MULTINET: A BENCHMARK FOR GENERALIST ACTION MODELS

Pranav Guruprasad
Metarch.ai
Manifold Research
pranav@metarch.ai

Harshvardhan Sikka
Georgia Tech
Metarch.ai
Manifold Research
harsh@metarch.ai

ABSTRACT

Recent advances in machine learning have demonstrated the potential of large-scale models to exhibit broad generalization capabilities across diverse tasks. However, the evaluation of these models remains fragmented, with different benchmarks focusing on specific modalities or capabilities. We present Multinet, a comprehensive benchmark designed to evaluate truly generalist models across vision, language, and action domains. Multinet consolidates diverse, high-quality datasets including OBELICS, COYO-700M, and OpenX-Embodiment, establishing standardized evaluation protocols for assessing both the action capabilities of Vision-Language Models (VLMs) and the multimodal understanding of Vision-Language-Action Models (VLAs). Our benchmark includes carefully curated training data spanning vision-language association (800M+ image-text pairs), language understanding (1.3T tokens), and control tasks (35+ TB of robotics and RL data). We provide evaluation protocols across multiple dimensions, including image captioning, visual question answering, commonsense reasoning, and robotic control. Additionally, we open-source a toolkit that standardizes the challenging process of obtaining and utilizing reinforcement learning and robotics data from various sources. Through systematic evaluation of state-of-the-art models, we aim to demonstrate significant gaps in current approaches: VLMs struggle with control tasks while VLAs show limited capabilities in pure vision-language understanding. These findings will highlight the need for more genuinely generalist models. Multinet serves as both a comprehensive evaluation framework and a foundation for developing the next generation of truly generalist AI systems.

1 Introduction

Motivations

Recent advances in machine learning have demonstrated the potential of large-scale models to exhibit broad generalization capabilities across diverse tasks. A particularly promising direction emerged with DeepMind's Gato [42], which provided the first glimpse of a truly generalist model capable of performing hundreds of different tasks across multiple domains. However, the datasets and model used to train Gato remain closed-source, limiting the research community's ability to build upon and extend this work.

Concurrent developments in Vision-Language-Action (VLA) models have shown impressive capabilities in grounding real-world actions with vision and natural langueg [25, 51]. These models can interpret visual scenes, understand language commands, and generate appropriate control actions. However, current VLA models are primarily focused on narrow domains like robotic manipulation, and are not proven to perform well on complex vision-language understanding tasks. This limitation could stem partly from their training data, which typically emphasizes a specific subset of capabilities rather than true generalist behavior.

Building genuinely generalist models requires training on diverse datasets that span multiple modalities and task types. Such models must excel not only at individual modalities (vision, language, or control) but also at tasks that require seamless integration across modalities - a requirement that better reflects real-world scenarios. Currently, there exists no large-scale, open-source dataset specifically designed for training and evaluating such generalist models. This gap motivated the development of Multinet.

Contributions

In this paper, we present Multinet, a comprehensive benchmark for developing and evaluating generalist action models. Our contributions include:

- The largest open-source generalist dataset, consolidating diverse modalities and tasks suitable for pre-training, fine-tuning, and evaluation
- A novel benchmark for assessing state-of-the-art Vision-Language Models (VLMs) and Vision-Language-Action Models (VLAs)
- A detailed analysis of the constituent datasets' validity and utility for generalist objectives
- An open-source software toolkit that facilitates dataset access and standardizes control data from various sources into a common TensorFlow format

Through Multinet, we aim to accelerate research in generalist models by providing the community with the necessary tools and benchmarks to develop and evaluate truly general-purpose AI systems. Our benchmark enables systematic comparison of different approaches and provides insights into the challenges and opportunities in building models that can seamlessly operate across multiple modalities and tasks.

2 Related Work

Recent advances in machine learning have produced several promising directions toward generalist AI systems. We organize our discussion of related work into three main categories: generalist agents, vision-language-action models, and existing benchmarks.

2.1 Generalist Agents

DeepMind's Gato [42] represented a significant milestone as the first truly multi-modal, multi-task, multi-embodiment agent. Operating with a single neural network and fixed weights, Gato demonstrated capabilities spanning Atari gameplay, image captioning, conversational interaction, and real-world robotic manipulation. The agent dynamically selects appropriate output modalities (text, joint torques, button presses, or other tokens) based on context. However, both the model and the majority of its training datasets remain closed-source, limiting its impact on the broader research community.

