# An Open-Source Software Toolkit & Benchmark Suite for the Evaluation and Adaptation of Multimodal Action Models

**Pranav Guruprasad** [*1 2] **Yangyue Wang** [*1 2] **Sudipta Chowdhury** [2] **Jaewoo Song** [2] **Harshvardhan Sikka** [1 2 3]

## Abstract

Recent innovations in multimodal action models represent a promising direction for developing general-purpose agentic systems, combining visual understanding, language comprehension, and action generation. We introduce MultiNet - a novel, fully open-source benchmark and surrounding software ecosystem designed to rigorously evaluate and adapt models across vision, language, and action domains. We establish standardized evaluation protocols for assessing vision-language models (VLMs) and vision-language-action models (VLAs), and provide open source software to download relevant data, models, and evaluations. Additionally, we provide a composite dataset with over a trillion tokens of image captioning, visual question answering, commonsense reasoning, robotic control, digital game-play, simulated locomotion/manipulation, and many more tasks. The MultiNet benchmark, framework, toolkit, and evaluation harness have been used in downstream research on the limitations of VLA generalization.

## 1. Introduction

Recent advances in machine learning have demonstrated the potential of large-scale models to exhibit broad generalization capabilities across diverse tasks. Vision-Language-Action (VLA) models in particular have showcased impressive abilities in grounding real-world actions with visual perception and natural language understanding (Kim et al., 2024) (Black et al., 2024) (Ichter et al., 2025). These models can interpret complex visual scenes, comprehend natural language commands, and generate appropriate control sequences. Nevertheless, contemporary VLA models are predominantly engineered for, and evaluated on, narrow domains such as robotic manipulation (Guruprasad et al., 2024). Their efficacy on multifaceted vision-language understanding tasks or their capacity for broader generalization remains largely unproven, a limitation potentially stemming from training regimens that prioritize specific capabilities over holistic, generalist behavior.

Developing functional agents capable of generalizing to a broad set of action tasks necessitates training on vast, diverse datasets encompassing multiple modalities (vision, language, control) and a wide array of task types. Such agents must not only excel within individual modalities but also seamlessly integrate information and execute actions across them, mirroring the multifaceted demands of real-world scenarios (Guruprasad et al., 2025). Presently, the community lacks a large-scale, open-source benchmark specifically architect ed for the rigorous training and comprehensive evaluation of these ambitious generalist models. Most benchmarks focus on narrow, very specific domains and tasks, and are typically closed-source (White et al., 2024) (Hendrycks et al., 2021b) (Liang et al., 2021) (Gulcehre et al., 2021) (Fu et al., 2021). This critical gap motivates our work.

In this paper, we introduce the MultiNet ecosystem - a comprehensive set of benchmarking software designed to catalyze the development and evaluation of generalist action models. Our contributions are multifaceted:

- **A Large-Scale Generalist Dataset**: Curation, standardization, and release of an extensive, open-source dataset that amalgamates diverse data sources (vision, language, and action) suitable for training and evaluating generalist models as seen in A.2.
- **An Open-Source Data Curation SDK**: Open-source access to a software toolkit to facilitate easy access to the consolidated dataset. This toolkit also standardizes control data (Reinforcement Learning and Robotics data) from a myriad of sources into a common, accessible format.
- **A Systematic Evaluation Harness**: A standardized, well-reasoned evaluation methodology including test splits and carefully designed metrics as seen in A.3.

[1]Metarch [2]Manifold Research [3]Georgia Institute of Technology. Correspondence to: Pranav Guruprasad <pranav@metarch.ai>.

*Proceedings of the ICML 2025 Workshop on Championing Open-source Development in Machine Learning (CODEML '25).* Copyright 2025 by the author(s).

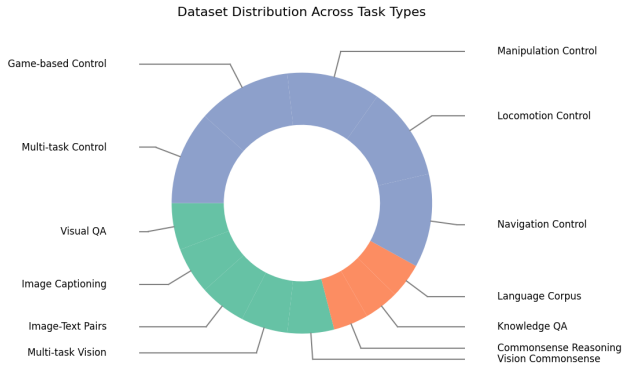


*Figure 1.* Control represents the largest portion (58%) due to the extensive OpenX-Embodiment collection, followed by Vision-Language (29%) and Language (13%) datasets

This framework is specifically engineered to provide easy access to the community to assess the generalization capabilities of state-of-the-art Vision-Language Models (VLMs) and Vision-Language-Action (VLA) models across a spectrum of familiar and novel domains, including real-world robotics tasks, and procedurally generated game environments.

- **Open-source adaptations for SoTA VLA models and VLMs**: Open-sourced adaptations of SoTA VLMs and VLAs, enabling them to operate effectively on the data formats and diverse domains in MultiNet, even those unseen during their original training, to further accelerate progress toward building generalist AI systems.
- **In-depth Experiments and Analysis**: Utilization of the MultiNet benchmark, framework, evaluation harness, and model adaptations to obtain evaluations and analysis of the performance of leading VLMs, VLAs, and emerging generalist models.

Through MultiNet, we aim to provide the community with the essential resources—datasets, tools, and standardized evaluation protocols—to systematically compare different approaches, gain deeper insights into the challenges and opportunities in building generalist AI, and ultimately accelerate the development of truly general-purpose intelligent systems.

## 2. A Large-scale Open-Source Generalist Dataset

Through this work, we provide a comprehensive dataset collection to facilitate the training and evaluation of generalist AI models, and fully open-source easy access to it. Our collection aggregates a diverse range of existing datasets, encompassing various domains such as vision-language understanding, language processing, reinforcement learning, and robotics. This collection provides a rich resource by bringing together data across multiple domains, modalities and tasks, aiming to support the development of more capable and versatile AI systems.

