acosta, Robyn Allsman, David Alonso, Yusra AlSayyad, Scott F. Anderson, John Andrew, James Roger P. Angel, George Z. Angeli, Reza Ansari, Pierre Antilogus, Constanza Araujo, Robert Armstrong, Kirk T. Arndt, Pierre Astier, Éric Aubourg, Nicole Auza, Tim S. Axelrod, Deborah J. Bard, Jeff D. Barr, Aurelian Barrau, James G. Bartlett, Amanda E. Bauer, Brian J. Bauman, Sylvain Baumont, Ellen Bechtol, Keith Bechtol, Andrew C. Becker, Jacek Becla, Cristina Beldica, Steve Bellavia, Federica B. Bianco, Rahul Biswas, Guillaume Blanc, Jonathan Blazek, Roger D. Blandford, Josh S. Bloom, Joanne Bogart, Tim W. Bond, Michael T. Booth, Anders W. Borgland, Kirk Borne, James F. Bosch, Dominique Boutigny, Craig A. Brackett, Andrew Bradshaw, William Nielsen Brandt, Michael E. Brown, James S. Bullock, Patricia Burchat, David L. Burke, Gianpietro Cagnoli, Daniel Calabrese, Shawn Callahan, Alice L. Callen, Jeffrey L. Carlin, Erin L. Carlson, Srinivasan Chandrasekharan, Glenaver Charles-Emerson, Steve Chesley, Elliott C. Cheu, Hsin-Fang Chiang, James Chiang, Carol Chirino, Derek Chow, David R. Ciardi, Charles F. Claver, Johann Cohen-Tanugi, Joseph J. Cockrum, Rebecca Coles, Andrew J. Connolly, Kem H. Cook, Asantha Cooray, Kevin R. Covey, Chris Cribbs, Wei Cui, Roc Cutri, Philip N. Daly, Scott F. Daniel, Felipe Daruich, Guillaume Daubard, Greg Daues, William Dawson, Francisco Delgado, Alfred Dellapenna, Robert de Peyster, Miguel de Val-Borro, Seth W. Digel, Peter Doherty, Richard Dubois, Gregory P. Dubois-Felsmann, Josef Durech, Frossie Economou, Tim Eifler, Michael Eracleous, Benjamin L. Emmons, Angelo Fausti Neto, Henry Ferguson, Enrique Figueroa, Merlin Fisher-Levine, Warren Focke, Michael D. Foss, James Frank, Michael D. Freemon, Emmanuel Gangler, Eric Gawiser, John C. Geary, Perry Gee, Marla Geha, Charles J. B. Gessner, Robert R. Gibson, D. Kirk Gilmore, Thomas Glanzman, William Glick, Tatiana Goldina, Daniel A. Goldstein, Iain Goodenow, Melissa L. Graham, William J. Gressler, Philippe Gris, Leanne P. Guy, Augustin Guyonnet, Gunther Haller, Ron Harris, Patrick A. Hascall, Justine Haupt, Fabio Hernandez, Sven Herrmann, Edward Hileman, Joshua Hoblitt, John A. Hodgson, Craig Hogan, James D. Howard, Dajun Huang, Michael E. Huffer, Patrick Ingraham, Walter R. Innes, Suzanne H. Jacoby, Bhuvnesh Jain, Fabrice Jammes, M. James Jee, Tim Jenness, Garrett Jernigan, Darko Jevremović, Kenneth Johns, Anthony S. Johnson, Margaret W. G. Johnson, R. Lynne Jones, Claire Juramy-Gilles, Mario Jurić, Jason S. Kalirai, Nitya J. Kallivayalil, Bryce Kalmbach, Jeffrey P. Kantor, Pierre Karst, Mansi M. Kasliwal, Heather Kelly, Richard Kessler, Veronica Kinnison, David Kirkby, Lloyd Knox, Ivan V. Kotov, Victor L. Krabbendam, K. Simon Krughoff, Petr Kubánek, John Kuczewski, Shri Kulkarni, John Ku, Nadine R. Kurita, Craig S. Lage, Ron Lambert, Travis Lange, J. Brian Langton, Laurent Le Guillou, Deborah Levine, Ming Liang, Kian-Tat Lim, Chris J. Lintott, Kevin E. Long, Margaux Lopez, Paul J. Lotz, Robert H. Lupton, Nate B. Lust, Lauren A. MacArthur, Ashish Mahabal, Rachel Mandelbaum, Thomas W. Markiewicz, Darren S. Marsh, Philip J. Marshall, Stuart Marshall, Morgan May, Robert McKercher, Michelle McQueen, Joshua Meyers, Myriam Migliore, Michelle Miller, and David J. Mills. LSST: From Science Drivers to Reference Design and Anticipated Data Products. , 873(2):111, March 2019. doi: 10.3847/1538-4357/ab042c.