2.2 Vision-Language-Action Models

Recent work has produced several notable vision-language-action (VLA) models, including Octo [51] and OpenVLA [25]. These models demonstrate impressive capabilities in grounding language instructions in robotic control actions through either pre-training from scratch or fine-tuning existing models. However, their training focuses exclusively on vision-language-grounded control tasks, neglecting pure vision-language understanding or generation tasks. This specialization limits their ability to serve as truly generalist models capable of operating across the full spectrum of modalities.

2.3 Benchmarks and Evaluation Frameworks

Existing benchmarks in the field can be categorized into control-focused, vision-language focused, and robotics-specific evaluations.

Control Benchmarks: RL Unplugged [18] provides a comprehensive suite of offline reinforcement learning benchmarks, spanning domains from Atari games to DM Control Suite tasks. It standardizes environments, datasets, and evaluation protocols to enhance reproducibility in offline RL research. Similarly, D4RL [14] offers over 40 standardized environments and datasets for offline RL, covering robotic manipulation, navigation, and autonomous driving tasks. The robotics community has developed numerous specialized benchmarks, particularly focusing on imitation learning and behavior cloning. THE COLOSSEUM[40] provides a systematic evaluation framework for robotic manipulation across 14 environmental perturbations, while FactorWorld[56] and KitchenShift[57] examine generalization across various environmental factors. Several task-specific benchmarks have emerged: RLBench[24] offers 100 simulated manipulation tasks, RAVENS[23] focuses on vision-based manipulation, and FurnitureBench[22] addresses long-horizon complex manipulation in real-world settings. Recent additions include LIBERO[32] for lifelong robot learning, FMB[35] for generalizable manipulation, DUDE[53] for document manipulation, and ProcTHOR[11] for procedurally generated embodied AI tasks.

Vision-Language Benchmarks: The evolution of multimodal evaluation has progressed significantly from single-task benchmarks like VQA[17], OK-VQA[36], and MSCOCO[31] to more comprehensive frameworks. MultiBench [30] presents a unified evaluation framework spanning 15 datasets, 10 modalities, and 20 prediction tasks across 6 research areas. MMLU [21] evaluates language model capabilities across 57 academic subjects, while MMMU[63] extends this to college-level multimodal understanding across technical disciplines. Recent developments include specialized benchmarks like MathVista[34] for mathematical reasoning and GAIA[37] for fundamental reasoning abilities, as well as more holistic evaluations through LAMM[59], LVLM-eHub[58], SEED[29], and MM-Vet[62]. LiveBench [55] specifically addresses test set contamination through continuously updated evaluation data.

While these benchmarks excel in their respective domains, they remain specialized to particular modalities or task types. Multinet addresses this limitation by providing a diverse collection of datasets specifically designed to evaluate and advance truly generalist models. Our benchmark enables the development of systems with strong capabilities across vision-language association, language understanding and generation, and reward-based action trajectories in varied environments. Additionally, Multinet serves as a comprehensive evaluation framework for current VLMs and VLAs, highlighting areas for improvement in the development of next-generation generalist models.

3 Coverage

3.1 Datasets

3.1.1 For Training

Vision-Language and Language

OBELICS OBELICS [27] is an open web-scale filtered dataset of interleaved image-text documents extracted from Common Crawl [1]. It comprises 141 million web pages, 353 million associated images, and 115 billion text tokens. The interleaved nature of the documents provides richer context compared to simple image-text pairs. The dataset occupies 666 GB in arrow format and 377 GB in Parquet format, and currently includes only training data.

COYO-700M COYO-700M [2] contains 747 million pairs of alt-text and associated images harvested from HTML documents. The dataset was curated from approximately 10 billion initial pairs collected from CommonCrawl between October 2020 and August 2021. It employs minimal filtering, resulting in a "noisier" dataset that can potentially improve model robustness while providing challenging evaluation scenarios.

MS-COCO Captions MS-COCO Captions [31] is a large-scale object detection, segmentation, and captioning dataset containing 330,000 images with 5 captions each, totaling 1.5 million object instances. The dataset's high-quality annotations make it particularly valuable for training and evaluating vision understanding tasks.