The collection includes a variety of datasets focused on different aspects of vision-language understanding: OBELICS (Laurençon et al., 2023), an open web-scale dataset of 141 million interleaved image-text web pages, 353 million images, and 115 billion text tokens; DataComp-1B (Gadre et al., 2023), a curated set of 1.4 billion image-text pairs; COYO-700M (coy), containing 747 million image-alt-text pairs with minimal filtering; MS-COCO Captions (Lin et al., 2015), featuring 330,000 images each with 5 captions; Conceptual Captions (Sharma et al., 2018), consisting of 3.3 million web-harvested images with filtered descriptions; A-OKVQA, (Schwenk et al., 2022) with 24,903 question/answer/rationale triplets requiring broad commonsense knowledge; VQA-V2 (Goyal et al., 2017), offering 265,000 images with open-ended questions that demand understanding of vision, language, and commonsense; Flickr30k (Plummer et al., 2016), with 31,000 images paired with five human-annotated sentences; TextVQA (Singh et al., 2019), containing 45,336 questions on 28,408 images to assess reading and reasoning about text within images; VizWiz (Gurari et al., 2018), a collection of images taken by blind individuals with associated questions; WinoGAViL (Bitton et al., 2022), which tests vision-and-language commonsense reasoning; ImageNet-R (Hendrycks et al., 2021a), featuring artistic renditions of 200 ImageNet classes; ObjectNet (Barbu et al., 2019), a real-world test set for object recognition with random backgrounds, rotations, and viewpoints;

Further enriching the collection are datasets focused on reinforcement learning, and robotics tasks: DM Lab (Beattie et al., 2016), providing frames from the DeepMind Lab environment annotated with agent-object distances; ALE Atari (Bellemare et al., 2013), with 57 Atari 2600 game environments and 500,000 interactions per game; BabyAI (Chevalier-Boisvert et al., 2019), a platform with 19 difficulty levels and 100,000 episodes for grounded language learning ; MuJoCo (Todorov et al., 2012), a benchmark suite of 11 continuous control tasks with 10,000 episodes per environment; DM Control Suite (Tassa et al., 2018), offering standardized continuous control environments; V-D4RL (Lu et al., 2023), a benchmark for continuous control from visual observations of DM Control Suite tasks ; Meta-World (Yu et al., 2021), providing the MT50 benchmark with 50 diverse robot manipulation tasks and 10,000 episodes per environment; Procgen (Cobbe et al., 2020), OpenAI's suite of 16 procedurally generated game-like environments; OpenX-Embodiment (Collaboration et al., 2023), the largest open-source real robot dataset with over 1 mil-

lion trajectories from 22 embodiments; and LocoMuJoCo (Al-Hafez et al., 2023), an imitation learning benchmark for locomotion using real noisy motion capture data.

The consolidated set also incorporates datasets that can help train and evaluate systems on advanced language capabilities: Fineweb-edu (Penedo et al., 2024), providing 1.3 trillion tokens of educational content filtered from the FineWeb dataset; HellaSwag (Zellers et al., 2019), comprising 70,000 multiple-choice questions for commonsense natural language inference; ARC (AI2 Reasoning Challenge)(Clark et al., 2018), providing science exam questions from grades 3-9; CommonsenseQA (Talmor et al., 2019), containing 12,247 multiple-choice questions requiring general world knowledge; and MMLU (Hendrycks et al., 2021b), a benchmark evaluating language models across 57 subjects with approximately 16,000 multiple-choice questions;

## 3. Open-Source Dataset SDK

Our open-source codebase [1] allows for the seamless download of any or all datasets in our collection, along with a toolkit for standardizing robotics and reinforcement learning data. Addressing the common issues of outdated formats, poor maintenance, and accessibility challenges found in many existing control datasets, our toolkit provides stable access methods for diverse RL and robotics datasets, converts control data of various data formats into a unified TensorFlow dataset format, and enables straightforward local storage and usage for training, fine-tuning, and evaluation.

## 4. Evaluation Harness and Metric Suite

To ensure the integrity of pre-training and fine-tuning processes, we introduce systematic test splits for the datasets within our collection as a part of our evaluation harness. These curated splits prevent data contamination and establish a reliable foundation for benchmarking, further enhancing the utility of MultiNet in advancing multimodal AI research.

We also introduce a carefully designed evaluation metrics suite to evaluate the generalization capabilities of SoTA VLMs and VLAs and capture their performance in a fair, quantifiable manner. This suite includes metrics such as Mean Squared Error, Brier Mean Absolute Error (Brier, 1950), Precision, Recall, F1 scores, Invalid output percentage to understand the performance of these models on expert offline trajectories of Reinforcement Learning and Robotics data. For multimodal understanding and generation capabilities, the evaluation suite includes metrics such as CIDEr (Vedantam et al., 2015) to evaluate image captioning and image-based text retrieval, VQA accuracy for visual question answering, Recall@K for image understanding and text-based image retrieval, and Accuracy for commonsense reasoning and text understanding.

By making the test splits and implementation of evaluation metrics Open-Source we hope to enable the community to conduct evaluations of the generalization capabilities of VLMs and VLAs, understand the gaps in the path toward building generalist AI systems, and push the field forward.

## 5. An Open-Source Universal Prompting Framework to Adapt VLMs

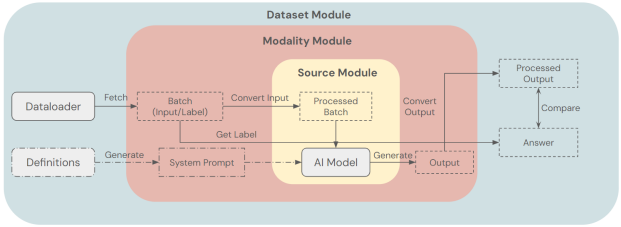


*Figure 2.* GenESIS is a modular framework that efficiently integrates diverse AI models and datasets into benchmark projects without disrupting existing components.

We introduce GenESIS [2] seen in Fig 2, a fully open-source, modular framework designed to streamline the integration of diverse VLMs across a multitude of tasks and datasets. GenESIS enables the efficient addition of new models and datasets to the MultiNet benchmark without disrupting existing components. It primarily helps adapt VLMs to datasets with output formats that differ from their training data. It maximizes scalability and engineering efficiency through the following core principles:

**Interchangeability**: Models and datasets are designed to be mutually supportive and easily swappable, facilitating rapid experimentation and evaluation.