[11] Kento Kawaharazuka, Tatsuya Matsushima, Andrew Gambardella, Jiaxian Guo, Chris Paxton, and Andy Zeng. Real-World Robot Applications of Foundation Models: A Review. *arXiv e-prints*, art. arXiv:2402.05741, February 2024. doi: 10.48550/arXiv.2402.05741.

[12] Nolan Koblischke, Liam Parker, Francois Lanusse, Irina Espejo Morales, Jo Bovy, and Shirley Ho. Semantic search for 100M+ galaxy images using AI-generated captions. *arXiv e-prints*, art. arXiv:2512.11982, December 2025. doi: 10.48550/arXiv.2512.11982.

[13] Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A Survey of State of the Art Large Vision Language Models: Alignment, Benchmark, Evaluations and Challenges. *arXiv e-prints*, art. arXiv:2501.02189, January 2025. doi: 10.48550/arXiv.2501.02189.

[14] Siddharth Mishra-Sharma, Yiding Song, and Jesse Thaler. PAPERCLIP: Associating Astronomical Observations and Natural Language with Multi-Modal Models. *arXiv e-prints*, art. arXiv:2403.08851, March 2024. doi: 10.48550/arXiv.2403.08851.

[15] E. O. Nadler, A. Drlica-Wagner, K. Bechtol, S. Mau, R. H. Wechsler, V. Gluscevic, K. Boddy, A. B. Pace, T. S. Li, M. McNanna, A. H. Riley, J. García-Bellido, Y.-Y. Mao, G. Green, D. L. Burke, A. Peter, B. Jain, T. M. C. Abbott, M. Aguena, S. Allam, J. Annis, S. Avila, D. Brooks, M. Carrasco Kind, J. Carretero, M. Costanzi, L. N. da Costa, J. De Vicente, S. Desai, H. T. Diehl, P. Doel, S. Everett, A. E. Evrard, B. Flaugher, J. Frieman, D. W. Gerdes, D. Gruen, R. A. Gruendl, J. Gschwend, G. Gutierrez, S. R. Hinton, K. Honscheid, D. Huterer, D. J. James, E. Krause, K. Kuehn, N. Kuropatkin, O. Lahav, M. A. G. Maia, J. L. Marshall, F. Menanteau, R. Miquel, A. Palmese, F. Paz-Chinchón, A. A. Plazas, A. K. Romer, E. Sanchez, V. Scarpine, S. Serrano, I. Sevilla-Noarbe, M. Smith, M. Soares-Santos, E. Suchyta, M. E. C. Swanson, G. Tarle, D. L. Tucker, A. R. Walker, W. Wester, and DES Collaboration. Constraints on Dark Matter Properties from Observations of Milky Way Satellite Galaxies. , 126(9):091101, March 2021. doi: 10.1103/PhysRevLett.126.091101.

[16] Wenke Ren, Hengxiao Guo, Wenwen Zuo, and Xiaoman Zhang. A systematic evaluation of vision-language models for observational astronomical reasoning tasks. *arXiv e-prints*, art. arXiv:2604.24589, April 2026. doi: 10.48550/arXiv.2604.24589.