Conceptual Captions Conceptual Captions [47] consists of 3.3 million web-harvested images with filtered descriptions derived from HTML alt-text attributes. The dataset includes 3,318,333 image-URL/caption pairs for training and 15,840 pairs for validation. Unlike COYO-700M, it undergoes more rigorous filtering to ensure higher data quality.

A-OKVQA A-OKVQA [46], the successor to OKVQA [36], contains 24,903 question/answer/rationale triplets requiring broad commonsense and world knowledge. The dataset is split into 17.1K/1.1K/6.7K for train, validation, and test. Questions cannot be answered by simply querying a knowledge base, making it particularly valuable for evaluating sophisticated reasoning capabilities.

VQA-V2 VQA-V2 [17] provides open-ended questions about images that require understanding of vision, language, and commonsense knowledge. The dataset includes 265,000 images with at least three questions per image, ten ground truth answers per question, and three plausible answers per question.

DataComp-1B DataComp-1B [15] consists of 1.4 billion image-text pairs curated from an initial pool of 12.8 billion pairs. Despite being smaller than alternatives like LAION-2B [44], it achieves better performance with fewer resources, demonstrating 79.2% zero-shot accuracy on ImageNet [12] while using 9x less compute.

Fineweb-edu Fineweb-edu [38] contains 1.3T tokens of educational content filtered from the FineWeb dataset using an educational quality classifier trained on LLama3-70B-Instruct [13] annotations. The dataset demonstrates superior performance on standard benchmarks compared to larger, unfiltered alternatives.

Reinforcement Learning and Robotics

DM Lab DM Lab [5] provides frames from the DeepMind Lab environment annotated with agent-object distances. The dataset contains 360x480 color images across 6 classes (combinations of close, far, very far and positive reward, negative reward). This 1.8 TB dataset is particularly valuable for training models to reason about spatial relationships and depth perception in 3D environments, making it crucial for robotics and augmented reality applications.

ALE Atari The Arcade Learning Environment (ALE) provides 57 Atari 2600 game environments for AI agent development. The dataset version included in Multinet contains 500,000 interactions per game generated by JAT [16], where dedicated agents were trained for 2 billion steps using asynchronous PPO [45]. The complete dataset occupies 66 GB.

BabyAI BabyAI [7] is a research platform comprising 19 levels of increasing difficulty, designed to investigate grounded language learning with humans in the loop. The platform teaches agents a combinatorially rich subset of English and includes a hand-crafted bot that simulates a human teacher. Our version contains 100,000 episodes across 39 available settings collected by JAT [16], totaling 148 GB.

MuJoCo The MuJoCo [52] benchmark suite contains 11 continuous control tasks of varying complexity. The dataset version included in Multinet includes 10,000 episodes per environment collected by JAT[16], generated by agents trained using asynchronous PPO [45] from Sample Factory [39]. These agents achieved scores meeting or exceeding current standards. The complete dataset occupies 33 GB.

DM Control Suite The DeepMind Control Suite [50] provides standardized continuous control environments powered by the MuJoCo physics engine. It includes diverse tasks ranging from simple (Pendulum, Cart-pole) to complex (Humanoid, Manipulator), with interpretable rewards. The dataset occupies 52 GB and is particularly valuable for training on simulated motor-control problems.

V-D4RL V-D4RL [33] is the first publicly available benchmark for continuous control from visual observations of DMControl Suite tasks featuring diverse behavioral policies. The 62 GB dataset specifically tests robustness to distractions, generalization across dynamics, and offline reinforcement learning at scale.

Meta-World Meta-World's MT50 benchmark [61] provides 50 diverse robot manipulation tasks. The dataset version included in Multinet includes a 15 GB dataset collected by JAT[16], and contains 10,000 episodes per environment (limited to 100 timesteps) generated by task-specific trained agents.

Procgen Procgen [9] is OpenAI's suite of 16 procedurally generated game-like environments designed to benchmark efficiency and generalization in RL. With diversity comparable to ALE, these environments require robust policies that avoid overfitting to narrow state spaces. The dataset occupies 739 GB.

OpenX-Embodiment OpenX-Embodiment [10] is currently the largest open-source real robot dataset, containing over 1M trajectories from 22 robot embodiments. For Multinet v0.1, we utilize 53 of the 72 available datasets, stored in the RLDS [41] format which accommodates various action spaces and input modalities. The training splits of these 53 datasets total 32 TB. Dataset selection involved careful curation based on robot morphology, gripper specifications, action spaces, and sensor configurations.