**Abstraction**: Common architectural patterns and logic are shared across modules, simplifying development, understanding, and testing. Base module classes significantly advance this principle by providing a standardized foundation for new components.

**Encapsulation**: Developers can focus on their specific model or dataset without requiring deep knowledge of unrelated modules, promoting modular development and reducing cognitive overhead.

**Prompt Engineering Framework**: A crucial intermediary that translates data of different domains (in this effort,

[1] https://github.com/ManifoldRG/MultiNet

[2] https://github.com/ManifoldRG/MultiNet/tree/main/src/modules

robotics data) into a rich, structured textual representation, priming the VLM for analyzable outputs.

The GenESIS framework comprises the following key elements:

**System-Level Instructions**: Defines the overall goal, inference constraints, and the nature of the interaction (e.g., a simulated environment or robotics embodiment, etc.).

**Task and Environment Context**: Provides explicit descriptions of the specific task, its rules, and relevant environment details.

**Multimodal Input Integration**: Integrates visual observations within the prompt structure, enabling the VLM to contextualize its actions based on visual input.

**Action Space Definition**: Specifies the available actions, their format, and their corresponding verbal descriptions, ensuring clarity in the model's decision-making process.

**Output Instructions**: Defines the precise output structure, enabling the model to handle the complexity of generating outputs of very specific formats.

## 6. Open-Source Architectural and Post Processing Adaptations for SoTA VLA Models

Additionally, as part of MultiNet, we provide a set of SoTA VLA and VLM models that have been adapted to a variety of OOD domains, modalities, environments, and tasks. These adaptations involved processing model inputs and outputs, architectural changes, and inference pipelines to handle the specific structural and statistical characteristics of domains such as offline robotics embodiment datasets and procedurally generated discrete action environments. Our approach emphasizes modular, reusable, and Open-Source adaptations, enabling broader application and benchmarking of VLAs beyond their original training regimes.

**JAT** To adapt HuggingFace's JAT model (Gallouédec et al., 2024) to offline trajectories of robotics embodiment datasets, the inputs to the model were configured by pre-processing input image observations and concatenating floating-point observations into a singular tensor.

**OpenVLA** To adapt OpenVLA to offline trajectories of robotics embodiment datasets, the setup centered on standardizing gripper commands and managing action space compatibility. When adapted to procedurally generated, open-ended discrete action environments, the autoregressive step of the model was restricted to one for predicting single-dimensional action vectors. The generated actions were then unnormalized using statistical data derived from the complete Procgen subdataset. Finally, these unnormalized actions were rounded to obtain discrete action values. For evaluation, logits were extracted from the Llama 2 backbone (Touvron et al., 2023), and probabilities were calculated. These probabilities were subsequently grouped by their associated integer action classes using cached mappings from Llama vocabulary tokens to Procgen integer actions.

**Pi0 Base**: To adapt the Pi0 Base model to procedurally generated discrete action environments, the action horizon was configured to a single timestep. The adaptation involved flow matching denoising over ten steps, utilizing a default action dimension of 32. The first dimension of the action output was selected, unnormalized using Procgen statistics, and then discretized by rounding to the nearest integer.

**Pi0 Fast**: To adapt the Pi0 Fast model to procedurally generated discrete action environments, the action horizon was set to one and autoregressive decoding was limited to four tokens. Probabilities were computed by mapping the backbone Paligemma (Beyer et al., 2024) token IDs to Procgen actions. To mitigate its comparatively slower inference speed (approximately 10 times slower than OpenVLA), embeddings for static zero-images were cached after SigLIP (Zhai et al., 2023) processing, which is the vision encoder model in the backbone. This optimization resulted in a doubling of the inference speed for Pi0 Fast.

## 7. Research Results Using our Open-Source MultiNet Framework

Our MultiNet framework was utilized to evaluate how SoTA VLMs, VLAs, and generalist models generalize to OOD data. The results show consistent failures to generalize to complex robotics and simulated digital action environments. Significant performance gaps highlight the importance of adaptable and platform-agnostic models. Factors such as training data distributions, architectural decisions, and processing techniques strongly influence the outcomes and lead to distinct behaviors in performance. We include some important results from these experiments in A.4 and A.5.

## 8. Conclusion

The MultiNet Benchmark, Framework, Toolkit, and Evaluation Harness are an early step towards the development of next-generation generalist AI systems. MultiNet and its associated tools are openly available, with the intention of empowering the broader research community, enhance reproducibility, and accelerate the collective pursuit of more capable and powerful generalist agents. We detail in future steps in A.1

## Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal

consequences of our work, none which we feel must be specifically highlighted here.

## References


Coyo-700m: Image-text pair dataset. URL https://github.com/kakaobrain/coyo-dataset.

Al-Hafez, F., Zhao, G., Peters, J., and Tateo, D. Locomujoco: A comprehensive imitation learning benchmark for locomotion, 2023. URL https://arxiv.org/abs/2311.02496.

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., and Katz, B. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/97af07a14cacba681feacf3012730892-Paper.pdf.

Beattie, C., Leibo, J. Z., Teplyashin, D., Ward, T., Wainwright, M., Küttler, H., Lefrancq, A., Green, S., Valdés, V., Sadik, A., Schrittwieser, J., Anderson, K., York, S., Cant, M., Cain, A., Bolton, A., Gaffney, S., King, H., Hassabis, D., Legg, S., and Petersen, S. Deepmind lab, 2016. URL https://arxiv.org/abs/1612.03801.

Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, June 2013. ISSN 1076-9757. doi: 10.1613/jair.3912. URL http://dx.doi.org/10.1613/jair.3912.

Beyer, L., Steiner, A., Pinto, A. S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., Unterthiner, T., Keysers, D., Koppula, S., Liu, F., Grycner, A., Gritsenko, A., Houlsby, N., Kumar, M., Rong, K., Eisenschlos, J., Kabra, R., Bauer, M., Bošnjak, M., Chen, X., Minderer, M., Voigtlaender, P., Bica, I., Balazevic, I., Puigcerver, J., Papalampidi, P., Henaff, O., Xiong, X., Soricut, R., Harmsen, J., and Zhai, X. Paligemma: A versatile 3b vlm for transfer, 2024. URL https://arxiv.org/abs/2407.07726.