[17] Simone Riggi, Thomas Cecconello, Andrea Pilzer, Simone Palazzo, Nikhel Gupta, Andrew Hopkins, Corrado Trigilio, and Grazia Umana. radio-llava: Advancing vision-language models for radio astronomical source analysis. , 42:e121, August 2025. doi: 10.1017/pasa.2025.10082.

[18] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, Alex Makelov, Alex Neitz, Alex Wei, Alexandra Barr, Alexandre Kirchmeyer, Alexey Ivanov, Alexi Christakis, Alistair Gillespie, Allison Tam, Ally Bennett, Alvin Wan, Alyssa Huang, Amy McDonald Sandjideh, Amy Yang, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrei Gheorghe, Andres Garcia Garcia, Andrew Braunstein, Andrew Liu, Andrew Schmidt, Andrey Mereskin, Andrey Mishchenko, Andy Applebaum, Andy Rogerson, Ann Rajan, Annie Wei, Anoop Kotha, Anubha Srivastava, Anushree Agrawal, Arun Vijayvergiya, Ashley Tyra, Ashvin Nair, Avi Nayak, Ben Eggers, Bessie Ji, Beth Hoover, Bill Chen, Blair Chen, Boaz Barak, Borys Minaiev, Botao Hao, Bowen Baker, Brad Lightcap, Brandon McKinzie, Brandon Wang, Brendan Quinn, Brian Fioca, Brian Hsu, Brian Yang, Brian Yu, Brian Zhang, Brittany Brenner, Callie Riggins Zetino, Cameron Raymond, Camillo Lugaresi, Carolina Paz, Cary Hudson, Cedric Whitney, Chak Li, Charles Chen, Charlotte Cole, Chelsea Voss, Chen Ding, Chen Shen, Chengdu Huang, Chris Colby, Chris Hallacy, Chris Koch, Chris Lu, Christina Kaplan, Christina Kim, CJ Minott-Henriques, Cliff Frey, Cody Yu, Coley Czarnecki, Colin Reid, Colin Wei, Cory Decareaux, Cristina Scheau, Cyril Zhang, Cyrus Forbes, Da Tang, Dakota Goldberg, Dan Roberts, Dana Palmie, Daniel Kappler, Daniel Levine, Daniel Wright, Dave Leo, David Lin, David Robinson, Declan Grabb, Derek Chen, Derek Lim, Derek Salama, Dibya Bhattacharjee, Dimitris Tsipras, Dinghua Li, Dingli Yu, DJ Strouse, Drew Williams, Dylan Hunn, Ed Bayes, Edwin Arbus, Ekin Akyurek, Elaine Ya Le, Elana Widmann, Eli Yani, Elizabeth Proehl, Enis Sert, Enoch Cheung, Eri Schwartz, Eric Han, Eric Jiang, Eric Mitchell, Eric Sigler, Eric Wallace, Erik Ritter, Erin Kavanaugh, Evan Mays, Evgenii Nikishin, Fangyuan Li, Felipe Petroski Such, Filipe de Avila Belbute Peres, Filippo Raso, Florent Bekerman, Foivos Tsimpourlas, Fotis Chantzis, Francis Song, Francis Zhang, Gaby Raila, Garrett McGrath, Gary Briggs, Gary Yang, Giambattista Parascandolo, Gildas Chabot, Grace Kim, Grace Zhao, Gregory Valiant, Guillaume Leclerc, Hadi Salman, Hanson Wang, Hao Sheng, Haoming Jiang, Haoyu Wang, Haozhun Jin, Harshit Sikchi, Heather Schmidt, Henry Aspegren, Honglin Chen, Huida Qiu, Hunter Lightman, Ian Covert, Ian Kivlichan, Ian Silber, Ian Sohl, Ibrahim Hammoud, Ignasi Clavera, Ikai Lan, Ilge Akkaya, Ilya Kostrikov, Irina Kofman, Isak Etinger, Ishaan Singal, Jackie Hehir, Jacob Huh, Jacqueline Pan, Jake Wilczynski, Jakub Pachocki, James Lee, James Quinn, Jamie Kiros, Janvi Kalra, Jasmyn Samaroo, Jason Wang, Jason Wolfe, Jay Chen, Jay Wang, and Jean Harb. OpenAI GPT-5 System Card. *arXiv e-prints*, art. arXiv:2601.03267, December 2025. doi: 10.48550/arXiv.2601.03267.