LocoMuJoCo LocoMuJoCo [3] is an open-source imitation learning benchmark focused on locomotion. It includes diverse environments (quadrupeds, bipeds, and musculoskeletal human models) with comprehensive datasets including real noisy motion capture data and ground truth expert data. Our version uses the "perfect dataset" containing ground truth states and actions from expert policies, occupying 690 MB.

3.1.2 For Evaluation

Beyond the validation and test splits available in our training datasets, we include several datasets specifically for evaluation purposes:

Flickr30k Flickr30k [60] contains 31,000 images from Flickr, each paired with five human-annotated descriptive sentences. The dataset includes rich annotations like coreference chains and bounding boxes, making it particularly valuable for evaluating sentence-based image description and grounded language understanding.

TextVQA TextVQA [48] evaluates models' ability to read and reason about text within images. It contains 45,336 questions across 28,408 images from OpenImages [26], specifically testing models' capability to incorporate text as a modality for question answering.

VizWiz VizWiz [19] provides a collection of images taken by blind individuals along with associated questions, presenting unique challenges due to poor image quality and conversational question formats. This dataset tests model robustness in real-world assistive technology scenarios where questions may not always be answerable.

WinoGAVil WinoGAViL [6] tests vision-and-language commonsense reasoning abilities by requiring models to identify associations beyond simple visual recognition. It provides both zero-shot and supervised evaluation settings, challenging models with tasks requiring general knowledge and abstraction capabilities.

ImageNet-R ImageNet-R [20] contains artistic renditions of 200 ImageNet [12] classes, including art, cartoons, graffiti, embroidery, and other stylized representations. The dataset evaluates models' ability to generalize beyond standard photographic images to diverse visual representations.

ObjectNet ObjectNet [4] is a real-world test set for object recognition featuring random backgrounds, rotations, and viewpoints. Without an associated training set, it specifically tests generalization capabilities by controlling for common dataset biases and presenting objects in unusual poses and cluttered scenes.

HellaSwag HellaSwag [64] comprises 70,000 multiple-choice questions designed to evaluate commonsense natural language inference. The dataset features adversarially generated endings that are difficult for AI but easy for humans, who achieve over 95% accuracy.

WinoGrande WinoGrande [43] contains 44,000 examples inspired by the Winograd Schema Challenge [28]. Each example presents a sentence with an ambiguous pronoun that requires commonsense reasoning and world knowledge to resolve correctly.

ARC The AI2 Reasoning Challenge [8] provides multiple-choice questions from grade 3-9 science exams, split into Easy and Challenge sets. The Challenge Set specifically tests advanced reasoning capabilities that many current algorithms struggle with.

CommonsenseQA CommonsenseQA [49] contains 12,247 multiple-choice questions, each with one correct answer and four distractors. The dataset evaluates models' ability to utilize general world knowledge beyond specific contexts.

MMLU The Massive Multitask Language Understanding benchmark [21] evaluates language models across 57 subjects through approximately 16,000 multiple-choice questions. It tests world knowledge and problem-solving abilities in zero-shot and few-shot settings across diverse academic disciplines.

3.2 Analysis

Multinet consists of datasets across several categories, encompassing a diverse set of modalities and tasks, thus testing a potential generalist model in many different ways. **All control datasets are simulated environments except for the OpenX-Embodiment collection which contains data from robots used in the real-world.**

4 The Multinet Benchmark

4.1 Motivation

While existing benchmarks excel at evaluating specific capabilities and modalities, there remains a notable gap in holistic evaluation frameworks that can assess both the action capabilities of Vision-Language Models (VLMs) and the multimodal understanding of Vision-Language-Action Models (VLAs). Multinet addresses this gap by providing a comprehensive benchmark that spans vision-language, language, and control tasks. Our work consolidates diverse, high-quality datasets and establishes standardized evaluation metrics to enable systematic comparison of state-of-the-art models.