Bitton, Y., Guetta, N. B., Yosef, R., Elovici, Y., Bansal, M., Stanovsky, G., and Schwartz, R. Winogavil: Gamified association benchmark to challenge vision-and-language models, 2022. URL https://arxiv.org/abs/2207.12576.

Black, K. et al. pi0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. URL https://arxiv.org/abs/2410.24164.

Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3, 1950. doi: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.

Chevalier-Boisvert, M., Bahdanau, D., Lahlou, S., Willems, L., Saharia, C., Nguyen, T. H., and Bengio, Y. Babyai: A platform to study the sample efficiency of grounded language learning, 2019. URL https://arxiv.org/abs/1810.08272.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Cobbe, K., Hesse, C., Hilton, J., and Schulman, J. Leveraging procedural generation to benchmark reinforcement learning, 2020. URL https://arxiv.org/abs/1912.01588.

Collaboration, O. X.-E., O'Neill, A., Rehman, A., Gupta, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandlekar, A., Jain, A., Tung, A., Bewley, A., Herzog, A., Irpan, A., Khazatsky, A., Rai, A., Gupta, A., Wang, A., Kolobov, A., Singh, A., Garg, A., Kembhavi, A., Xie, A., Brohan, A., Raffin, A., Sharma, A., Yavary, A., Jain, A., Balakrishna, A., Wahid, A., Burgess-Limerick, B., Kim, B., Schölkopf, B., Wulfe, B., Ichter, B., Lu, C., Xu, C., Le, C., Finn, C., Wang, C., Xu, C., Chi, C., Huang, C., Chan, C., Agia, C., Pan, C., Fu, C., Devin, C., Xu, D., Morton, D., Driess, D., Chen, D., Pathak, D., Shah, D., Büchler, D., Jayaraman, D., Kalashnikov, D., Sadigh, D., Johns, E., Foster, E., Liu, F., Ceola, F., Xia, F., Zhao, F., Frujeri, F. V., Stulp, F., Zhou, G., Sukhatme, G. S., Salhotra, G., Yan, G., Feng, G., Schiavi, G., Berseth, G., Kahn, G., Yang, G., Wang, G., Su, H., Fang, H.-S., Shi, H., Bao, H., Amor, H. B., Christensen, H. I., Furuta, H., Bharadhwaj, H., Walke, H., Fang, H., Ha, H., Mordatch, I., Radosavovic, I., Leal, I., Liang, J., Abou-Chakra, J., Kim, J., Drake, J., Peters, J., Schneider, J., Hsu, J., Vakil, J., Bohg, J., Bingham, J., Wu, J., Gao, J., Hu, J., Wu, J., Wu, J., Sun, J., Luo, J., Gu, J., Tan, J., Oh, J., Wu, J., Lu, J., Yang, J., Malik, J., Silvério, J., Hejna, J., Booher, J., Tompson, J., Yang, J., Salvador, J., Lim, J. J., Han, J., Wang, K., Rao, K., Pertsch, K., Hausman, K., Go, K., Gopalakrishnan, K., Goldberg, K., Byrne, K., Oslund, K., Kawaharazuka, K., Black, K., Lin, K., Zhang, K., Ehsani, K., Lekkala, K., Ellis, K., Rana, K., Srinivasan, K., Fang, K., Singh, K. P.,

Zeng, K.-H., Hatch, K., Hsu, K., Itti, L., Chen, L. Y., Pinto, L., Fei-Fei, L., Tan, L., Fan, L. J., Ott, L., Lee, L., Weihs, L., Chen, M., Lepert, M., Memmel, M., Tomizuka, M., Itkina, M., Castro, M. G., Spero, M., Du, M., Ahn, M., Yip, M. C., Zhang, M., Ding, M., Heo, M., Srirama, M. K., Sharma, M., Kim, M. J., Kanazawa, N., Hansen, N., Heess, N., Joshi, N. J., Suenderhauf, N., Liu, N., Palo, N. D., Shafiullah, N. M. M., Mees, O., Kroemer, O., Bastani, O., Sanketi, P. R., Miller, P. T., Yin, P., Wohlhart, P., Xu, P., Fagan, P. D., Mitrano, P., Sermanet, P., Abbeel, P., Sundaresan, P., Chen, Q., Vuong, Q., Rafailov, R., Tian, R., Doshi, R., Mart'in-Mart'in, R., Baijal, R., Scalise, R., Hendrix, R., Lin, R., Qian, R., Zhang, R., Mendonca, R., Shah, R., Hoque, R., Julian, R., Bustamante, S., Kirmani, S., Levine, S., Lin, S., Moore, S., Bahl, S., Dass, S., Sonawani, S., Tulsiani, S., Song, S., Xu, S., Haldar, S., Karamcheti, S., Adebola, S., Guist, S., Nasiriany, S., Schaal, S., Welker, S., Tian, S., Ramamoorthy, S., Dasari, S., Belkhale, S., Park, S., Nair, S., Mirchandani, S., Osa, T., Gupta, T., Harada, T., Matsushima, T., Xiao, T., Kollar, T., Yu, T., Ding, T., Davchev, T., Zhao, T. Z., Armstrong, T., Darrell, T., Chung, T., Jain, V., Kumar, V., Vanhoucke, V., Zhan, W., Zhou, W., Burgard, W., Chen, X., Chen, X., Wang, X., Zhu, X., Geng, X., Liu, X., Liangwei, X., Li, X., Pang, Y., Lu, Y., Ma, Y. J., Kim, Y., Chebotar, Y., Zhou, Y., Zhu, Y., Wu, Y., Xu, Y., Wang, Y., Bisk, Y., Dou, Y., Cho, Y., Lee, Y., Cui, Y., Cao, Y., Wu, Y.-H., Tang, Y., Zhu, Y., Zhang, Y., Jiang, Y., Li, Y., Li, Y., Iwasawa, Y., Matsuo, Y., Ma, Z., Xu, Z., Cui, Z. J., Zhang, Z., Fu, Z., and Lin, Z. Open X-Embodiment: Robotic learning datasets and RT-X models. https://arxiv.org/abs/2310.08864, 2023.