[19] Fiorenzo Stoppa, Turan Bulmus, Steven Bloemen, Stephen J. Smartt, Paul J. Groot, Paul Vreeswijk, and Ken W. Smith. Textual interpretation of transient image classifications from large language models. *Nature Astronomy*, 9:1869–1878, December 2025. doi: 10.1038/s41550-025-02670-z.

[20] C. Y. Tan, A. Drlica-Wagner, A. B. Pace, W. Cerny, E. O. Nadler, A. Doliva-Dolinsky, D. Anbajagane, T. S. Li, J. D. Simon, A. K. Vivas, A. R. Walker, M. Adamów, K. Bechtol, J. L. Carlin, Q. O. Casey, C. Chang, A. Chaturvedi, T.-Y. Cheng, A. Chiti, Y. Choi, D. Crnojević, P. S. Ferguson, R. A. Gruendl, A. P. Ji, G. Limberg, G. E. Medina, B. Mutlu-Pakdil, N. E. D. Noël, K. Overdeck, V. M. Placco, A. H. Riley, D. J. Sand, J. Sharp, N. F. Sherman, G. S. Stringfellow, R. H. Wechsler, M. Aguena, S. Allam, O. Alves, D. Bacon, D. Brooks, D. L. Burke, R. Camilleri, J. A. Carballo-Bello, A. Carnero Rosell, J. Carretero, L. N. da Costa, M. E. da Silva Pereira, T. M. Davis, J. de Vicente, S. Desai, S. Everett, B. Flaugher, J. Frieman, J. García-Bellido, D. Gruen, G. Gutierrez, K. Herner, S. R. Hinton, D. L. Hollowood, D. J. James, K. Kuehn, O. Lahav, S. Lee, J. L. Marshall, C. E. Martínez-Vázquez, P. Massana, J. Mena-Fernández, R. Miquel, J. Muir, J. Myles, R. L. C. Ogando, A. A. Plazas Malagón, A. Porredon, E. Sanchez, D. Sanchez Cid, I. Sevilla-Noarbe, M. Smith, E. Suchyta, M. E. C. Swanson, C. To, E. J. Tollerud, D. L. Tucker, V. Vikram, N. Weaverdyck, M. Yamamoto, A. Zenteno, Delve Collaboration, and Des Collaboration. DELVE Milky Way Satellite Galaxy Census. I. Satellite Population and Survey Selection Function in DES, DELVE, and Pan-STARRS. , 1000(1):87, March 2026. doi: 10.3847/1538-4357/ae4479.

[21] Dimitrios Tanoglidis and Bhuvnesh Jain. At First Sight! Zero-shot Classification of Astronomical Images with Large Multimodal Models. *Research Notes of the American Astronomical Society*, 8(10):265, October 2024. doi: 10.3847/2515-5172/ad887a.

[22] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). *arXiv e-prints*, art. arXiv:2309.17421, September 2023. doi: 10.48550/arXiv.2309.17421.

[23] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. *arXiv e-prints*, art. arXiv:2311.16502, November 2023. doi: 10.48550/arXiv.2311.16502.

[24] Sharaf Zaman, Michael J. Smith, Pranav Khetarpal, Rishabh Chakrabarty, Michele Ginolfi, Marc Huertas-Company, Maja Jabłońska, Sandor Kruk, Matthieu Le Lain, Sergio José Rodríguez Méndez, and Dimitrios Tanoglidis. AstroLLaVA: towards the unification of astronomical data and natural language. *arXiv e-prints*, art. arXiv:2504.08583, April 2025. doi: 10.48550/arXiv.2504.08583.