Dataset	Task/Data type	Modality
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
Fineweb-edu	High quality text corpus	Language

Table 1: Training, fine-tuning, and evaluation datasets for vision-language and language tasks

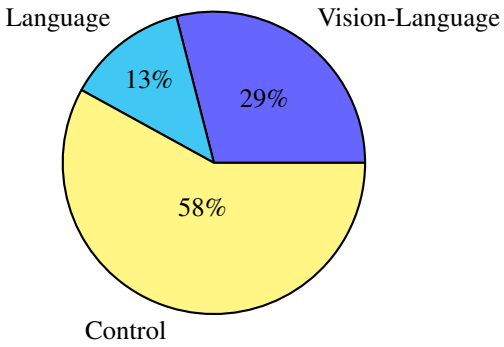

Figure 1: Distribution of datasets across modalities in Multinet. Control represents the largest portion (58%) due to the extensive OpenX-Embodiment collection, followed by Vision-Language (29%) and Language (13%) datasets.

4.2 Evaluation Metrics

We employ several complementary metrics to evaluate model performance across different modalities and tasks:

CIDEr The Consensus-based Image Description Evaluation (CIDEr) metric [54] evaluates image captioning quality by comparing generated captions to reference captions using TF-IDF-weighted n-grams. For a candidate caption c_i and reference captions S_i, the CIDEr score for n-grams of length n is calculated as:

$$\text{CIDEr}_n(c_i, S_i) = \left(\frac{1}{m} \sum_j \text{sim}(c_i, s_{ij}) \right) \tag{1}$$

where similarity is computed as:

$$\text{sim}(c_i, s_{ij}) = \frac{g^n(c_i) \cdot g^n(s_{ij})}{||g^n(c_i)|| \cdot ||g^n(s_{ij})||} \tag{2}$$

In the equation above $g^n(c_i)$ is the TF-IDF vector for a candidate description and $g^n(s_{ij})$ is the TF-IDF vector for the reference sentence j of image i. The final CIDEr score averages across different n-gram lengths:

$$\text{CIDEr}(c_i, S_i) = \frac{1}{N} \sum_n^N \text{CIDEr}_n(c_i, S_i) \tag{3}$$

VQA Accuracy This metric evaluates visual question answering performance through exact string matching between predicted and reference answers:

$$\text{VQA accuracy} = \frac{\sum_{i=1}^{N} \min\left(1, \frac{\text{number of annotators who agree with model's answer}}{3}\right)}{N}$$

Recall@K For image-text retrieval tasks, Recall@K measures the proportion of relevant items retrieved within the top K results:

$$\text{Recall@K} = \frac{\text{Number of relevant items retrieved within top K}}{\text{Total number of relevant items}}$$

Accuracy For commonsense reasoning and text understanding tasks, we use standard classification accuracy:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \times 100 \tag{4}$$

Mean Squared Error For evaluating action prediction on offline robotics trajectories, we employ Mean Squared Error:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5}$$

MSE is particularly suitable for this task due to its non-negativity, sensitivity to large errors, and incorporation of both bias and variance components. When evaluating models on the OpenX dataset, we use MSE to measure the accuracy of predicted actions given observation states, image observations, and language instructions at each timestep. This metric is especially valuable for offline evaluation where direct robot deployment is not possible.

Metric	**Evaluation Categories**
CIDEr	• Image Captioning • Image-based Text Retrieval
VQA Accuracy	Visual Question Answering
Recall@K	• Image Understanding • Text-based Image Retrieval
Accuracy	• Visual Question Answering • Commonsense Reasoning • Text Understanding
Mean Squared Error	• Reinforcement Learning • Robotics

Table 2: Metrics used in the Multinet benchmark and their corresponding evaluation categories. Each metric is designed to assess specific aspects of model performance across different modalities and tasks.

4.3 Benchmark Evaluation Protocol

The first release of Multinet establishes evaluation protocols for state-of-the-art models including GPT-4, JAT, and OpenVLA on the OpenX-Embodiment datasets [10]. Our evaluation strategy adapts to the varying availability of dataset splits:

Pre-existing Test Splits For datasets with established test splits, we conduct evaluations directly on these splits, maintaining consistency with previous benchmarking efforts.

Validation Split Usage When test splits are unavailable but validation splits exist, we use the validation splits for evaluation to avoid potential training data contamination.