Fu, J., Kumar, A., Nachum, O., Tucker, G., and Levine, S. D4rl: Datasets for deep data-driven reinforcement learning, 2021. URL https://arxiv.org/abs/2004.07219.

Gadre, S. Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., Orgad, E., Entezari, R., Daras, G., Pratt, S., Ramanujan, V., Bitton, Y., Marathe, K., Mussmann, S., Vencu, R., Cherti, M., Krishna, R., Koh, P. W., Saukh, O., Ratner, A., Song, S., Hajishirzi, H., Farhadi, A., Beaumont, R., Oh, S., Dimakis, A., Jitsev, J., Carmon, Y., Shankar, V., and Schmidt, L. Datacomp: In search of the next generation of multimodal datasets, 2023. URL https://arxiv.org/abs/2304.14108.

Gallouédec, Q., Beeching, E., Romac, C., and Dellandréa, E. Jack of all trades, master of some, a multi-purpose transformer agent, 2024. URL https://arxiv.org/abs/2402.09844.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017. URL https://arxiv.org/abs/1612.00837.

Gulcehre, C., Wang, Z., Novikov, A., Paine, T. L., Colmenarejo, S. G., Zolna, K., Agarwal, R., Merel, J., Mankowitz, D., Paduraru, C., Dulac-Arnold, G., Li, J., Norouzi, M., Hoffman, M., Nachum, O., Tucker, G., Heess, N., and de Freitas, N. Rl unplugged: A suite of benchmarks for offline reinforcement learning, 2021. URL https://arxiv.org/abs/2006.13888.

Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J., and Bigham, J. P. Vizwiz grand challenge: Answering visual questions from blind people, 2018. URL https://arxiv.org/abs/1802.08218.

Guruprasad, P., Sikka, H., Song, J., Wang, Y., and Liang, P. P. Benchmarking vision, language, & action models on robotic learning tasks, 2024. URL https://arxiv.org/abs/2411.05821.

Guruprasad, P., Wang, Y., Chowdhury, S., and Sikka, H. Benchmarking vision, language, & action models in procedurally generated, open ended action environments, 2025. URL https://arxiv.org/abs/2505.05540.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. The many faces of robustness: A critical analysis of out-of-distribution generalization, 2021a. URL https://arxiv.org/abs/2006.16241.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding, 2021b. URL https://arxiv.org/abs/2009.03300.

Ichter, B. et al. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025. URL https://arxiv.org/abs/2501.09747.

Kim, M. J. et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A. M., Kiela, D., Cord, M., and Sanh, V. Obelics: An open web-scale filtered dataset of interleaved image-text documents, 2023. URL https://arxiv.org/abs/2306.16527.

Liang, P. P., Lyu, Y., Fan, X., Wu, Z., Cheng, Y., Wu, J., Chen, L., Wu, P., Lee, M. A., Zhu, Y., Salakhutdinov,

R., and Morency, L.-P. Multibench: Multiscale benchmarks for multimodal representation learning, 2021. URL https://arxiv.org/abs/2107.07502.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Lu, C., Ball, P. J., Rudner, T. G. J., Parker-Holder, J., Osborne, M. A., and Teh, Y. W. Challenges and opportunities in offline reinforcement learning from visual observations, 2023. URL https://arxiv.org/abs/2206.04779.

Penedo, G., Kydlíček, H., allal, L. B., Lozhkov, A., Mitchell, M., Raffel, C., Werra, L. V., and Wolf, T. The fineweb datasets: Decanting the web for the finest text data at scale, 2024. URL https://arxiv.org/abs/2406.17557.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, 2016. URL https://arxiv.org/abs/1505.04870.

Schwenk, D., Khandelwal, A., Clark, C., Marino, K., and Mottaghi, R. A-okvqa: A benchmark for visual question answering using world knowledge, 2022. URL https://arxiv.org/abs/2206.01718.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Gurevych, I. and Miyao, Y. (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238. URL https://aclanthology.org/P18-1238.

Singh, A., Natarjan, V., Shah, M., Jiang, Y., Chen, X., Parikh, D., and Rohrbach, M. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326, 2019.

Talmor, A., Herzig, J., Lourie, N., and Berant, J. Commonsenseqa: A question answering challenge targeting commonsense knowledge, 2019. URL https://arxiv.org/abs/1811.00937.

Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., de Las Casas, D., Budden, D., Abdolmaleki, A., Merel, J., Lefrancq, A., Lillicrap, T., and Riedmiller, M. Deepmind control suite, 2018. URL https://arxiv.org/abs/1801.00690.

Todorov, E., Erez, T., and Tassa, Y. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012. doi: 10.1109/IROS.2012.6386109.

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., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Vedantam, R., Zitnick, C. L., and Parikh, D. Cider: Consensus-based image description evaluation, 2015. URL https://arxiv.org/abs/1411.5726.

White, C., Dooley, S., Roberts, M., Pal, A., Feuer, B., Jain, S., Shwartz-Ziv, R., Jain, N., Saifullah, K., Naidu, S., Hegde, C., LeCun, Y., Goldstein, T., Neiswanger, W., and Goldblum, M. Livebench: A challenging, contamination-free llm benchmark, 2024. URL https://arxiv.org/abs/2406.19314.

Yu, T., Quillen, D., He, Z., Julian, R., Narayan, A., Shively, H., Bellathur, A., Hausman, K., Finn, C., and Levine, S. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning, 2021. URL https://arxiv.org/abs/1910.10897.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. Hellaswag: Can a machine really finish your sentence?, 2019. URL https://arxiv.org/abs/1905.07830.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training, 2023. URL https://arxiv.org/abs/2303.15343.