Custom Split Creation For datasets containing only training data, we create evaluation splits by:

- Reserving approximately 20% of the dataset shards for evaluation
- Ensuring each shard contains 512 timesteps of data
- Extending the evaluation split boundary when necessary to include complete episodes (i.e., if an episode extends beyond the 20% boundary, we include all timesteps until episode completion)

This protocol ensures comprehensive evaluation while maintaining episode integrity and preventing the artificial truncation of behavioral sequences.

5 Importance of Multinet

The development and release of Multinet represents a significant step toward advancing generalist AI systems. Its importance spans several key dimensions:

Advancing Generalist Foundation Models Multinet establishes a comprehensive benchmark for evaluating truly generalist models that can operate across multiple modalities, tasks, and environments. Our initial findings [] demonstrate a significant capability gap in current state-of-the-art models: while VLMs and VLAs excel in their primary domains, they struggle to maintain consistent performance across a diverse set of real-world robotics tasks that they have not been exposed to before.

Enabling Next-Generation VLA Models Current Vision-Language-Action models typically excel at vision-language-grounded actions but may underperform in pure vision-language or language tasks. Multinet provides pre-training scale data across all these modalities, enabling the development of models that achieve state-of-the-art performance across all constituent tasks, not just their primary domain. As future versions expand the dataset, these capabilities will only grow stronger.

Comprehensive Evaluation of Robotics Foundation Models Traditional evaluations of robotics foundation models often focus narrowly on control tasks, neglecting to assess their capabilities in understanding and generating content across individual modalities. Multinet's comprehensive evaluation framework highlights opportunities to enhance these models' fundamental capabilities, paving the way for more powerful, truly generalist robotic systems.

Standardizing Robotics Data A significant contribution of this work is our open-source toolkit for standardizing robotics and reinforcement learning data. Many existing datasets suffer from outdated formats, poor maintenance, and accessibility issues. Our toolkit addresses these challenges by:

- Providing stable access methods for diverse RL and robotics datasets
- Converting various data formats to a unified TensorFlow dataset format
- Enabling easy local storage and usage for training, fine-tuning, and evaluation

Fostering Advanced Evaluation Methods While we currently use Mean Squared Error between predicted and ground-truth actions to evaluate performance on the OpenX-Embodiment dataset, we acknowledge this metric's limitations. By highlighting the need for more sophisticated evaluation methods, we aim to encourage:

- Development of more advanced simulation environments
- Creation of evaluation methods that better approximate real-world performance
- Innovation in offline RL and robotics task assessment

Through these contributions, Multinet not only provides immediate value for evaluating and developing generalist AI systems but also highlights critical areas for future research and development in the field.

6 Future Work

The release of Multinet marks an important first step toward a new paradigm of foundation models, but significant opportunities remain for expansion and improvement. Our vision for future iterations of Multinet encompasses several ambitious directions.

A key priority is deepening our understanding of how control-task training affects model capabilities. While current VLAs show promising results on control tasks, we plan to systematically evaluate their performance on pure vision-language and language tasks to assess whether fine-tuning or co-fine-tuning on control tasks compromises their capabilities in individual modalities. This investigation will provide crucial insights into the trade-offs involved in developing truly generalist models.

We also aim to broaden our evaluation scope beyond the OpenX-Embodiment dataset. By incorporating the diverse control tasks described in this paper, we can better understand how VLAs and generalist models perform on completely out-of-distribution data. This expansion will help identify both the strengths and limitations of current approaches while suggesting directions for improvement.

While our current profiling efforts focus on zero-shot performance, future work will explore few-shot learning and fine-tuning scenarios. Of particular interest is the potential transfer of VLA capabilities to novel domains. For instance, we are exploring how these models might be adapted to software environments, potentially enabling more capable digital agents by leveraging insights from embodied learning.

Finally, we envision transforming Multinet from its current offline form into an online benchmark. This evolution may include the development of simulation environments for both 2D and 3D control tasks, enabling more dynamic and interactive evaluation of model capabilities. Such an extension would provide richer insights into model behavior and better approximate real-world deployment scenarios.

Through these future developments, we aim to establish Multinet as a comprehensive and evolving platform for advancing the field of generalist AI systems. By providing increasingly sophisticated tools and benchmarks, we hope to accelerate progress toward more capable and versatile artificial intelligence.

7 References

References

[1] Common crawl. URL https://commoncrawl.org/.

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

[3] Firas Al-Hafez, Guoping Zhao, Jan Peters, and Davide Tateo. Locomujoco: A comprehensive imitation learning benchmark for locomotion, 2023. URL https://arxiv.org/abs/2311.02496.

[4] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *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.

[5] Charles Beattie, Joel Z. Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, Julian Schrittwieser, Keith Anderson, Sarah York, Max Cant, Adam Cain, Adrian Bolton, Stephen Gaffney, Helen King, Demis Hassabis, Shane Legg, and Stig Petersen. Deepmind lab, 2016. URL https://arxiv.org/abs/1612.03801.

[6] Yonatan Bitton, Nitzan Bitton Guetta, Ron Yosef, Yuval Elovici, Mohit Bansal, Gabriel Stanovsky, and Roy Schwartz. Winogavil: Gamified association benchmark to challenge vision-and-language models, 2022. URL https://arxiv.org/abs/2207.12576.

[7] Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. Babyai: A platform to study the sample efficiency of grounded language learning, 2019. URL https://arxiv.org/abs/1810.08272.

[8] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

[9] Karl Cobbe, Christopher Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning, 2020. URL https://arxiv.org/abs/1912.01588.

[10] Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Andrey Kolobov, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Felipe Vieira Frujeri, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guangwen Yang, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homanga Bharadhwaj, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jay Vakil, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi "Jim" Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick "Tree" Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Mart'in-Mart'in, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shubham Tulsiani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vikash Kumar, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiangyu Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yansong Pang, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yongqiang Dou, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open x-embodiment: Robotic learning datasets and rt-x models, 2024. URL https://arxiv.org/abs/2310.08864.

[11] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. *Advances in Neural Information Processing Systems*, 35:5982–5994, 2022.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi:10.1109/CVPR.2009.5206848.

[13] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra,

Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best,

Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

[14] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning, 2021. URL https://arxiv.org/abs/2004.07219.

[15] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets, 2023. URL https://arxiv.org/abs/2304.14108.

[16] Quentin Gallouédec, Edward Beeching, Clément Romac, and Emmanuel Dellandréa. Jack of all trades, master of some, a multi-purpose transformer agent, 2024. URL https://arxiv.org/abs/2402.09844.

[17] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017. URL https://arxiv.org/abs/1612.00837.

[18] Caglar Gulcehre, Ziyu Wang, Alexander Novikov, Tom Le Paine, Sergio Gomez Colmenarejo, Konrad Zolna, Rishabh Agarwal, Josh Merel, Daniel Mankowitz, Cosmin Paduraru, Gabriel Dulac-Arnold, Jerry Li, Mohammad Norouzi, Matt Hoffman, Ofir Nachum, George Tucker, Nicolas Heess, and Nando de Freitas. Rl unplugged: A suite of benchmarks for offline reinforcement learning, 2021. URL https://arxiv.org/abs/2006.13888.

[19] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people, 2018. URL https://arxiv.org/abs/1802.08218.

[20] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization, 2021. URL https://arxiv.org/abs/2006.16241.

[21] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

[22] Minho Heo, Youngwoon Lee, Doohyun Lee, and Joseph J Lim. Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation. *arXiv preprint arXiv:2305.12821*, 2023.

[23] Jie Huang, Wei Ping, Peng Xu, Mohammad Shoeybi, Kevin Chen-Chuan Chang, and Bryan Catanzaro. Raven: In-context learning with retrieval augmented encoder-decoder language models. *arXiv preprint arXiv:2308.07922*, 2023.

[24] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. Rlbench: The robot learning benchmark & learning environment. arxiv e-prints, art. *arXiv preprint arXiv:1909.12271*, 2019.

[25] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. 2024. URL https://arxiv.org/abs/2406.09246.

[26] Ivan Krasin, Tom Duerig, Neil Alldrin, Andreas Veit, Sami Abu-El-Haija, Serge Belongie, David Cai, Zheyun Feng, Vittorio Ferrari, and Victor Gomes. Openimages: A public dataset for large-scale multi-label and multi-class image classification., 01 2016.

[27] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents, 2023. URL https://arxiv.org/abs/2306.16527.

[28] Hector J. Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press, 2012. ISBN 9781577355601.

[29] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023. URL https://arxiv.org/abs/2307.16125.