# A. Appendix

## A.1. Future directions for MultiNet

- **Comprehensive Modality Analysis**: We will systematically investigate the interplay between control-task training and the emergent vision-language capabilities within Vision-Language Agents (VLAs). This exploration is critical for understanding architectural trade-offs and guiding the design of truly versatile, multi-modal generalist models.
- **Expanded Evaluation Horizons**: To rigorously assess and push the boundaries of generalization, MultiNet will integrate a more diverse array of control tasks, extending beyond current benchmarks like OpenX-Embodiment and Procgen. This expansion will allow for a thorough examination of model performance in entirely novel environments, pinpointing architectural bottlenecks and the frontiers of current generalization capabilities.
- **Advanced Transfer Learning Paradigms**: Our research will move beyond zero-shot evaluations to explore the efficacy of few-shot learning and fine-tuning strategies. A key focus will be on transferring learned skills to disparate domains, including complex software environments, thereby deepening our understanding of how to cultivate truly transferable representations across both embodied and digital tasks.
- **An Evolving Open-Source Simulation Benchmark**: We are actively working to transform MultiNet from its current offline evaluation framework into a dynamic, interactive, and open-source online benchmark. This will be powered by sophisticated world models underpinning state-of-the-art 2D and 3D simulation environments, enabling real-time assessment of agent capabilities in responsive and evolving settings. This open benchmark will serve as a community resource for robust and continuous evaluation.
- **Cultivating Multi-Domain Expertise**: A crucial research thrust involves developing advanced cross-domain adaptation mechanisms. The goal is to enable AI agents to seamlessly transfer and apply knowledge across a wide spectrum of environmental contexts, from robotic manipulation and navigation to software interaction and gameplay in diverse virtual worlds.

## A.2. Dataset Spec

*Table 1.* Overview of Datasets

| DATASET | DESCRIPTION / TASK TYPE | CATEGORY |
|---|---|---|
| OBELICS | INTERLEAVED IMAGE-TEXT | VISION-LANGUAGE |
| COYO-700M | IMAGE-TEXT PAIRS | VISION-LANGUAGE |
| MS COCO | OBJECT DETECTION, SEGMENTATION, KEY-POINT DETECTION, CAPTIONING | VISION-LANGUAGE |
| CONCEPTUAL CAPTIONS | IMAGE CAPTIONING | VISION-LANGUAGE |
| A-OKVQA | VISUAL QUESTION ANSWERING | VISION-LANGUAGE |
| VQA-V2 | VISUAL QUESTION ANSWERING | VISION-LANGUAGE |
| DATACOMP-1B | IMAGE-TEXT PAIRS | VISION-LANGUAGE |
| FLICKR30K | IMAGE CAPTIONING | VISION-LANGUAGE |
| TEXTVQA | VISUAL QUESTION ANSWERING | VISION-LANGUAGE |
| VIZWIZ | VISUAL QUESTION ANSWERING | VISION-LANGUAGE |
| WINOGAVIL | VISION-BASED COMMONSENSE REASONING | VISION-LANGUAGE |
| IMAGENET-R | IMAGE-TEXT PAIRS | VISION-LANGUAGE |
| OBJECTNET | IMAGE-TEXT PAIRS | VISION-LANGUAGE |
| FINEWEB-EDU | HIGH QUALITY TEXT CORPUS | LANGUAGE |
| HELLASWAG | COMMONSENSE REASONING | LANGUAGE |
| ARC | COMPLEX REASONING AND KNOWLEDGE APPLICATION | LANGUAGE |
| COMMONSENSEQA | COMMONSENSE REASONING | LANGUAGE |
| MMLU | KNOWLEDGE-INTENSIVE QUESTION ANSWERING | LANGUAGE |
| DM LAB | TEACH RL AGENTS 3D VISION (NAVIGATION-BASED CONTROL) | CONTROL |
| DM CONTROL SUITE | PHYSICS-BASED SIMULATION ENVIRONMENTS (LOCOMOTION-BASED CONTROL) | CONTROL |
| ALE ATARI | ATARI GAMES (GAME-BASED CONTROL) | CONTROL |
| BABY AI | LANGUAGE-GROUNDED NAVIGATION (NAVIGATION-BASED CONTROL) | CONTROL |
| MUJOCO | MULTI-JOINT DYNAMICS (LOCOMOTION-BASED CONTROL) | CONTROL |
| META-WORLD | META-RL AND MULTI-TASK LEARNING (MANIPULATION-BASED CONTROL) | CONTROL |
| V-D4RL | PIXEL-BASED ANALOGUES OF DM CONTROL SUITE (LOCOMOTION-BASED CONTROL) | CONTROL |
| PROCGEN | PROCEDURALLY GENERATED ATARI-LIKE ENVIRONMENTS (GAME-BASED CONTROL) | CONTROL |
| OPEN X EMBODIMENT | REAL-WORLD ROBOTICS TASKS (MANIPULATION & LOCOMOTION CONTROL) | CONTROL |
| LOCOMUJOCO | IMITATION LEARNING FOR LOCOMOTION (LOCOMOTION-BASED CONTROL) | CONTROL |

## A.3. MultiNet Benchmark Metrics

*Table 2.* MultiNet Benchmark metrics & categories

| METRIC | EVALUATION CATEGORY |
|---|---|
| MEAN SQUARED ERROR | RL, ROBOTICS |
| BRIER MEAN ABSOLUTE ERROR | RL, ROBOTICS |
| PRECISION (MICRO, MACRO, AND CLASS-WISE VARIANTS) | RL, ROBOTICS |
| RECALL (MICRO, MACRO, AND CLASS-WISE VARIANTS) | RL, ROBOTICS |
| F1 SCORE (MICRO, MACRO, AND CLASS-WISE VARIANTS) | RL, ROBOTICS |
| INVALIDS PERCENTAGE | RL, ROBOTICS |
| CIDER | IMAGE CAPTIONING, IMAGE-BASED TEXT RETRIEVAL |
| VQA ACCURACY | VISUAL QUESTION ANSWERING |
| RECALL@K | IMAGE UNDERSTANDING, TEXT-BASED IMAGE RETRIEVAL |
| ACCURACY | VQA, COMMONSENSE REASONING, TEXT UNDERSTANDING |