[30] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A. Lee, Yuke Zhu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Multibench: Multiscale benchmarks for multimodal representation learning, 2021. URL https://arxiv.org/abs/2107.07502.

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

[32] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[33] Cong Lu, Philip J. Ball, Tim G. J. Rudner, Jack Parker-Holder, Michael A. Osborne, and Yee Whye Teh. Challenges and opportunities in offline reinforcement learning from visual observations, 2023. URL https://arxiv.org/abs/2206.04779.

[34] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts, 2024. URL https://arxiv.org/abs/2310.02255.

[35] Jianlan Luo, Charles Xu, Fangchen Liu, Liam Tan, Zipeng Lin, Jeffrey Wu, Pieter Abbeel, and Sergey Levine. Fmb: A functional manipulation benchmark for generalizable robotic learning. *The International Journal of Robotics Research*, page 02783649241276017, 2023.

[36] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge, 2019. URL https://arxiv.org/abs/1906.00067.

[37] Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants, 2023. URL https://arxiv.org/abs/2311.12983.

[38] Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale, 2024. URL https://arxiv.org/abs/2406.17557.

[39] Aleksei Petrenko, Zhehui Huang, Tushar Kumar, Gaurav Sukhatme, and Vladlen Koltun. Sample factory: Egocentric 3D control from pixels at 100000 FPS with asynchronous reinforcement learning. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7652–7662. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/petrenko20a.html.

[40] Wilbert Pumacay, Ishika Singh, Jiafei Duan, Ranjay Krishna, Jesse Thomason, and Dieter Fox. The colosseum: A benchmark for evaluating generalization for robotic manipulation, 2024. URL https://arxiv.org/abs/2402.08191.

[41] Sabela Ramos, Sertan Girgin, Léonard Hussenot, Damien Vincent, Hanna Yakubovich, Daniel Toyama, Anita Gergely, Piotr Stanczyk, Raphael Marinier, Jeremiah Harmsen, Olivier Pietquin, and Nikola Momchev. Rlds: an ecosystem to generate, share and use datasets in reinforcement learning, 2021. URL https://arxiv.org/abs/2111.02767.

[42] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A generalist agent. 2022. URL https://arxiv.org/abs/2205.06175.

[43] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale, 2019. URL https://arxiv.org/abs/1907.10641.

[44] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. URL https://arxiv.org/abs/2210.08402.

[45] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

[46] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge, 2022. URL https://arxiv.org/abs/2206.01718.

[47] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi:10.18653/v1/P18-1238. URL https://aclanthology.org/P18-1238.

[48] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read, 2019. URL https://arxiv.org/abs/1904.08920.

[49] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge, 2019. URL https://arxiv.org/abs/1811.00937.

[50] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. Deepmind control suite, 2018. URL https://arxiv.org/abs/1801.00690.

[51] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy, 2024. URL https://arxiv.org/abs/2405.12213.

[52] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033, 2012. doi:10.1109/IROS.2012.6386109.

[53] Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, et al. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19528–19540, 2023.

[54] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015. URL https://arxiv.org/abs/1411.5726.

[55] Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. Livebench: A challenging, contamination-free llm benchmark, 2024. URL https://arxiv.org/abs/2406.19314.

[56] Annie Xie, Lisa Lee, Ted Xiao, and Chelsea Finn. Decomposing the generalization gap in imitation learning for visual robotic manipulation, 2023. URL https://arxiv.org/abs/2307.03659.

[57] Eliot Xing, Abhinav Gupta, Sam Powers, and Victoria Dean. Kitchenshift: Evaluating zero-shot generalization of imitation-based policy learning under domain shifts. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

[58] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models, 2023. URL https://arxiv.org/abs/2306.09265.

[59] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, Jing Shao, and Wanli Ouyang. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark, 2023. URL https://arxiv.org/abs/2306.06687.

[60] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi:10.1162/tacl_a_00166. URL https://aclanthology.org/Q14-1006.

[61] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Avnish Narayan, Hayden Shively, Adithya Bellathur, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning, 2021. URL https://arxiv.org/abs/1910.10897.

[62] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities, 2023. URL https://arxiv.org/abs/2308.02490.

[63] 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, 2024. URL https://arxiv.org/abs/2311.16502.

[64] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence?, 2019. URL https://arxiv.org/abs/1905.07830.