## A.4. Model Performance on Out-of-Distribution Robotic Environments

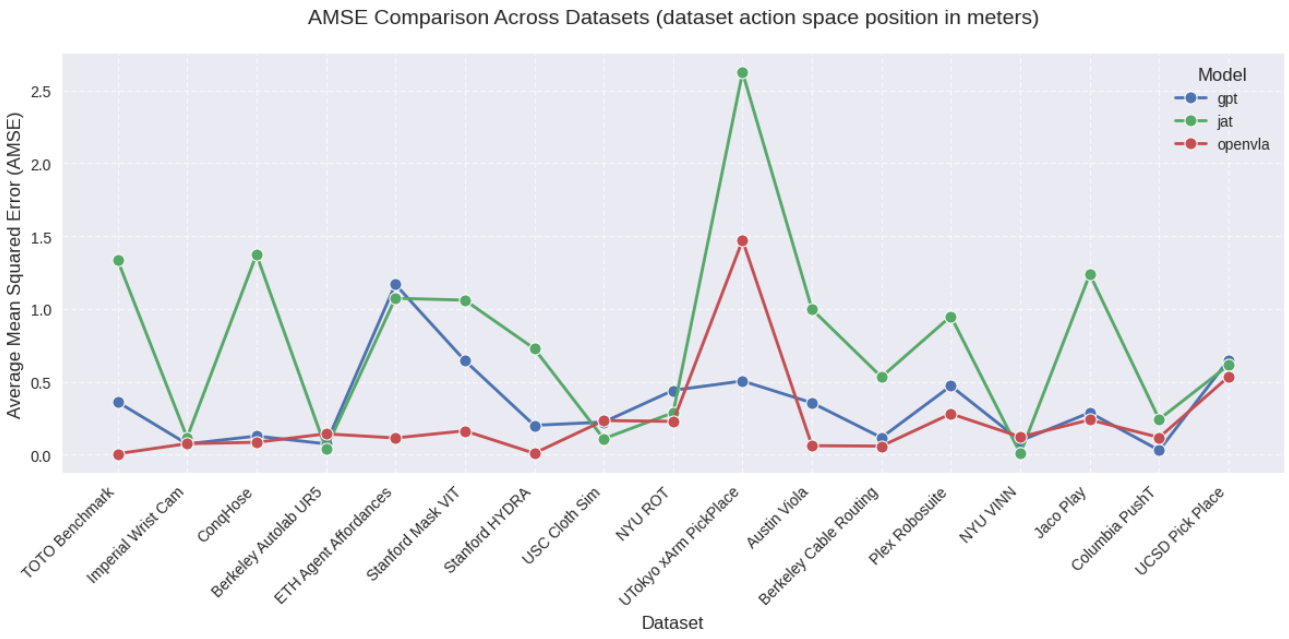


*Figure 3.* AMSE values of GPT-4o, JAT, and OpenVLA Across 20 OpenX Datasets. JAT displays the poorest performance out of the 3 models with higher AMSE scores, while OpenVLA and GPT-4o demonstrate similar performance. OpenVLA displays consistent performance across most datasets.

## A.5. Model Performance on Out-of-Distribution Procedurally Generated Environments

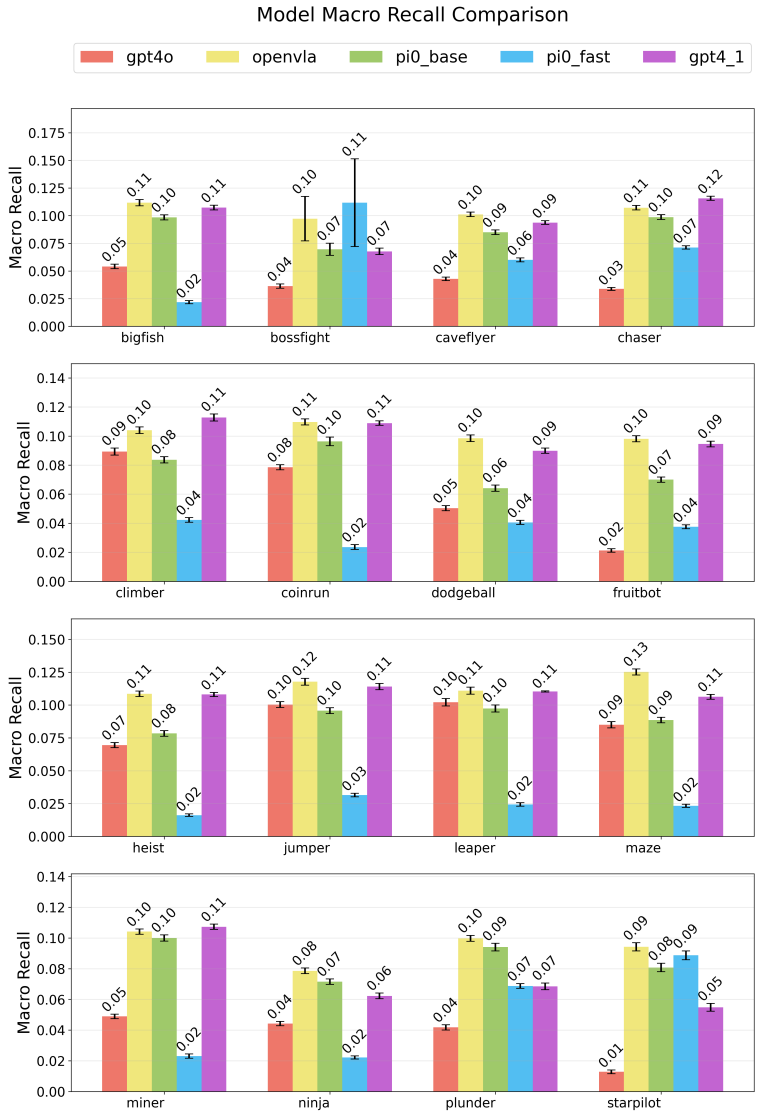


*Figure 4.* Macro recall across all 5 models. OpenVLA performs better when considering macro recall compared to macro precision, indicating a high number of false positives. GPT 4o shows lower macro recall than precision, indicating biased performance towards specific minority classes. GPT 4.1 and Pi0 Base show relatively less biased and moderate performance, whereas Pi0 FAST showed consistent low recall.

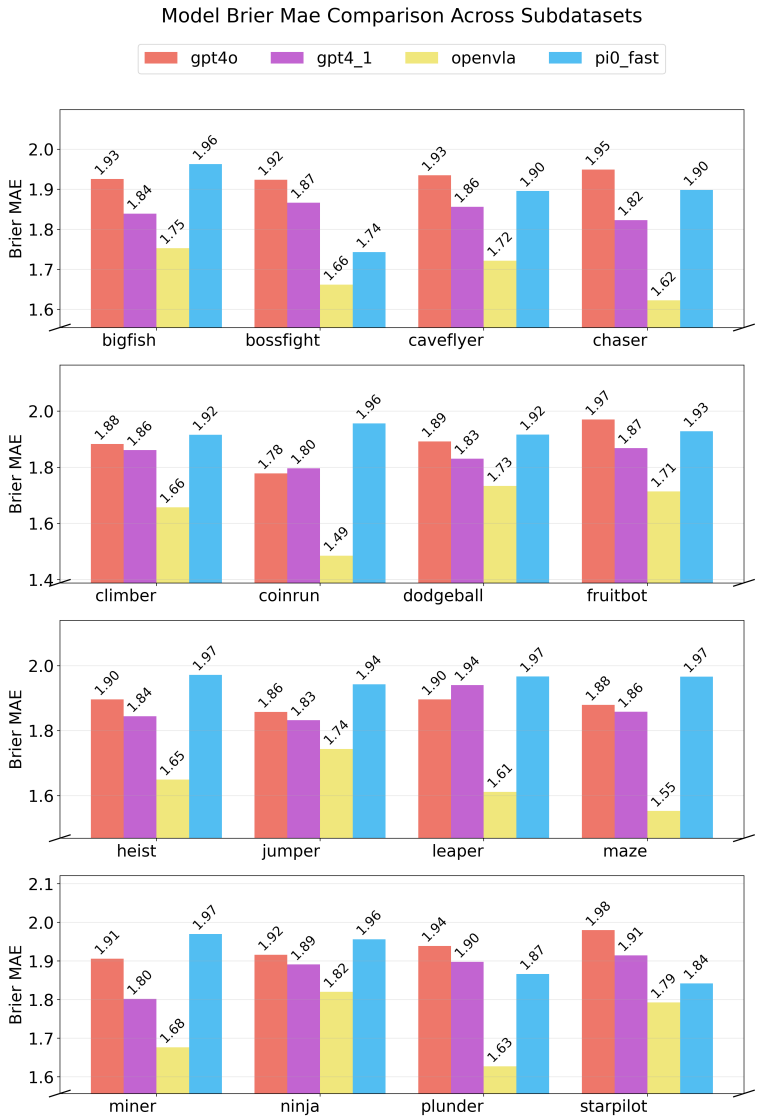


*Figure 5.* Brier Mean Absolute Error scores across 4 models - GPT 4o, OpenVLA, GPT 4.1, and Pi0 FAST. Pi0 Base is a diffusion-based model and can not be evaluated using Brier MAE due to a lack of logits in its inference architecture. All models display Brier MAE close to 2, indicating poor performance.

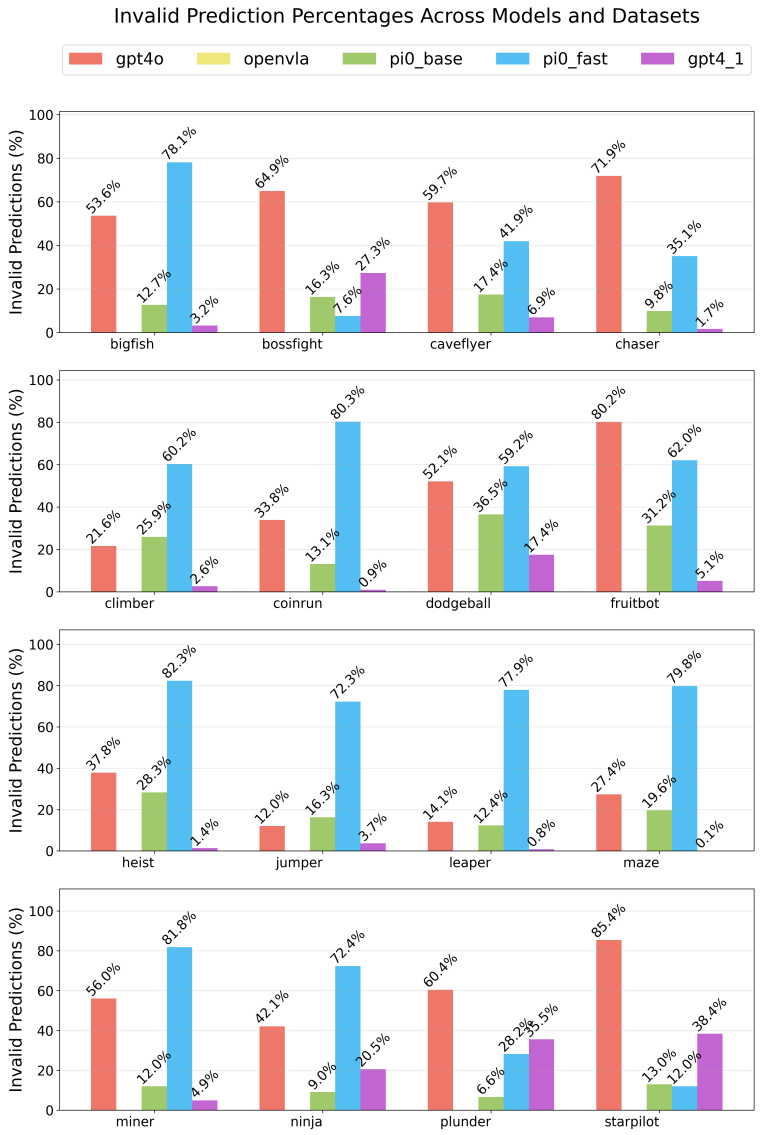


*Figure 6.* Percentage invalids across all 5 models. Invalids refer to model predictions that are not valid actions in the subdataset's action space. Pi0 FAST and GPT 4o struggle to produce valid actions irrespective of the subdataset.